2008春季 無母數統計講義

張福春

中山大學應用數學系

fuchuen@gmail.com

PRACTICAL NONPARAMETRIC STATISTICS

Third Edition (1999)

W. J. Conover

2008-02-18 ∼ 2008-06-15

This page intentionally left blank

Contents

PREFACE

Contents

無母數統計課程介紹（2008春季）

上課時間	週一10:10∼12:00、週三11:10∼12:00
上課地點	理4009-1
討論時間	週四、五　10:00∼12:00
授課教師	張福春 教授 理學院 3002-4 Tel: (O) (07) 525-2000 ext 3823 Email: `fuchuen@gmail.com`
講授方式	課堂講授

教材課本	Conover, W. J. (1999).
	Practical Nonparametric Statistics, 3rd edition.
	`http://as.wiley.com/WileyCDA/WileyTitle/`
	`productCd-0471160687.html`
	ISBN: 0471160687　Publisher: Wiley
	中山大學復文書局　Tel: (07) 525-0930

成績評量	作業(每章繳一次): 25%; 二次考試：60%; 報告: 15% 第一次考試(第一章∼第三章) 30%　第二次考試(第四章∼第五章) 30%
軟體	R, Splus, SPSS, SAS, Statistica,. . .
教學內容	Chap. 1: Probability Theory Chap. 2: Statistical Inference Chap. 3: Some Tests Based on the Binomial Distribution Chap. 4: Contingency Tables Chap. 5: Some Tests Based on Ranks

教學設備	硬體：電腦、單槍投影機 軟體：Yap

作業

第一章	*E1.10, P1.1, E2.12, E2.16, E3.6, P3.1, E4.10, E4.12, E5.4, E5.10*
第二章	*E1.4, E1.6, P1.2, E2.2, E2.6, E2.8, E3.2, E3.6, E4.4, E4.6*
第三章	*E1.2, E1.8, E2.2, E2.6, E3.4, E3.10, E4.2, E4.4, E5.4, E5.6*
第四章	*E1.5, E1.8, E2.4, E2.8, E3.2, E3.6, E4.4, E5.2, E5.6, E6.2*
第五章	*E1.1, E1.5, E2.2, E2.6, E3.1, E3.3, E4.2, E4.4, E5.1, E5.2*

哥林多前書10:23　凡事都可行，但不都有益處；凡事都可行，但不都建造人。

1 Corinthians 10:23　All things are lawful, but not all things are profitable; all things are lawful, but not all things build up.

有用網頁

1. Statistical Computing (UCLA Academic Technology Services)
   `http://www.ats.ucla.edu/stat/`

Part I

Lecture Notes

Chapter 1

Probability Theory

Contents

Preliminary remarks

Nonparametric statistical methods: Not necessary to be an expert in probability theory to understand the theory behind the methods.

With a few easily learned, elementary concepts, the basic fundamentals underlying most nonparametric statistical methods become quite accessible.

Recommended procedure for study: Read the text, pencil through the examples, work the exercises and problems.

1.1 Counting

Process of computing probabilities often depends on being able to count. Some sophisticated methods of counting are developed to handle those complicated situations.

- Toss a coin once: H or T
- Toss a coin twice: HH, HT, TH or TT
- Toss a coin n times: 2^n possible outcomes

Experiment: A process of following a well-defined set of rules, where the result of following those rules is not known prior to the experiment.
Model:

- The value of coin tossing is that it serves as a prototype for many different models in many different situations.

- ▶ Good models: Tossing coins, rolling dice, drawing chips from a jar, placing balls into boxes.
- ▶ They serve as useful and simple prototypes of many more complicated models arising from experimentation in diverse areas.
- ▶ Excellent study of the diversity of models above is given by Feller (1968).

Event: Possible outcomes of an experiment.

Rule 1.1.1 *If an experiment consists of n trials where each trial may result in one of k possible outcomes, there are k^n possible outcomes of the entire experiment.*

Example 1.1.1 *Suppose an experiment is composed of seven trials, where each trial consist of throwing a ball into one of the three boxes.*

- ▶ First throw: 3 different outcomes.
- ▶ First two throws: $3^2 = 9$ outcomes.
- ▶ Seven throws: $3^7 = 2187$ different outcomes. □

Rule 1.1.2 (Permutation) *There are $n!$ ways of arranging n distinguishable objects into a row.*

Example 1.1.2 *Consider the number of ways of arranging the letters A, B and C in a row.*

- ▶ First letter can be any of the three letters.
- ▶ Second letter can be chosen two different ways once the first letter is selected.
- ▶ The remaining letter becomes the final letter selected.
- ▶ Total: $(3)(2)(1) = 6$ different arrangements:

$$ABC, ACB, BAC, BCA, CAB, CBA.$$ □

Example 1.1.3 *Suppose that in a horse race there are eight horses. If you correctly predict which horse will win the race and which horse will come in second and wager to that effect, you are said to "win the exacta".*

- ▶ Win the exacta: Need to purchase $(8)(7) = 56$ betting tickets.
- ▶ Outcomes of all eight positions: $8! = 40320$ different ways. □

Rule 1.1.3 (Multinomial coefficient) *If a group of n objects is composed of n_1 objects of type 1, n_2 identical objects of type $2, \ldots, n_r$, identical objects of type r, the number of distinguishable arrangements into a row, denoted by*

$$\binom{n}{n_1, \ldots, n_r} = \frac{n!}{n_1! \ldots n_r!}.$$

In particular, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ if $n_1 = k$ and $n_2 = n - k$.

Example 1.1.4 (In example 2) *Suppose A and B are identical. We will denote them by the letter X, then*

- Original $3! = 6$ arrangements.
- Reduce to $\binom{3}{2} = 3$ distinguishable arrangements, XXC, XCX and CXX. □

Example 1.1.5 *In a coin tossing experiment where a coin is tossed five times, the result is two heads and three tails.*

- The number of different sequences of two heads and three tails equals the number of distinguishable arrangements of two objects of one kind and three objects of another, which is $\binom{5}{2} = 10$.

$$\begin{gathered} HHTTT, THHTT, TTHHT, HTHTT, THTHT, \\ TTHTH, HTTHT, THTTH, TTTHH, HTTTH. \end{gathered}$$

- How many different groups of k objects may be formed from n objects? $\binom{n}{k}$ □

Example 1.1.6 *Consider again the three letters A, B and C. The number of ways of selecting two of these letters is $\binom{3}{2} = 3$, that is, AB, BC and BC.*

- To see how this relates to the previous discussion, we will "tag" two of the three letters with an asterisk (*) denoting the tag.

$$\begin{aligned} A^*B^*C &\text{ gives } AB \\ A^*BC^* &\text{ gives } AC \\ \text{and } AB^*C^* &\text{ gives } BC \end{aligned}$$ □

Binomial coefficient

- *Binomial coefficient:* $\binom{n}{i}$
- *Binomial expansion:*

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

- *Multinomial coefficient:* $\binom{n}{n_1,\ldots,n_r}$
- *Multinomial expansion:*

$$(x_1 + \cdots + x_r)^n = \sum_{n_1+\cdots+n_r=n} \binom{n}{n_1,\ldots,n_r} x_1^{n_1} \cdots x_r^{n_r}$$

- Evaluate $(2+3)^4$ by binomial expansion:

$$(2+3)^4 = \sum_{i=0}^{4} \binom{4}{i} 2^i 3^{4-i} = 625$$

1.2 Probability

Definition 1.2.1 (Sample space) *The sample space is the collection of all possible different outcomes of an experiment.*

Definition 1.2.2 (Sample point) *A point in the sample space is a possible outcome of an experiment.*

Example 1.2.1 *If an experiment consists of tossing a coin twice, the sample space consists of the four points HH, HT, TH and TT.* □

Example 1.2.2 *An examination consisting of 10 "true or false" questions is administered to one student as an experiment. There are $2^{10} = 1024$ points in the sample space, where each point consists of the sequence of possible answers to the ten successive questions, such as "TTFTFFTTTT".* □

Definition 1.2.3 (Event) *An event is any set of points in the sample space.*

- *Empty set:* A set with no points in it.
- *Sure event:* The event consisting of all points in the sample space.
- *Mutually exclusive events:* If two events have no points in common.
- *Contained in:* $A \subseteq B$

Probability

- To each point in the sample space there corresponds a number called *the probability of the point or the probability of the outcome*.

Definition 1.2.4 (Probability of an event) *If A is an event associated with an experiment, and if n_A represents the number of times A occurs in n independent repetitions of the experiment, the probability of the event A, denoted by $P(A)$, is given by*

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \tag{1}$$

which is read "the limit of the ratio of the number of times A occurs to the number of times the experiment is repeated, as the number of repetitions approaches infinity".

Probability function

- The set of probabilities associated with a particular sample space is seldom known, but the probabilities are assigns according to the experimenter's preconceived notions.

Example 1.2.3 *If an experiment consisting of the single toss of an unbiased coin, it is reasonable to assume that the outcome H will occur about half the time.*

- $P(H) = 1/2$ and $P(T) = 1/2$. □

Example 1.2.4 *If an experiment consisting of three tosses of an unbiased coin, it is reasonable to assume that each of the* $2^3 = 8$ *outcomes* $HHH, HHT, HTH, HTT, THH, THT, TTH, TTT$ *is equally likely.*

- The probability of each outcome is $1/8$.
- $P(3\ \text{tails}) = 1/8$, $P(\text{at least one head}) = 7/8$, and $P(\text{more heads than tails}) = P(\text{at least two heads}) = 4/8 = 1/2$. □

Definition 1.2.5 (*Probability function*) *A probability function is a function that assigns probabilities to the various events in the sample space.*

Conditional probability

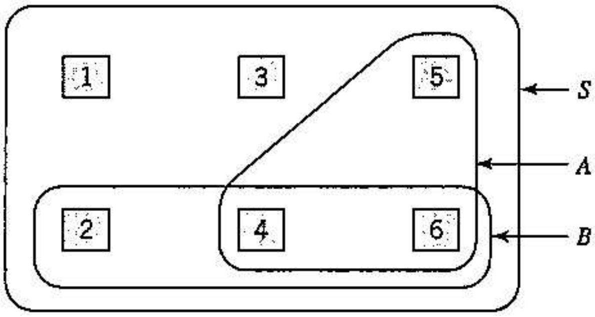


Fig. 1

Definition 1.2.6 (*Probability of joint events*) *If A and B are two events in a sample space S, the event "both A and B occur", representing those points in the sample space that are in both A and B at the same time, is called the joint event A and B and is represented by AB. The probability of the joint event is represented by $P(AB)$.*

$$P(A|B) = \lim_{n\to\infty} \frac{n_{AB}}{n_B} = \lim_{n\to\infty} \frac{n_{AB}/n}{n_B/n} = \frac{P(AB)}{P(B)} \tag{3}$$

Definition 1.2.7 (*Conditional probability*) *The conditional probability of A given B is the probability that A occurred given that B occurred and is given by*

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{4}$$

where $P(B) > 0$. If $P(B) = 0$, $P(A|B)$ is not defined.

Example 1.2.5 *Consider the rolling of a fair die, let A be the event "a 4,5, or 6 occurs" and let B be the event "an even number occurs".*

- $P(AB) = P(4 \text{ or } 6) = 2/6 = 1/3$.
- $P(B) = 3/6 = 1/2$.
- $P(A|B) = \frac{P(AB)}{P(B)} = \frac{1/3}{1/2} = \frac{2}{3}$. □

Independent events

Definition 1.2.8 (Independent events) *Two events A and B are independent if*

$$P(AB) = P(A)P(B). \tag{5}$$

Example 1.2.6 *In an experiment consisting of two tosses of a balanced coin, the four points in the sample space are assumed to have equal probabilities. Let A be the event "a head occurs on the first toss" and let B be the event "a head occurs on the second toss."*

- A: HH and HT.
- B: HH and TH.
- AB: HH.
- $P(A) = 2/4$, $P(B) = 2/4$ and $P(AB) = 1/4$.
- A and B are independent. □

Example 1.2.7 *Consider the experiment consisting of one roll of a balanced die, where the sample space consists of the six equally likely points.*

- A: "an even number occurs"
- B: "at least a 4 occurs"
- C: "at least a 5 occurs"
- A and B are not independent because $P(A)P(B) = (1/2)(1/2)$, or $1/4$ while $P(AB) = 1/3$.
- A and C are independent, because $P(A)P(C) = 1/6$, the same as $P(AC)$. □

Independent experiments

Definition 1.2.9 (Two independent experiments) *Two experiments are independent if for every event A associated with one experiment and every event B associated with the second experiment,*

$$P(AB) = P(A)P(B).$$

It is equivalent to define two experiments as independent if every event associated with one experiment is independent of every event associated the other experiment.

Definition 1.2.10 (n independent experiments) *n experiments are mutually independent if for every set of n events, formed by considering one event from each of the n experiments, the following equation is true:*

$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2)\cdots P(A_n) \tag{6}$$

where A_i represents an outcome of the ith experiment, for $i = 1, 2, \ldots, n$.

Example 1.2.8 *Consider a biased coin, where H: probability p, and T: probability $q = 1 - p$.*

- Consider three independent repetitions of the experiment,

$$\begin{aligned} P(H_1T_2H_3) &= P(H_1)P(T_2)P(H_3) = pqp \\ P(\text{exactly two heads}) &= \binom{3}{2}p^2q = 3p^2q \\ P(\text{exactly } k \text{ heads}) &= \binom{n}{k}p^kq^{n-k} \end{aligned}$$

where $P(H) = p$. □

1.3 Random variables

Definition 1.3.1 (Random variable) *A random variable is a function that assigns real numbers to the points in a sample space.*

Notation

- Random variables: W, X, Y or Z (capital cases)
- Observed values: w, x, y or z (lower cases)

Example 1.3.1 *A consumer is given a choice of three products, soap, detergent, or Brand A. Let the random variable assign the number 1 to the choice "Brand A" and the number 0 to the other two possible outcomes. Then $P(X = 1)$ equals the probability that the consumer chooses Brand A.* □

Example 1.3.2 *Six girls and eight boys are each asked whether they communicate more easily with their mother or their father.*

- X: Number of girls who feel they communicate more easily with their mother.
- Y: Total number of children who feel they communicate more easily with their mother.
- $X = 3$: The event "3 girls feel they communicate more easily with their mothers.
- $Y = 7$: The event "3 girls and 4 boys feel they communicate more easily with their mothers. □

Example 1.3.3 *Toss a coin twice. Let X denote the number of heads*

- $X = 1$: The event contains only the points HT and TH. □

Definition 1.3.2 (Conditional probability) *The conditional probability of X given Y, written*

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)},$$

is the probability that the random variable X assumes the value x, given that the random variable Y has assumed the value y.

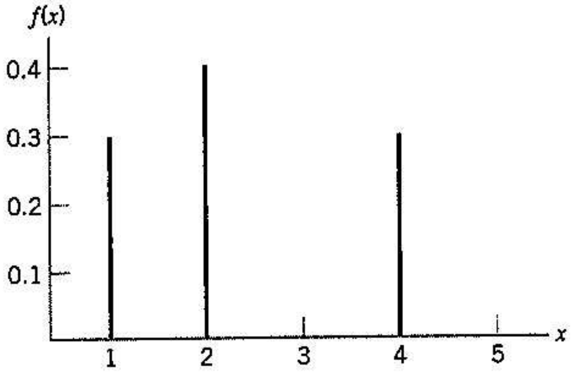
Example 1.3.4 (In Example 2.) *Let $Z = Y - X$ be the number of boys,*

$$\begin{aligned} P(X = 3, Y = 7) &= P(X = 3, Z = 4) \\ &= P(X = 3)P(Z = 4) \\ &= \binom{6}{3} p^3 (1-p)^3 \binom{8}{4} p^4 (1-p)^4 \end{aligned}$$

$$\begin{aligned} P(Y = 7) &= \binom{14}{7} p^7 (1-p)^7 \\ P(X = 3 | Y = 7) &= \frac{P(X = 3, Y = 7)}{P(Y = 7)} = .408 \end{aligned} \qquad \square$$

Definition 1.3.3 (*Probability mass function, pmf*) *The probability mass function of the random variable X, usually denoted by $f(x)$, is the function that gives the probability of X assuming the value x, for any real number x. In other words*

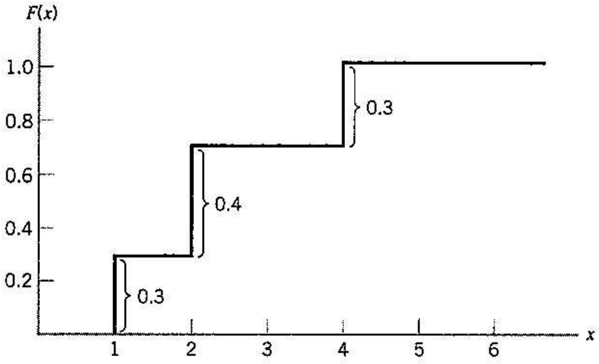
$$f(x) = P(X = x). \tag{5}$$



Definition 1.3.4 (*Cumulative distribution function, cdf*) *The cumulative distribution function of a random variable X, usually denoted by $F(x)$, is the function that gives the probability of X being less than or equal to any real number x. In other words,*

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t) \tag{6}$$

where the summation extends over all values of t that do not exceed x.



Definition 1.3.5 (Binomial distribution) *Let X be a random variable. The binomial distribution is the probability distribution represented by the probability function*

$$\begin{aligned} f(x) &= P(X = x) \\ &= \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \dots, n \end{aligned} \tag{7}$$

where n is a positive integer, $0 \le p \le 1$, and $q = 1 - p$. Note that we are using the usual convention that $0! = 1$.

Example 1.3.5 *An experiment consists of n independent trials where each trial may result in one of two outcomes, "success" or "failure", with probabilities p and q, respectively, such as with the tossing of a coin.*

- X: Total number of success in the n trials.
- $P(X = x) = \binom{n}{x} p^x q^{n-x}$
- Thus X has the binomial distribution. □

Definition 1.3.6 (Discrete uniform distribution) *Let X be a random variable. The discrete uniform distribution is the probability distribution represented by the probability function*

$$f(x) = \frac{1}{N}, \qquad x = 1, 2, \dots, N. \tag{9}$$

Example 1.3.6 *A jar has N plastic chips, numbered 1 to N. An experiment consists of drawing one chip from the jar, where each chip is equally likely to be drawn. Let X equal the number on the drawn chip. Then X has the discrete uniform distribution.* □

Definition 1.3.7 (Joint probability mass function) *The joint probability function $f(x_1, x_2, \dots, x_n)$ of the random variables $X_1, X_2, \dots, X_n$ is the probability of the joint occurrence of $X_1 = x_1, X_2 = x_2, \dots,$ and $X_n = x_n$. Stated differently,*

$$f(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n). \tag{10}$$

Definition 1.3.8 (Joint distribution function) *The joint distribution function $F(x_1, x_2, \dots, x_n)$ of the random variables $X_1, X_2, \dots, X_n$ is the probability of the joint occurrence of $X_1 \le x_1, X_2 \le x_2, \dots,$ and $X_n \le x_n$. Stated differently,*

$$F(x_1, x_2, \dots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n). \tag{11}$$

Example 1.3.7 *Consider X and Y in Example 2:*

$$\begin{aligned} f(3,7) &= P(X = 3, Y = 7) = \binom{6}{3}\binom{8}{4} p^7 (1-p)^7 \\ F(3,7) &= P(X \le 3, Y \le 7) = \sum_{\substack{0 \le x \le 3 \\ x \le y \le 7}} f(x,y) \end{aligned}$$

where

$$f(x,y) = \binom{6}{x} p^x (1-p)^{6-x} \binom{8}{y-x} p^{y-x} (1-p)^{8-(y-x)} \qquad \square$$

Definition 1.3.9 (*Conditional probability*) *The conditional probability function of X given Y, $f(x|y)$, is*

$$f(x|y) = P(X = x|Y = y) = \frac{f(x,y)}{f(y)}. \tag{14}$$

Example 1.3.8 *As a continuation of Example 7, let $f(x|y)$ denote the conditional probability function of X given $Y = y$. Then*

- $f(3|7) = P(X = 3|Y = 7) = .408$
- $f(y) = P(Y = y) = \binom{14}{y} p^y (1-p)^{14-y}$

$$f(x|y) = \frac{f(x,y)}{f(y)} = \frac{\binom{6}{x}\binom{8}{y-x}}{\binom{14}{y}} \qquad \begin{array}{l} 0 \le x \le 6, \\ 0 \le y - x \le 8 \end{array} \tag{16}\square$$

Definition 1.3.10 (*Hypergeometric distribution*) *Let X be a random variable. The hypergeometric distribution is the probability distribution represented by the probability function*

$$f(x) = P(X = x) = \frac{\binom{A}{x}\binom{B}{k-x}}{\binom{A+B}{k}} \qquad \begin{array}{l} 0 \le x \le A, \\ 0 \le k - x \le B \end{array} \tag{17}$$

where A, B and k are nonnegative integers and $k \le A + B$.

Definition 1.3.11 (*Mutually independent*) *Let $X_1, X_2, \ldots, X_n$ be random variables with the respective probability functions $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ and with the joint probability function $f(x_1, x_2, \ldots, x_n)$. Then $X_1, X_2, \ldots, X_n$ are mutually independent if*

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n) \tag{18}$$

for all combinations of values of $x_1, x_2, \ldots, x_n$.

Example 1.3.9 (In Example 8.)

- The probability function of X, the number of girls who feel they communicate more easily with their mothers, out of 6 girls, is given by

$$f_1(x) = P(X = x) = \binom{6}{x} p^x (1-p)^{6-x}.$$

- The probability function of Y, the total number of children who feel they communicate more easily with their mothers, out of 14 children, is given by

$$f_2(y) = P(Y = y) = \binom{14}{y} p^y (1-p)^{14-y}.$$

- The joint probability function of X and Y being given by

$$f(x,y) = P(X = x|Y = y)P(Y = y) = \binom{6}{x}\binom{8}{y-x}p^y(1-p)^{14-y}.$$

- $f(x,y) \neq f_1(x)f_2(y)$ therefore, X and Y are not independent. □

1.4 Some properties of random variables

- We have already discussed some of the properties associated with random variables, such as their probability functions and their distribution functions.

- The distribution function describes all of the properties of a random variable that are of interest, because the distribution function reveals the possible values the random variable may assume and the probability associated with each value.

- We will now introduce some other properties of random variables.

Quantile v.s. median, quartile, decile and percentile

- The most common method used in this book for summarizing the distribution of a random variable is by giving some selected quantiles of the random variable.

- The term "quantile" is not as well known as the terms "median," "quartile," "decile," and "percentile".

Definition 1.4.1 (Quantile) *The number x_p for a given value of p between 0 and 1, is called the pth quantile of the random variable X, if $P(X < x_p) \leq p$ and $P(X > x_p) \leq 1-p$.*

- If more than one number satisfies the definition of the pth quantile, we will use that *x_p equals the average of the largest and the smallest number that satisfy Definition 1*.

- Median: 0.5 quantile

- Third decile: 0.3 quantile

- Upper and lower quartiles: 0.75 and 0.25 quantiles

- Sixty-third percentile: 0.63 quantile

- The easiest method of finding the pth quantile involves using the graph of the distribution function of the random variable.

Example 1.4.1 *Let X be a random variable*

$$P(X = 0) = \frac{1}{4}, \quad P(X = 1) = \frac{1}{4}, \quad P(X = 2) = \frac{1}{3}, \quad P(X = 3) = \frac{1}{6}.$$

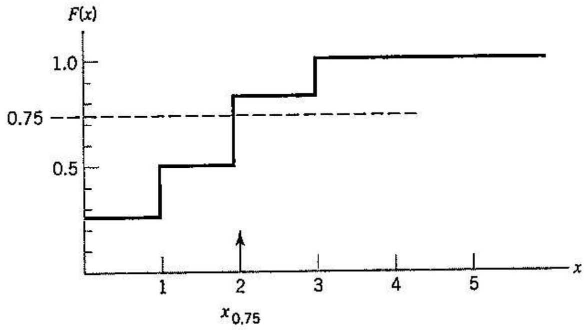


Fig. 2

- .75 quantile: $x_{.75} = 2$
- Median: $(1 + 2)/2 = 1.5$ □

Test statistics

- Certain random variables called "test statistics" play an important role in most statistical procedures.
- Test statistics are useless unless their distribution functions are at least partially known.
- Most of the tables in the appendix give information concerning the distribution functions of various test statistics used in nonparametric statistics.
- This information is condensed with the aid of quantiles.

Definition 1.4.2 (Expected value) *Let X be a random variable with the probability function $f(x)$ and let $u(X)$ be a real valued function of X. The expected value of $u(X)$, written $E[u(X)]$, is*

$$E[u(X)] = \sum_x u(x)f(x). \tag{1}$$

Our interest is confined mainly to two special expected values, the mean and the variance of X.

Definition 1.4.3 (Mean) *Let X be a random variable with the probability function $f(x)$. The mean of X,usually denoted by μ, is*

$$\mu = E(X). \tag{2}$$

Example 1.4.2 (Mean of Bernoulli distribution) *Consider a simple experiment.*

- p: Success
- $1 - p$: Failure
- $X \sim$ binomial with $n = 1$
- $E(X) = 1(p) + 0(1 - p) = p$ □

Example 1.4.3 *Consider a businessman who always eats lunch at the same restaurant, which has lunches priced at \$4.00, \$4.50, \$5.00 and \$5.50. Let X be the price of the lunch.*

- $P(X = 4) = .25, P(X = 4.5) = .35, P(X = 5) = .20, P(X = 5.5) = .20$
- $E(X) = 4(.25) + (4.5)(.35) + 5(.20) + (5.5)(.20) = 4.675$ □

Scale

- The properties of the random variable that measure the amount of spread, or variability, of the random variable are called "measures of scale."
- Measure of location: mean and median.
- Measure of scale: *interquartile range* $= x_{.75} - x_{.25}$, range and *standard deviation*.
- The most common measure of scale is the standard deviation, which equals the square root of the variance.

Definition 1.4.4 (Variance) *Let X be a random variable with mean μ and the probability function $f(x)$. The variance of X, usually denoted by σ^2 or by Var(X), is*

$$\sigma^2 = E[(X-\mu)^2] = E[X^2] - \mu^2. \tag{5}$$

Example 1.4.4 *(In Example 2.)*

- $\sigma^2 = (1-p)^2(p) + (0-p)^2(1-p) = p(1-p) = pq$
- $E(X^2) = (1)^2(p) + (0)^2(1-p) = p$
- $\sigma^2 = E(X^2) - \mu^2 = p - p^2 = p(1-p)$
- $\sigma = \sqrt{p(1-p)}$ □

Example 1.4.5 (Mean and variance of a fair die) *There are six identical chips numbered 1 to 6. A monkey has been trained to select one chip and give its trainer. Let X be the number on the chip. If each chip has probability 1/6 of being selected then X has the discrete uniform distribution.*

- $E(X) = \sum_{k=1}^{6} k(1/6) = 3\frac{1}{2}$
- $E(X^2) = \sum_{k=1}^{6} k^2(1/6) = 15\frac{1}{6}$
- $\text{Var}(X) = E(X^2) - \mu^2 = 15\frac{1}{6} - 3\frac{1}{2} = 2\frac{11}{12}$ □

Definition 1.4.5 (Expected value of multivariate distribution) *Let $X_1, X_2, \ldots, X_n$ be random variables with the joint probability function $f(x_1, x_2, \ldots, x_n)$, and let $u(X_1, X_2, \ldots, X_n)$ be a real valued function of $X_1, X_2, \ldots, X_n$. Then the expected value of $u(X_1, X_2, \ldots, X_n)$ is*

$$E[u(X_1, X_2, \ldots, X_n)] = \sum u(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n).$$

Theorem 1.4.1 (Mean of sum of random variables) *Let $X_1, X_2, \ldots, X_n$ be random variables and let*

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then

$$E(Y) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

Example 1.4.6 (Mean of binomial distribution) *$Y \sim$ binomial with parameters n and p. Then*

$$\begin{aligned} E(Y) &= E(X_1) + E(X_2) + \cdots + E(X_n) \\ &= np. \end{aligned}$$

□

Lemma 1.4.1 $\sum_{i=a}^{N} i = \frac{(N+a)(N-a+1)}{2}$ *and* $\sum_{i=1}^{N} i = \frac{(N+1)N}{2}$.

Example 1.4.7 (Mean of discrete uniform distribution) *N chips in a jar, numbered from 1 to N. Y the sum of the numbers on the n drawn chips and X_i is the number on the ith chip drawn.*

- $P(X_i = k) = \frac{1}{N}, \ k = 1, 2, \ldots, N$
- $E(X_i) = \sum_{k=1}^{N} k \left(\frac{1}{N}\right) = \frac{N+1}{2}$
- $E(Y) = E(X_1) + \cdots + E(X_n) = n\frac{N+1}{2}$ □

Definition 1.4.6 (Covariance) *Let X_1 and X_2 be two random variables with means μ_1 and μ_2, probability functions $f_1(x_1)$ and $f_2(x_2)$, respectively, and joint probability function $f(x_1, x_2)$. The covariance of X_1 and X_2 is*

$$\begin{aligned} Cov(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E(X_1X_2) - \mu_1\mu_2. \end{aligned}$$

$$\text{Cov}(X_1, X_2) = \sum (x_1 - \mu_1)(x_2 - \mu_2) f(x_1, x_2)$$

Example 1.4.8 *Any particular person having an automobile accident within a given year is about* 0.1.

- This probability becomes .3 if it is known that the person had an automobile accident the previous year.
- $X_1 = 0, 1$ dependent on whether a particular person has no accident or at least one accident during the first year of his or her insurance period.
- X_2 defined for the second year.
- $P(X_1 = 0) = .9, \ P(X_1 = 1) = .1$
- $E(X_1) = .1, \ E(X_2) = .1$
- $f(1, 1) = P(X_1 = 1, X_2 = 1) = P(X_2 = 1|X_1 = 1)P(X_1 = 1) = (.3)(.1) = .03$

- $E(X_1X_2) = (1)(1)f(1,1) = .03$
- $\mathrm{Cov}(X_1, X_2) = E(X_1X_2) - E(X_1)E(X_2) = .03 - (0.1)(0.1) = .02$ □

Correlation coefficient

- *Correlation coefficient*: a measure of linear dependence between two random variables.
- Although we will not prove it here, the correlation coefficient is always between −1 and +1.
- It equals zero when the two random variables are independent, although it may equal zero in other cases also.

Definition 1.4.7 (*Correlation coefficient*) *The correlation coefficient between two random variables is their covariance divided by the product of their standard deviations. That is, the correlation coefficient, usually denoted by ρ, between two random variables X_1 and X_2 is given by*

$$\rho = \frac{Cov(X_1, X_2)}{\sqrt{Var(X_1)\,Var(X_2)}}.$$

Lemma 1.4.2 (Sum of the first N positive integers) $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$.

Proof. (Construct a telescoping series)

- Consider $(i+1)^3 - i^3 = 3i^2 + 3i + 1$.
- Sum of the previous identities for $i = 0, 1, \ldots, N$, yields

$$\begin{aligned}\sum_{i=0}^{N}(i+1)^3 - i^3 &= \sum_{i=0}^{N} 3i^2 + 3i + 1\\ (N+1)^3 - 0^3 &= 3\sum_{i=0}^{N} i^2 + 3\sum_{i=0}^{N} i + \sum_{i=0}^{N} 1\\ &= 3\sum_{i=0}^{N} i^2 + \frac{3N(N+1)}{2} + (N+1)\end{aligned}$$

- It gives that $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$. □

Example 1.4.9 *Example 7 continued. N plastic chips numbered 1 to N, X_i equals the number on the ith chip drawn from the jar,*

- $E(X_i) = \frac{N+1}{2}$
- $\mathrm{Var}(X_i) = E(X_i^2) - [E(X_i)]^2 = \frac{(N+1)(N-1)}{12}$
- $f(x_i, x_j) = \frac{1}{N-1} \cdot \frac{1}{N},\ x_i \neq x_j$

$$\mathrm{Cov}(X_i, X_j) = E[X_iX_j] - E(X_i)E(X_j)$$

- ▶ Cov(() $X_i, X_j) = \sum_{k \neq s} \frac{ks}{(N-1)N} - \left(\frac{N+1}{2}\right)^2 = -\frac{N+1}{12}$ □

Theorem 1.4.2 (Independence implies zero covariance) *If X_1 and X_2 are independent random variables, the covariance of X_1 and X_2 is zero.*

- ▶ The converse of Theorem 2 is not necessarily true as the following example illustrates.

Example 1.4.10 (Dependent random variables with zero covariance) *Define* $P(X = 0, Y = 0) = 1/2,\ P(X = 1, Y = 1) = 1/4,\ P(X = -1, Y = 1) = 1/4,$ *then*

- ▶ $P(X = 0) = 1/2,\ P(X = 1) = 1/4,$
  $P(X = -1) = 1/4$
- ▶ $P(Y = 0) = 1/2,\ P(Y = 1) = 1/2$
- ▶ $E(X) = 0,\ E(Y) = 1/2$
- ▶ $\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = (1)(1/4) + (-1)(1/4) - (0)(1/2) = 0$
- ▶ X and Y are not independent, because

$$P(X = 0, Y = 0) = 1/2 \neq P(X = 0)P(Y = 0) = 1/4.$$ □

Theorem 1.4.3 (Variance of sum of random variables) *Let $X_1, X_2, \ldots, X_n$ be random variables and let*

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then

$$Var(Y) = \sum_{i=1}^{n} Var(X_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} Cov(X_i, X_j).$$

Proof.
To show that $\text{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m \text{Cov}(X_i, Y_j)$.

Let $\mu_i = E[X_i]$ and $\nu_j = E[Y_j]$. Then $E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mu_i$ and $E\left[\sum_{j=1}^m Y_j\right] = \sum_{j=1}^m \nu_j$

$$\begin{aligned}
\text{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) &= E\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right)\left(\sum_{j=1}^{m} Y_j - \sum_{j=1}^{m} \nu_j\right)\right] \\
&= E\left[\left(\sum_{i=1}^{n} (X_i - \mu_i)\right)\left(\sum_{j=1}^{m} (Y_j - \nu_j)\right)\right] \\
&= E\left[\sum_{i=1}^{n} \sum_{j=1}^{m} (X_i - \mu_i)(Y_j - \nu_j)\right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} E\left[(X_i - \mu_i)(Y_j - \nu_j)\right]
\end{aligned}$$

Let $Y_j = X_j, j = 1, \ldots, n$. Then

$$\begin{aligned}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) &= \operatorname{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right)\\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \operatorname{Cov}(X_i, X_j)\\ &= \sum_{i=1}^{n} \operatorname{Var}(X_i) + \sum_{i\neq j} \operatorname{Cov}(X_i, X_j) \qquad \square\end{aligned}$$

Example 1.4.11 (In Example 9) *Let X_i equal the number on the ith chip drawn as before, and let Y equal the sum of the X_is as in Example 7. Then*

$$\begin{aligned}Var(Y) &= n\frac{(N+1)(N-1)}{12} + n(n-1)\left(-\frac{N+1}{12}\right)\\ &= \frac{n(N+1)(N-n)}{12}. \qquad \square\end{aligned}$$

Example 1.4.12 (Variance of binomial distribution) *Consider n independent trials with success probability p. Let X_i equal to 0 or 1, depending on whether the ith trial results in "failure" or "success", respectively. Then*

$$\begin{aligned}Var(Y) &= \sum_{i=1}^{n} Var(X_i)\\ &= npq. \qquad \square\end{aligned}$$

Theorem 1.4.4 (Mean and variance of binomial distribution) *Let X be a random variable with the binomial distribution*

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

Then the mean and variance of X are given by

$$\begin{aligned}E(X) &= np\\ Var(X) &= npq.\end{aligned}$$

Theorem 1.4.5 *Let X be the sum of n integers selected at random, without replacement, from the first N integers 1 to N. Then the mean and variance of X are given by*

$$\begin{aligned}E(X) &= \frac{n(N+1)}{2}\\ Var(X) &= \frac{n(N+1)(N-n)}{12}.\end{aligned}$$

Example 1.4.13 *An advertising agency drew 12 samples magazine ads for one of their customers and ranked the ads from 1 to 12 on the basis of the agency's opinion of which ads would be the most effective in selling the product. The "most effective" ad was given the rank 1, and so on. The customer, the manufacture of the product, selected 4 ads for purchase. They were ranked $4, 6, 7,$ and 11 by the agency.*

- Assuming that the customer's choice and the agency's ranking were independent, the sum of the ranks on the selected ads should be distributed the same as the sum of the numbers on 4 chips selected at random out of 12 chips numbered 1 to 12.
- X: Sum of the ranks of 4 ads.
- From Theorem 5: $E(X) = \frac{(4)(12+1)}{2} = 26$
- $\mathrm{Var}(X) = \frac{(4)(12+1)(12-4)}{12} = 34\frac{2}{3}$
- $\sigma = \sqrt{\mathrm{Var}(X)} = 5.9$
- The observed value of X is

$$X = 4 + 6 + 7 + 11 = 28$$

which is close to $E(X) = 26$. □

1.5 Continuous random variables

- All of the random variables that we have introduced so far in this chapter have one property in common: their possible values can be listed.
- The list of possible values assumed by the binomial variables is $0, 1, 2, 3, 4, \ldots, n-1, n$. No other values may be assumed by the binomial random variable.
- The list of values that may be assumed by the discrete uniform random variable could be written as $1, 2, 3, \ldots, N$.
- Similar lists could be made for each random variable introduced in the previous definitions and examples.
- A more precise way of stating that the possible values of a random variable may be listed is to say that there exists a one-to-one correspondence between the possible values of the random variable and some or all of the positive integers.
- This means that to each possible value there corresponds one and only one positive integer, and that positive integer does not correspond to more than one possible value of the random variable. Random variables with this property are called discrete.
- All of the random variables we have considered so far are discrete random variables.
- However, the theorems we have proven hold for all random variables, even though we proved them only for discrete random variables.

Definition 1.5.1 (*Discrete random variable*) *A random variable X is discrete if there exists a one to one correspondence between the possible values of X and some or all of the positive integers.*

- The distribution function of a discrete random variable is a *step function*.

- ▶ The graph of the distribution function of a *continuous* random variable has no steps but rises only gradually.

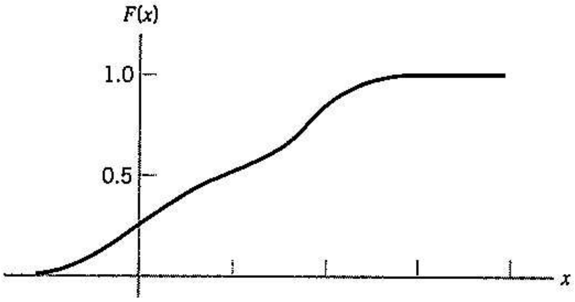


Fig. 3

Definition 1.5.2 (Continuous random variable) *A random variable X is continuous if no two quantiles x_{p_1} and x_{p_2} of X are equal to each other, where p_1 is not equal to p_2. Equivalently, a random variable X is continuous if $P(X \leq x)$ equals $P(X < x)$ for all numbers x.*

Example 1.5.1 *The distribution function graphed in Figure 4 is a continuous distribution function.*

- ▶ Typical continuous random variables: measuring time, weight, distance, volume, and so forth.

- ▶ In practice, no actual random variable is continuous. □

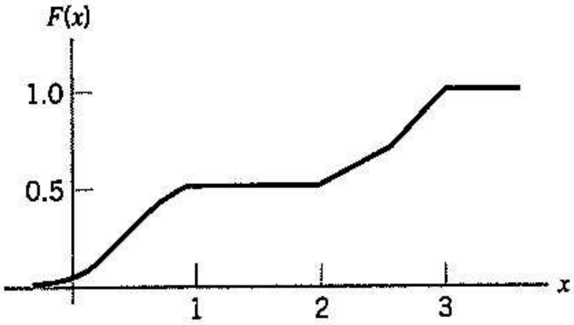


Fig. 4

Example 1.5.2 *The time it takes a racehorse to run a mile race is a continuous quantity, because time is generally a continuous quantity. In practice, the time is measured to the nearest 1/5 seconds. It is not unusual for a horse to run two races in identical lengths of times. The actual lengths of time will be exactly equal with probability zero; therefore it is reasonable to assume that the time of a rave, measured exactly, is a continuous random variable that is approximately equal to the measured time of the race.* □

- ▶ Another reason for considering continuous random variables is that the distribution function of a discrete random variable sometimes may be approximated by a continuous distribution function, resulting in a convenient method for computing desired probabilities associated with the discrete random variable.

- ▶ Two continuous distribution functions commonly used for this purpose are the normal distribution and the chi-squared distribution.

Normal distribution

Definition 1.5.3 (Normal distribution) *Let X be a random variable. Then X is said to have the normal distribution if the distribution of X is given by*

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \, \mathrm{d}y.$$

- The normal distribution function cannot be evaluated directly, and so Table A1 may be used to find approximate probabilities associated with normal random variables.
- Table A1 may be used to find approximate probabilities associated with normal random variables.
- Quantiles of normal random variables with mean μ and variance σ^2 may be found from Table A1, with the aid of the equations gives in the following theorem.

Theorem 1.5.1 (Normalization) *For a given value of p, let x_p be the pth quantile of a normal random variable with mean μ and variance σ^2, and let z_p be the pth quantile of a standard normal random variable. The quantile x_p may be obtained from z_p by using the relationship*

$$x_p = \mu + \sigma z_p.$$

Similarly,

$$z_p = \frac{x_p - \mu}{\sigma}.$$

Example 1.5.3 *Let $Z \sim N(0,1)$. Find $P(Z \leq 1.42)$.*

- $P(Z \leq 1.4187) = .922, P(Z \leq 1.4255) = .923.$
- Interpolate to get $P(Z \leq 1.42) \cong .9222.$ □

Example 1.5.4 *Let X be the IQ of a person selected at random from a large group of people. Assume that $X \sim N(100, 15^2)$. Find $P(X > 125)$.*

- $P(X > 125) = 1 - P(X \leq 125)$
- $z_p = \frac{x_p - \mu}{\sigma} = \frac{125-100}{15} = 1.67$
- $P(X \leq 125) = .95,\ P(X > 125) = .05$
- $P(X \leq x_{.99}) = .99$
- $x_{.99} = \mu + \sigma z_{.99} = 134.9$ □

Example 1.5.5 *A railroad company has observed over a period of time that the number X of people taking a certain train seems to follow $N(540, 32^2)$. How many seats should the company provide on the train if it wants to be 95% certain that everyone will have a seat?*

- $x_{.95} = \mu + \sigma z_{.95} = 592.6$ □

Central Limit Theorem

- The so-called central limit theorem appears in many different forms.
- All forms have in common the purpose of stating conditions under which the sum of several random variables may be approximated by a normal random variable.

Theorem 1.5.2 *(Central Limit Theorem) Let Y_n be the sum of the n random variables $X_1, X_2, \ldots, X_n$, let μ_n be the mean of Y_n and let σ^2 be the variance of Y_n. As n, the number of random variables, goes to infinity, the distribution function of the random variable*

$$\frac{Y_n - \mu_n}{\sigma_n}$$

approaches the standard normal distribution function.

- In practice, the number of random variables summed never goes to infinity.
- But the value of the central limit theorem is that in situations where the theorem PROBABILITY THEORY holds, the normal approximation is usually considered to be "reasonably good" as long as n is "large."
- The terms "reasonably good" and "large" are subjective terms; therefore much latitude exists in the practice of using the normal approximation.
- Usually if n is greater than 30 the normal approximation is satisfactory. However, sometimes when n is as small as 5 or 10 the normal approximation can be quite good.
- If one of the following sets of conditions holds:

  1. The X_i are independent and identically distributed, with $0 < \text{Var}(X_i) < \infty$.
  2. The X_i are independent but not necessarily identically distributed, but $E(X_i^3)$ exists for all i and satisfies certain conditions.
  3. The X_i are neither independent nor identically distributed, but represent the successive drawings, without replacement, of values from a finite population of size N, where N is greater than $2n$. Also a condition stated in Fisz (1963, p.523) should be satisfied.

Example 1.5.6 (Binomial distribution approximated by standard normal distribution) *Let $Y_n \sim Binomial(n, p)$. Then for large n*

$$\frac{Y_n - np}{\sqrt{npq}} \approx N(0,1).$$

- $Y_n = X_1 + X_2 + \cdots + X_n$ where $X_i \overset{\text{iid}}{\sim} \text{Binomial}(1, p)$
- $E(X_i) = p$ and $\text{Var}(X_i) = pq$
- $E(Y_n) = np$ and $\text{Var}(Y_n) = npq$ □

Example 1.5.7 *Consider the sampling scheme where n integers are selected at random, without replacement, from the first N integers, 1 to N.*

- $$Y_n = X_1 + X_2 + \cdots + X_n$$

  the set C conditions hold and the distribution function

  $$\frac{Y_n - \frac{n(N+1)}{2}}{\sqrt{\frac{n(N+1)(N-n)}{12}}} \approx N(0,1).$$

- The expectation and variance of Y_n are given in Theorem 1.4.5. □

Chi-squared distribution

Definition 1.5.4 (*Chi-squared distribution*) *A random variable X has the chi-squared distribution with k degrees of freedom if the distribution function of X is given by*

$$\begin{aligned} F(x) &= P(X \le x) \\ &= \begin{cases} \int_0^x \frac{y^{(k/2)-1}e^{-y/2}}{2^{k/2}\Gamma(k/2)}\,\mathrm{d}y & \text{if } x > 0, \\ 0 & \text{if } x \le 0. \end{cases} \end{aligned}$$

$E(X) = k$ and $Var(X) = 2k$.

- $X \sim \text{Gamma}(k/2, 1/2)$
- $E(X) = k$ and $\text{Var}(X) = 2k$
- Table A2 gives some selected quantiles of a chi-squared random variables.

Theorem 1.5.3 (Distribution of sum of squares of independent standard normal random *Let $X_1, X_2, \ldots, X_k$ be k independent and identically distributed standard normal random variables. Let Y be the sum of the squares of the X_i.*

$$Y = X_1^2 + X_2^2 + \cdots + X_k^2.$$

Then Y has the chi-squared distribution with k degrees of freedom.

Example 1.5.8 *A child psychologist asks each of 100 children to tell which of two trucks they would rather play with.*

- 42 children selected green trucks
- 58 children selected red trucks
- X: represents the number of children who selected green trucks
- Assume $X \sim \text{Binomial}(100, .5)$
- $\frac{X-50}{5} \approx N(0,1)$
- $X^* = \left(\frac{X-50}{5}\right)^2 \approx \chi_1^2$

- $P(X^* \leq 2.56) = 0.88$ □

Example 1.5.9 (In Example 8.) *The psychologist obtains two toy telephones, identical except that one is white and the other is blue. She asks each of 25 children to choose one to play with. Seventeen children chose the white telephone, and the other 8 preferred the blue telephone. Let Y be the random variable equal to the number of 25 children selecting the white telephone.*

- $\frac{Y-np}{\sqrt{npq}} = \frac{Y-(1/2)(25)}{5/2} \approx N(0,1)$
- $Y^* = \left(\frac{Y-(1/2)(25)}{5/2}\right)^2 \approx \chi_1^2$
- Combine data in Examples 8 and 9:

$$W = X^* + Y^* \sim \chi_2^2$$

- $W = 2.56 + 3.24 = 5.80$
- $P(W \geq 5.80) = 0.06$ □

Theorem 1.5.4 (Distribution of sum of independent chi-squared random variables) *Let $X_1, X_2, \ldots, X_n$ be independent chi-squared random variables with $k_1, k_2, \ldots, k_n$ degrees of freedom, respectively. Let Y equal the sum of the X_i. Then Y is a chi-squared random variable with k degrees of freedom, where*

$$k = k_1 + k_2 + \cdots + k_n.$$

- $M_{X_i}(t) = 1/(1-2t)^{k_i/2}$
- $M_X(t) = 1/(1-2t)^{\sum_{i=1}^n k_i/2}$
- $W = W_1 + W_2 + \cdots + W_k$ where $W_i \overset{\text{iid}}{\sim} \chi_1^2$.
- $E(W) = k$ and $\text{Var}(W) = 2k$.
- CLT: $Z = \frac{W-k}{\sqrt{2k}} \approx N(0,1)$ if $W \sim \chi_k^2$.
- $w_p = k + \sqrt{2k} z_p$
- $w_p = \left(z_p + \sqrt{2k-1}\right)^2$ and $k\left(1 - \frac{2}{9k} + z_p\sqrt{\frac{2}{9k}}\right)^3$ are better approximations given at the bottom of Table A2.

1.6 Summary

1. *Experiment:* A process of following a well-defined set of rules, where the result of following those rules is not known prior to the experiment. 2

2. *Event:* Possible outcomes of an experiment. 3

3. If an experiment consists of n trials where each trial may result in one of k possible outcomes, there are k^n possible outcomes of the entire experiment. 3

4. *Permutation:* There are $n!$ ways of arranging n distinguishable objects into a row. 3

5. *Multinomial coefficient:* If a group of n objects is composed of n_1 objects of type 1, n_2 identical objects of type $2, \ldots, n_r$, identical objects of type r, the number of distinguishable arrangements into a row, denoted by

$$\binom{n}{n_1, \ldots, n_r} = \frac{n!}{n_1! \ldots n_r!}.$$

In particular, $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ if $n_1 = k$ and $n_2 = n - k$. 3

6. *Binomial coefficient:* $\binom{n}{i}$ 4

7. *Binomial expansion:*

$$(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i y^{n-i}$$

. . . . 4

8. *Multinomial coefficient:* $\binom{n}{n_1,\ldots,n_r}$ 4

9. *Multinomial expansion:*

$$(x_1 + \cdots + x_r)^n = \sum_{n_1+\cdots+n_r=n} \binom{n}{n_1, \ldots, n_r} x_1^{n_1} \cdots x_r^{n_r}$$

. . . . 4

10. *Sample space*: The collection of all possible different outcomes of an experiment. . 5

11. *Sample point:* A possible outcome of an experiment. 5

12. *Event:* Any set of points in the sample space. 5

13. *Empty set:* A set with no points in it. 5

14. *Sure event:* The event consisting of all points in the sample space. 5

15. *Mutually exclusive events:* If two events have no points in common. 5

16. *Contained in:* $A \subseteq B$ 5

17. *Probability of an event:* If A is an event associated with an experiment, and if n_A represents the number of times A occurs in n independent repetitions of the experiment, the probability of the event A, denoted by $P(A)$, is given by

$$P(A) = \lim_{n \to \infty} \frac{n_A}{n} \tag{1}$$

which is read "the limit of the ratio of the number of times A occurs to the number of times the experiment is repeated, as the number of repetitions approaches infinity". 5

18. *Probability function:* A function that assigns probabilities to the various events in the sample space. 6

19. *Probability of joint events:* If A and B are two events in a sample space S, the event "both A and B occur", representing those points in the sample space that are in both A and B at the same time, is called the joint event A and B and is represented by AB. The probability of the joint event is represented by $P(AB)$. 6

20. *Conditional probability:* A given B is the probability that A occurred given that B occurred and is given by
$$P(A|B) = \frac{P(AB)}{P(B)} \tag{4}$$
where $P(B) > 0$. If $P(B) = 0$, $P(A|B)$ is not defined. 6

21. Two events A and B are independent if
$$P(AB) = P(A)P(B). \tag{5}$$
. . . . 7

22. *Two independent experiments:* Two experiments are independent if for every event A associated with one experiment and every event B associated with the second experiment,
$$P(AB) = P(A)P(B)$$
It is equivalent to define two experiments as independent if every event associated with one experiment is independent of every event associated the other experiment.7

23. *n independent experiments:* n experiments are mutually independent if for every set of n events, formed by considering one event from each of the n experiments, the following equation is true:
$$P(A_1A_2 \cdots A_n) = P(A_1)P(A_2) \cdots P(A_n) \tag{6}$$
where A_i represents an outcome of the ith experiment, for $i = 1, 2, \ldots, n$. 7

24. *Random variable:* A function that assigns real numbers to the points in a sample space. 8

25. *Conditional probability:* The conditional probability of X given Y, written
$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)},$$
is the probability that the random variable X assumes the value x, given that the random variable Y has assumed the value y. 8

26. *Probability mass function:* The probability mass function of the random variable X, usually denoted by $f(x)$, is the function that gives the probability of X assuming the value x, for any real number x. In other words
$$f(x) = P(X = x). \tag{5}$$
. . . . 9

27. *Distribution function:* The distribution function of a random variable X, usually denoted by $F(x)$, is the function that gives the probability of X being less than or equal to any real number x. In other words,

$$F(x) = P(X \leq x) = \sum_{t \leq x} f(t) \tag{6}$$

where the summation extends over all values of t that do not exceed x. 9

28. *Binomial distribution:* Let X be a random variable. The binomial distribution is the probability distribution represented by the probability function

$$\begin{aligned} f(x) &= P(X = x) \\ &= \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, \ldots, n \end{aligned} \tag{7}$$

where n is a positive integer, $0 \leq p \leq 1$, and $q = 1 - p$. Note that we are using the usual convention that $0! = 1$. 10

29. *Discrete uniform distribution:* Let X be a random variable. The discrete uniform distribution is the probability distribution represented by the probability function

$$f(x) = \frac{1}{N}, \qquad x = 1, 2, \ldots, N. \tag{9}$$

. . . . 10

30. *Joint probability mass function:* The joint probability function $f(x_1, x_2, \ldots, x_n)$ of the random variables $X_1, X_2, \ldots, X_n$ is the probability of the joint occurrence of $X_1 = x_1, X_2 = x_2, \ldots,$ and $X_n = x_n$. Stated differently,

$$f(x_1, x_2, \ldots, x_n) = P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n). \tag{10}$$

. . . . 10

31. *Joint distribution function:* The joint distribution function $F(x_1, x_2, \ldots, x_n)$ of the random variables $X_1, X_2, \ldots, X_n$ is the probability of the joint occurrence of $X_1 \leq x_1, X_2 \leq x_2, \ldots,$ and $X_n \leq x_n$. Stated differently,

$$F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n). \tag{11}$$

. . . . 10

32. *Conditional probability function:* The conditional probability function of X given Y, $f(x|y)$, is

$$f(x|y) = P(X = x|Y = y) = \frac{f(x, y)}{f(y)} \tag{14}$$

. . . . 11

33. *Hypergeometric distribution:* Let X be a random variable. The hypergeometric distribution is the probability distribution represented by the probability function

$$f(x) = P(X = x) = \frac{\binom{A}{x}\binom{B}{k-x}}{\binom{A+B}{k}} \qquad \begin{aligned} &0 \leq x \leq A, \\ &0 \leq k - x \leq B \end{aligned} \tag{17}$$

where A, B and k are nonnegative integers and $k \leq A + B$. 11

34. *Mutually independent:* Let $X_1, X_2, \ldots, X_n$ be random variables with the respective probability functions $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$ and with the joint probability function $f(x_1, x_2, \ldots, x_n)$. Then $X_1, X_2, \ldots, X_n$ are mutually independent if

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1)f_2(x_2)\cdots f_n(x_n) \tag{18}$$

for all combinations of values of $x_1, x_2, \ldots, x_n$. 11

35. *Quantile:* The number x_p for a given value of p between 0 and 1, is called the pth quantile of the random variable X, if $P(X < x_p) \leq p$ and $P(X > x_p) \leq 1 - p$. . . . 12

36. *Expected value:* Let X be a random variable with the probability function $f(x)$ and let $u(X)$ be a real valued function of X. The expected value of $u(X)$, written $E[u(X)]$, is

$$E[u(X)] = \sum_x u(x)f(x). \tag{1}$$

. . . . 13

37. *Mean:* Let X be a random variable with the probability function $f(x)$. The mean of X, usually denoted by μ, is

$$\mu = E(X). \tag{2}$$

. . . . 13

38. *Variance:* Let X be a random variable with mean μ and the probability function $f(x)$. The variance of X, usually denoted by σ^2 or by $\mathrm{Var}(X)$, is

$$\sigma^2 = E[(X - \mu)^2] = E(X^2) - \mu^2. \tag{5}$$

. . . . 14

39. *Expected value:* Let $X_1, X_2, \ldots, X_n$ be random variables with the joint probability function $f(x_1, x_2, \ldots, x_n)$, and let $u(X_1, X_2, \ldots, X_n)$ be a real valued function of $X_1, X_2, \ldots, X_n$. Then the expected value of $u(X_1, X_2, \ldots, X_n)$ is

$$E[u(X_1, X_2, \ldots, X_n)] = \sum u(x_1, x_2, \ldots, x_n)f(x_1, x_2, \ldots, x_n).$$

. . . . 14

40. *Mean of sum of random variables:* Let $X_1, X_2, \ldots, X_n$ be random variables and let

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then

$$E(Y) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

. . . . 15

41. $\sum_{i=a}^{N} i = \frac{(N+a)(N-a+1)}{2}$ and $\sum_{i=1}^{N} i = \frac{(N+1)N}{2}$. 15

42. *Covariance:* Let X_1 and X_2 be two random variables with means μ_1 and μ_2, probability functions $f_1(x_1)$ and $f_2(x_2)$, respectively, and joint probability function $f(x_1, x_2)$.

The covariance of X_1 and X_2 is

$$\begin{aligned}\mathrm{Cov}(X_1, X_2) &= E[(X_1 - \mu_1)(X_2 - \mu_2)] \\ &= E(X_1X_2) - \mu_1\mu_2.\end{aligned}$$

. . . . 15

43. *Correlation coefficient:* The correlation coefficient between two random variables is their covariance divided by the product of their standard deviations. That is, the correlation coefficient, usually denoted by ρ, between two random variables X_1 and X_2 is given by

$$\rho = \frac{\mathrm{Cov}(X_1, X_2)}{\sqrt{\mathrm{Var}(X_1)\mathrm{Var}(X_2)}}$$

. . . . 16

44. *Sum of the first N positive integers:* $\sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6}$. 16

45. *Independence implies zero covariance:* If X_1 and X_2 are independent random variables, the covariance of X_1 and X_2 is zero. 17

46. *Variance of sum of random variables:* Let $X_1, X_2, \ldots, X_n$ be random variables and let

$$Y = X_1 + X_2 + \cdots + X_n.$$

Then

$$\mathrm{Var}(Y) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}(X_i, X_j).$$

. . . . 17

47. *Mean and variance of binomial distribution:* Let X be a random variable with the binomial distribution

$$P(X = k) = \binom{n}{k} p^k q^{n-k}.$$

Then the mean and variance of X are given by

$$\begin{aligned}E(X) &= np \\ \mathrm{Var}(X) &= npq.\end{aligned}$$

. . . . 18

48. Let X be the sum of n integers selected at random, without replacement, from the first N integers 1 to N. Then the mean and variance of X are given by

$$\begin{aligned}E(X) &= \frac{n(N+1)}{2} \\ \mathrm{Var}(X) &= \frac{n(N+1)(N-n)}{12}.\end{aligned}$$

. . . . 18

49. *Discrete random variable:* A random variable X is *discrete* if there exists a one to one correspondence between the possible values of X and some or all of the positive integers. 19

50. *Continuous random variable:* A random variable X is continuous if no two quantiles x_{p_1} and x_{p_2} of X are equal to each other, where p_1 is not equal to p_2. Equivalently, a random variable X is continuous if $P(X \leq x)$ equals $P(X < x)$ for all numbers x.20

51. *Normal distribution:* Let X be a random variable. Then X is said to have the normal distribution if the distribution of X is given by

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\,\mathrm{d}y.$$

. 21

52. *Normalization:* For a given value of p, let x_p be the pth quantile of a normal random variable with mean μ and variance σ^2, and let z_p be the pth quantile of a standard normal random variable. The quantile x_p may be obtained from z_p by using the relationship

$$x_p = \mu + \sigma z_p.$$

Similarly,

$$z_p = \frac{x_p - \mu}{\sigma}$$

. 21

53. *Central Limit Theorem:* Let Y_n be the sum of the n random variables $X_1, X_2, \ldots, X_n$, let μ_n be the mean of Y_n and let σ^2 be the variance of Y_n. As n, the number of random variables, goes to infinity, the distribution function of the random variable

$$\frac{Y_n - \mu_n}{\sigma_n}$$

approaches the standard normal distribution function. 22

54. *Chi-squared distribution:* A random variable X has the chi-squared distribution with k degrees of freedom if the distribution function of X is given by

$$\begin{aligned} F(x) &= P(X \leq x) \\ &= \begin{cases} \int_0^x \frac{y^{(k/2)-1}e^{-y/2}}{2^{k/2}\Gamma(k/2)}\,\mathrm{d}y & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases} \end{aligned}$$

$E(X) = k$ and $\mathrm{Var}(X) = 2k$. 23

55. *Distribution of sum of squares of independent standard normal random variables:* Let $X_1, X_2, \ldots, X_k$ be k independent and identically distributed standard normal random variables. Let Y be the sum of the squares of the X_i.

$$Y = X_1^2 + X_2^2 + \cdots + X_k^2.$$

Then Y has the *chi-squared distribution* with k degrees of freedom. 23

56. *Distribution of sum of independent chi-squared random variables:* Let $X_1, X_2, \ldots, X_n$ be independent chi-squared random variables with $k_1, k_2, \ldots, k_n$ degrees of freedom, respectively. Let Y equal the sum of the X_i. Then Y is a chi-squared random variable with k degrees of freedom, where

$$k = k_1 + k_2 + \cdots + k_n.$$

Chapter 2

Statistical Inference

Contents

Preliminary remarks

- The concepts of probability theory introduced in the previous chapter do not cover the entire field of probability theory.
- Bridge the gap between probability theory and its application to data analysis.
- *Statistic:* Concepts of the basic science for data analysis.
- The field of statistics owes many of its significant ideas to people in the applied sciences who had difficult questions concerning their data.

2.1 Populations, samples, and statistics

- Much of our knowledge concerning the world we live in is the result of samples.
- E.g. We eat at a restaurant once and we form an opinion concerning the quality of the food and the service at that restaurant.
- We know 12 people from England and we feel we know the English people.
- Quite often the opinions we form from the sample are not accurate.
- However, samples that are obtained according to scientifically derived methods can give very accurate information about the entire population.
- The process of forming scientific opinions is often placed within the framework of an *experiment*.

- An experiment is the process of following a well-defined procedure, where the outcome of following the procedure is not known prior to the experiment.

- The collection of all elements under investigation is called the *population*.

- A *sample* is a collection of some elements of a population.

  - *Convenient sample*: A collection of the elements that are easiest to obtain, such as " citizen on the street " interviews, or TV call-in surveys.
  - Probability sample: Allows accurate statements to be made about the unknown population parameters.
  - Random sample: Later

- The population about which information is wanted is called the *target population*.

- The population to be sampled is called the *sampled population*.

- Random sample of size n from a population with N elements with (without) replacement.

Definition 2.1.1 *A sample from a finite population is a random sample if each of the possible samples was equally likely to be obtained.*

- $\binom{N}{n}$ possible samples of size n for without replacement.

- N^n possible samples of size n for replacement.

Definition 2.1.2 *A random sample of size n is a sequence of n independent and identically distributed random variables* $X_1, X_2, \dots, X_n$.

Definitions 1 and 2 are identical only if the drawing in Definition 1 is with replacement, for then and only then are the observations independent.
Multivariate random variable k-variate

$$X_i = (Y_{i1}, Y_{i2}, \dots, Y_{ik})$$

- X_is are independent and identically distributed.

- Y_{ij} random variables within each X_i may or may not be independent and/ or identically distributed.

- Special case: Bivariate random variate

  - Y_{i1}: Number of dreams in the ith night.
  - Y_{i2}: Length of sleep in the ith night.
  - $X_i = (Y_{i1}, Y_{i2})$
  - X_i and X_j are independent.

Example 2.1.1 *A psychologist would like to obtain four subjects for individual training and examination. He advertises and 20 volunteers respond. He has several ways of selecting a sample of 4 from his sampled population of size 20.*

- Two ways to select 4 people.
- $\binom{20}{4} = 4845$ pieces of paper that are identical and writes 4 names on each piece of paper.
- Another way of obtaining a random sample: Write each of the names on a slip of paper, 20 slips in all, and one by one draw 4 slips in some random manner. □

1. *Measurement scale*: An excellent paper by Stevens (1946).

   - *Nominal scale*: Use numbers merely as a means of separating the properties of elements into different classes or categories.

     The number assigned to the observation serves only as a "name" for the category to which the observation belongs, hence the title "nominal."
     - Toss a coin: 0 for tail and 1 for head.
     - Blue, yellow, red.
     - A, B, C.
   - *Ordinal scale*: Refer to measurements where only the comparisons "greater", "less", or "equal" between measurements are relevant.
     - Assign the number 1 to the most preferred of three brands, number 3 to the least preferred.
     - It is this ability to order the elements, on the basis of the relative size of their measurements, that gives the name of the ordinal scale.
   - *Interval scale*: Consider as pertinent information not only the relative order of the measurements as in the ordinal scale but also the size of the interval between measurements.
     - Scale used in temperature.
     - The actual numerical value of the temperature is merely a comparison with an arbitrary point called "zero degree".
     - The interval scale of measurement requires a zero point as well as a unit distance.
     - Two scales: Fahrenheit (F) and Celsius (C)

$$F = \frac{9}{5}C + 32$$

   - *Ratio scale*: Not only the order and interval size are important, but also the ratio between two measurements is meaningful.
     - Crop yields, distances, weights, heights, income.

   - Most of the usual parametric statistical methods require an *interval (or stronger) scale* of measurement.
   - Most nonparametric methods assume either the *nominal scale or the ordinal scale* to be appropriate.
   - Statistical methods requiring only a weaker scale may be used with the stronger scales also.

Statistics

- The word statistic originally referred to numbers published by the state, where the numbers were the result of a summarization of data collected by the government.
- Extend our idea of a statistic from being only a number to being a rule for finding the number.
- A statistic also conveys the idea of a summarization of data, so usually a statistic is considered to be a random variable that is a function of several other random variables.

Definition 2.1.3 *A statistic is a function which assigns real numbers to the points of a sample space, where the points of the sample space are possible values of some multivariate random variable. In other words, a statistic is a function of several random variables.*

Example 2.1.2 *Let $X_1, X_2, \ldots, X_n$ represent test scores of n students. Then each X_i is a random variable. Let W equal the average of these test scores.*

- $W = \frac{X_1+X_2+\cdots+X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- W is a statistic.
- If $X_1 = 76, X_2 = 84$, and $X_3 = 85$, then $W = 81\frac{2}{3}$.
- *Sample mean*: W □

Ordered observation

$$x^{(1)} \le x^{(2)} \le \cdots \le x^{(n)}$$

Definition 2.1.4 *The order statistic of rank k, $X^{(k)}$, is the statistic that takes as its value the kth smallest element $x^{(k)}$ in each observation $(x_1, x_2, \ldots, x_n)$ of $(X_1, X_2, \ldots, X_n\ )$.*

If $(X_1, X_2, \ldots, X_n)$ is a random sample, sometimes $(X^{(1)}, X^{(2)}, \ldots, X^{(n)})$ is called the *ordered random sample*.

2.2 Estimation

- One of the primary purposes of a statistic is to estimate unknown properties of the population.
- The estimate is based on a sample.
- The estimate is an educated guess concerning some unknown property of the probability distribution of a random variable.
- For example, we might use the proportion of defective items in a sample of transistors as an estimate the unknown proportion of defective transistors in some population of transistors.
- A statistic that is used to estimate is called an *estimator*.
- Sample mean, sample variance and sample quantiles.

The true distribution function $F(x)$ of a random variable is almost never known. Use the empirical function $S(x)$ to estimate $F(x)$.

Definition 2.2.1 *Let $X_1, X_2, \dots, X_n$ be a random sample. The empirical distribution function $S(x)$ is a function of x, which equals the fraction of X_is that are less than or equal to x, $-\infty < x < \infty$.*

$$S(x) = \frac{\text{number of } X_i\text{'s} \leq x}{\text{number of } X_i\text{'s}}$$

Example 2.2.1 *In a physical fitness study five boys were selected at random. They were asked to run a mile, and the time it took each of them to run the mile was recorded. The times were $6.23, 5.58, 7.06, 6.42, 5.20$. The empirical distribution function $S(x)$ is represented graphically in Figure 1.*

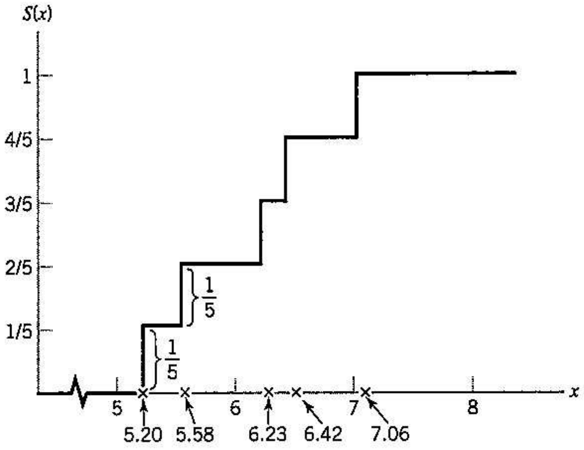


Fig. 1

□

- $S(x)$ is a random function.
- The empirical function is always a step function with jumps of height $1/n$ at each of the n numbers $x_1, x_2, \dots, x_n$.

Example 2.2.2 *The random variable, which has a distribution function identical to the function $S(x)$ of Example 1, is the random variable X with the following probability distribution.*

$$P(X = 5.20) = .2, \quad P(X = 5.58) = .2, \quad P(X = 6.23) = .2,$$
$$P(X = 6.42) = .2, \quad P(X = 7.06) = .2$$

- $E[X] = \sum_x xf(x) = 6.098$
- $\text{Var}(X) = \sum_x (x - E[X])^2 f(x) = .424$ □

Estimators

Sample (population) mean, variance, quantiles

Definition 2.2.2 *Let $X_1, X_2, \dots, X_n$ be a random sample. The pth sample quantile is that number Q_p satisfies the two conditions:*

1. *The fraction of the X_is that are less than Q_p is $\leq p$.*
2. *The fraction of the X_is that exceed Q_p is $\leq 1 - p$.*

$$Q_p = \begin{cases} \frac{X_{(np)}+X_{(np+1)}}{2} & \text{if } np \text{ is an integer,} \\ X_{(\lceil np \rceil)} & \text{otherwise.} \end{cases}$$

Example 2.2.3 *Six married women were selected at random and the number of children belonging to each was recorded.*

$$0, 2, 1, 2, 3, 4$$

- Empirical distribution function given in Figure 2.
- $Q_{.5} = 2,\ Q_{.25} = 1,\ Q_{.75} = 3$
- $Q_{1/3} = \frac{x_{(2)}+x_{(3)}}{2} = 1.5$

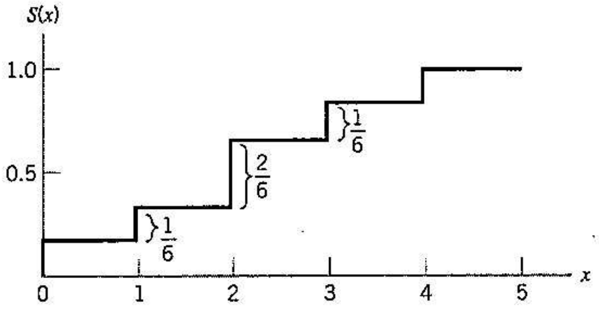


Fig. 2

□

Definition 2.2.3 *Let $X_1, X_2, \ldots, X_n$ be a random sample. The sample mean $\bar{X}$ is defined by*

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

The sample variance S^2 is

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2.$$

Sample standard deviation S is the square root of the sample variance.

Example 2.2.4 *In the random sample 0, 2, 1, 2, 3, 4 of Example 3.*

- $\bar{X} = \frac{1}{6}(0 + 2 + 1 + 2 + 3 + 4) = 2$
- $S^2 = \frac{1}{6}(2^2 + 0 + 1^2 + 0 + 1^2 + 2^2) = 1\frac{2}{3}$
- Estimate of the unknown mean is 2 and estimate of the unknown variance is 5/3. □

Point estimation (above) v.s. interval estimation

- We are 95% confident that the unknown mean lies between 1.3 and 2.7.
- An *interval estimator* consists of two statistics, one for each end of the interval, and the *confidence coefficient*, which is the probability that the interval estimator will contain the unknown population quantity.

- To make a point estimate we need only to think of a number.
- Criteria for comparing point estimators may be found in any introductory text in probability or statistics.

Definition 2.2.4 *An estimator $\hat{\theta}$ is an unbiased estimator of a population parameter θ if $E\left[\hat{\theta}\right] = \theta$.*

Theorem 2.2.1 *Let $X_1, X_2, \ldots, X_n$ be independent random variables from a population with mean μ and variance σ^2. Then $E\left[\bar{X}\right] = \mu$ and $Var(\bar{X}) = \sigma^2/n$.*

Proof.

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum X_i\right] = \frac{1}{n}\sum E[X_i] = \mu$$

$$\mathrm{Var}\left(\bar{X}\right) = \mathrm{Var}\left(\frac{1}{n}\sum X_i\right) = \frac{1}{n^2}\sum \mathrm{Var}(X_i) = \frac{\sigma^2}{n} \qquad \square$$

The standard deviation of an estimator is called its *standard error*. For example, standard error of $\bar{X}$: $\mathrm{std}(\bar{X}) = \sigma/\sqrt{n}$
S^2 is not an unbiased estimator for σ^2,
Unbiased estimator for σ^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Approximate $1-\alpha$ confidence interval

- Central Limit Theorem says if $X_1, X_2, \ldots, X_n$ are independent random variables, each with mean μ and variance σ^2, then

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \approx N(0,1).$$

- $$P\left(\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha$$
- For most practical purposes sample sizes larger than 30 are considered "large enough".

Example 2.2.5 *Large litters of pig mean more profit for a hog farmer. A state experiment station is studying a new method of raising hogs that may result in large litter sizes. In 55 litters the average number of surviving pigs was 9.8, with $s = 1.4$.*

- A 95% confidence interval:

$$9.8 \pm 1.96\frac{1.4}{\sqrt{55}} = \bar{X} - z_{1-\alpha/2}\frac{s}{\sqrt{n}} = (9.43, 10.17) \qquad \square$$

Bootstrap

- Many statistics of estimator used are difficult to derive. The bootstrap is a way to simulate these statistics.

- The *bootstrap method* samples n values with replacement from the observations in the original random sample of size n.
- Everything in the bootstrap procedure depends on the original sample values.
- For simply estimating the mean and the standard error of an estimator, the number of bootstrap replications seldom needs to be more 100 or 200, and as few as 25 replications can be very informative.
- Larger numbers of replications are needed to find confidence intervals.
- One method of obtaining an approximate confidence interval for θ is to use the $\alpha/2$ and $1-\alpha/2$ sample quantiles from the bootstrap sample estimator $\hat{\theta}^*$.
- For approximate $1-\alpha$ confidence interval Efron and Tibshirane (1986) recommend a minimum of 250 bootstrap replications.

Example 2.2.6 *In Example 5 we found an approximate 95% confidence interval for the mean number of pigs per litter using the central limit theorem that applies to $\bar{X}$. Now use the bootstrap method to find an approximate 95% confidence interval for σ.*

-

Obs.	1	2	3	4	...	55
Litter size	9	9	8	6	...	11

$s = 1.4$ (original sample)

-

Obs.	4	17	4	28	...	16
Litter size	6	9	6	10	...	9

$s_1^* = 1.6$ (Bootstrap sample #1)

-

Obs.	28	23	3	16	...	39
Litter size	10	10	8	9	...	8

$s_2^* = 1.8$ (Bootstrap sample #2)

- And so on, to

-

Obs.	6	1	55	14	...	17
Litter size	10	9	11	11	...	9

$s_{250}^* = 1.1$ (Bootstrap sample #250)

- Approximate 95% confidence interval:

$$(s^{*(\lceil .025(250)\rceil)}, s^{*(\lceil .975(250)\rceil)}) = (s^{*(7)}, s^{*(244)}) = (1.0, 2.0)$$

- The expected value of s is estimated by computing on the 250 values of s^*. □

Daison and Hinkley (1997, resampling) contains a library of routines for use with *S-Plus*.

Parameter estimation in general
Two main questions in estimating an unknown parameter θ:

1. What estimator should be used? *empirical distribution*
   - $\hat{\mu} = \bar{X}, \hat{\sigma} = S$ and the quantile estimator $\hat{x}_p = Q_p$.

2. How good is the estimator? *standard error*
   For example, $\text{std}(\bar{X}) = \sigma/\sqrt{n}$

Theory

- $S(x) \overset{n\to\infty}{\longrightarrow} F(x)$ in probability.
- In most cases of interest those estimators will approach the parameters they are estimating.
- For an introduction to the bootstrap concept see Efron and Tibshirani (1986).

Survival function:

$$P(x) = 1 - F(x)$$

- Useful in life testing, medical follow-up, and other fields.
- Natural estimator: The empirical survival function

  $$\hat{P}(x) = 1 - S(x)$$

  is the relative frequency of the sample $X_1, X_2, \ldots, X_n$ that exceeds x in value.

The Kaplan-Meier estimator (1958)

- X: time to death
- X may be unobservable in some cases because of the loss of the item from the experiment (subjects moving away, subjects entering the experiment late, the experiment ending before all subjects die, etc.)..
- Kaplan and Meier (1958) provide a method for using some information from the lost data, namely that death did not occurred before the loss.
- They use the fact that if death occurs after time x, then the death also occurred after all times prior to x.
- $P(X > x_1) = P(X > x_1, X > x_0) = P(X > x_1 | X > x_0)P(X > x_0)$
- Suppose that 100 items are put on test at the beginning of Year 1, and at the end of Year 1 only 30 survive.
- Then at the beginning of Year 2 suppose an additional 1000 items are put on test.
- At the end of Year 2, 250 of the 1000 items survive, and 10 of the 30 items survive from the original 100 items.
  - $\hat{P}(1) = \hat{P}(X > 1) = 30/100$
  - $\hat{P}(2) = \hat{P}(X > 2) = 10/100$ (use only the original 100 items)
  - Since a total of 1100 items have been on test one year, with a total of $250+30 = 280$ survivors, an improved estimate of $P(1)$ would be

    $$\hat{P}(1) = \hat{P}(X > 1) = 280/1100 \qquad \text{(use both sets of data)}$$

  - Unfortunately we are unable to improve our estimate of $P(X > 2/X > 1)$ because we don't how what happens to the 1000 observations in the second year of the test. So we use the estimator

    $$\hat{P}(X > 2|X > 1) = 10/30.$$

  - Use this improved estimate of $P(1)$ to get an improved estimate of $P(2)$, one which uses the fact that 280 of 1100 items survived past 1 year

    $$\hat{P}(2) = \hat{P}(X > 2|X > 1)\hat{P}(X > 1) = (10/30)(280/1100) = 0.085.$$

- Let $u_1 < u_2 < \cdots < u_k$ represent k " lifetimes ".

- $p_i = P(X > u_i|X > u_{i-1})$

- $\hat{p}_i = \frac{\text{Number of items known to survive past time } u_i}{\text{Number of items known to survive past time } u_{i-1}}$

- *Kaplan-Meier estimator* of $P(x)$:

  $$\hat{P}(x) = \begin{cases} 1 & \text{for } x < u_1 \\ \prod_{u_i \le x} \hat{p}_i & \text{for } x \ge u_1 \end{cases}$$

- This estimator is a decreasing step function that takes steps only at observed deaths.

- $S(x) = 1 - \hat{P}(x)$

Computer assistance: *Splus*

Example 2.2.7 *Ten fanbelts are tested by placing them on cars and records are kept of the mileage on each car when the fanbelt breaks. At the end of the test five fanbelts have broken, with lifetimes of 77, 47, 81, 56, and 80. The other five fanbelts are still unbroken, and the mileage on those cars are 62, 60, 43, 71, and 37. The Kaplan-Meier estimate of the survival function is found as follows.*

i	u_i	result	$\hat{p}_i$	$\hat{P}(u_i)$
1	37	lost	10/10	1
2	43	lost	9/9	1
3	47	death	7/8	0.875
4	56	death	6/7	0.75
5	60	lost	6/6	0.75
6	62	lost	5/5	0.75
7	71	lost	4/4	0.75
8	77	death	2/3	0.5
9	80	death	1/2	0.25
10	81	death	0/1	0

Graph of $\hat{P}(x)$ in Fig. 3.

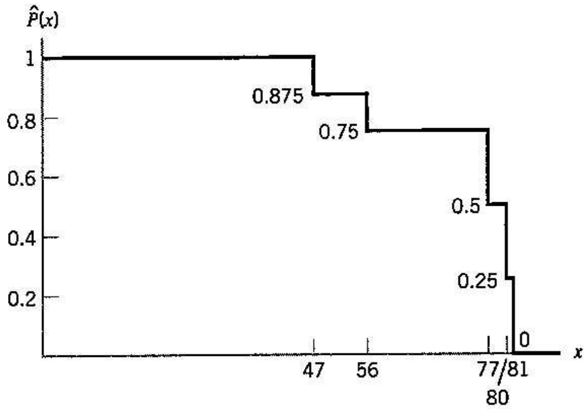


Fig. 3

□

- The Kaplan-Meier is the same as $1 - S(x)$ if there are no loses.
- There are losses as well as deaths, $\hat{P}(x)$ starts at 1.0, but the decreasing steps may no longer be of uniform height.
- If there is a loss after the final known death, $\hat{P}(x)$ will not decrease to zero, and is not defined for x beyond the final known loss.
- In this case $S(x)$ is not suitable for the estimation of some parameters associated with $F(x)$ such as the mean and the variance using the usual methods described earlier in this section, but may be used to estimate some quantiles.

Point estimation is always a nonparametric statistical method. It is more difficult to tell whether the methods of forming confidence intervals are parametric or nonparametric.

- If no knowledge of the form of the distribution function is required in order to find a confidence interval that method is nonparametric.
- If the method requires that the unknown distribution function be a normal distribution function or some specified form the method is parametric.

2.3 Hypothesis testing

Hypothesis testing is the process of inferring from a sample whether or not a given statement about the population appears to be true.

1. Women are more likely than man to have automobile accidents.
2. Nursery school helps a child achieve better marks in elementary school.
3. The defendant is guilty.
4. Toothpaste A is more effective in preventing cavities than toothpaste B.

Outline of the steps involved in a test

1. *Alternative (research) hypothesis*: The statement that the experimenter would like to prove.

- The new product is better than the old product.
- This medication is effective in curing the illness.

2. *Null hypothesis*: The negation of the alternative hypothesis.

   - The new product is no better than the old product.
   - This medication is not effective in curing the illness.

   If the data in the sample strongly disagree with the null hypothesis, the null hypothesis is rejected. If the data in the sample do not conflict with the null hypothesis, or if there are insufficient data to show a conflict with the null hypothesis, the experimenter "*fails to reject*" the null hypothesis.

3. *Test statistic*: A good test statistic is a sensitive indicator of whether the data agree or disagree with the null hypothesis.

4. *Decision rule*: Decide whether to accept or reject the null hypothesis.

5. On the basis of a random sample from the population, the test statistic is evaluated, and a decision is made to accept or reject the null hypothesis.

Example 2.3.1 *A certain machine manufactures parts. The machine is considered to be operating properly if 5% or less of the manufactured parts are defective. If more than 5% of the parts are defective the machine needs remedial attention.*

- *Null hypothesis*:
$$H_0 : \text{The machine is operating properly}$$
- *Alternative hypothesis*:
$$H_1 : \text{The machine needs attention}$$
- $H_0 : p \leq .05$ and $H_1 : p > .05$
- T: Total number of defective items.
- $T \sim \text{Binomial}(10, p)$
- Under H_0, $P(T \leq 2) \geq 0.9885$
- Critical region: $T > 2$
- Suppose a random sample consisting of 10 machined parts is observed and 4 of the parts are found to be defective.
- Then $T = 4$ and the null hypothesis is rejected. We conclude that the machine needs attention. □

Definition 2.3.1 *The hypothesis is simple if the assumption that the hypothesis is true leads to only one probability function defined on the sample space. The hypothesis is composite if the assumption that the hypothesis is true leads to only two or more probability functions defined on the sample space.*

Definition 2.3.2 *A test statistic is a statistic used to make the decision in a hypothesis test.*

- *Upper-tailed test*: The rejection region corresponds to the largest values of the test statistic.
- *Lower-tailed test*: The rejection region corresponds to the smallest values of the test statistic.
- *One-tailed test*: Previous two cases.
- *Two-tailed test*: If the test statistic is selected so that the largest values of the test statistic and the smallest values of the test statistic.

Definition 2.3.3 *The critical (rejection) region is the set of all points in the sample space that result in the decision to reject the null hypothesis.*

Definition 2.3.4 *The type I error is the error of rejecting a true null hypothesis.*

Definition 2.3.5 *The type II error is the error of accepting a false null hypothesis.*

Definition 2.3.6 *The level of significance, or α, is the maximum probability of rejecting a true null hypothesis.*

Definition 2.3.7 *The null distribution of the test statistic is its probability distribution when the null hypothesis is assumed to be true.*

Definition 2.3.8 *The power, denoted by $1-\beta$, is the probability of rejecting a false null hypothesis.*

		The decision	
		Accept H_0	Reject H_0
The true situation	H_0 is true	Correct decision $1-\alpha$	Type I error α
	H_0 is false	Type II error β	Correct decision $1-\beta$

Definition 2.3.9 *The p-value is the smallest significance level at which the null hypothesis would be rejected for the given observation.*

- Let t_{obs} represent the observed value of a test statistic T.
- In an upper-tailed test the p-value is $P(T \geq t_{\text{obs}})$ computed using the null distribution of T.
- In a lower-tailed test the p-value is $P(T \leq t_{\text{obs}})$.
- In a two-tailed test, the p-value can be stated as twice the smaller of the one-tailed p-values.

Example 2.3.2 *In order to see if children with nursery school experience perform differently academically than children without nursery school experience, 12 third-grade students are selected, 4 of whom attended nursery school. The hypothesis to be tested is*
H_0: The academic performance of third-grade children does not depend on whether or not they attended nursery school
H_1: There is a dependence between academic performance and attendance at nursery school.

- Assume that the 12 children are a random sample of all third-grade children, and also that the children can be ranked from 1 to 12 (best to worst) academically.

H_0:	The ranks of the four children with nursery school experience are a random sample of ranks from 1 to 12
H_1:	The ranks of the four children with nursery school experience tend to be higher or lower as a group than a random sample of 4 ranks out of 12

- T: Sum of the ranks of the 4 children who attended nursery school.
- $\binom{12}{4} = 495$ points in the sample space.
- Observed ranks: $2, 5, 6$, and 9 ($T = 22$).
- Under H_0: $E[T] = 26$ and $\text{Var}(T) = 34.67$
- $P(T \leq 22) \approx P\left(Z \leq \frac{22-26}{5.888}\right) = 0.248$
- Above procedure is called the Mann-Whitney test or the Wilcoxon test and will be discussed extensively in Chap. 5 along with its many variations.
- The data in Example 2 have the ordinal scale of measurement.
- Example 1 illustrated the analysis of nominal type data. □

Computer assistance

- Most statistics computer packages perform hypothesis tests.
- In some packages, the user specifies the null hypothesis and the alternative hypothesis, and the package returns the correct p-value.
- In other packages, the computer always returns a p-value for a two-sided test and the user must decide if that is the desired p-value, or if it must be halved to obtain a one-tailed p-value.
- If the p-value is less than or equal to the desired level of significance, which is selected by the user, then the null hypothesis is rejected.
- Many computer packages use approximate methods for finding p-values.
- More and more computer packages are following the example of StatXacf, which computes exact p-values or uses monte carlo simulation to approximate the exact p-value when the exact p-values are impractical to obtain.

2.4 Some properties of hypothesis tests

Once the hypotheses are formulated, there are usually several hypothesis tests available for testing the null hypotheses.
Select one test:

- Are these assumptions of this test valid assumptions in my experiment?
- If the answer is "No," that test probably should be discarded.
- For example, in most parametric tests one of the stated assumptions is that the random variable being examined has a normal distribution.

The use of a test in a situation where the assumptions of the test are not valid is dangerous for two reasons.

- The data may result in rejection of the null hypothesis, not because the data indicate that the null hypothesis is false, but because the data indicate that one of the assumptions of the test is invalid.
- Sometimes the data indicate strongly that the null hypothesis is false, and a false assumption in the model is also affecting the data, but these two effects neutralize each other.

Properties of a good test:

1. The test should be *unbiased*.
2. The test should be *consistent*.
3. The test should be more *efficient* in some sense than the other tests.

Power function
If H_1 is composite, the power may vary as the probability function varies. If H_1 is stated in terms of some unknown parameter, the power usually may be given as a function of that parameter.

Example 2.4.1 (Calculate power function) *In Example 2.3.1 the critical region consisted of all points with more than two defectives in the 10 items examined.*

- $P(\text{reject } H_0) = \sum_{i=3}^{10} \binom{10}{i} p^i (1-p)^{10-i} = 1 - \sum_{i=0}^{2} \binom{10}{i} p^i (1-p)^{10-i}$

p	$P(\text{reject } H_0)$	p	$P(\text{reject } H_0)$
0	0.0000	0.50	0.9453
0.05	0.0115	0.55	0.9726
0.10	0.0702	0.60	0.9877
0.15	0.1798	0.65	0.9952
0.20	0.3222	0.70	0.9984
0.25	0.4744	0.75	0.9996
0.30	0.6172	0.80	0.9999
0.35	0.7384	0.85	1.0000
0.40	0.8327	0.90	1.0000
0.45	0.9004	1.00	1.0000

- Under H_0, $0 < p < 0.05$ (Fig. 4)
- The power ranges from 0.0115 for p close to 0.05 to 1.0000 for p equal to 1.0.

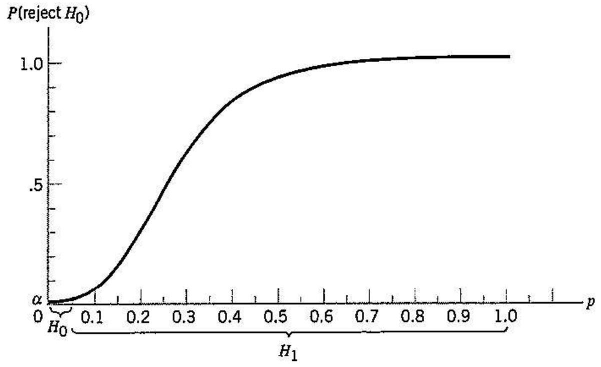


Fig. 4

□

Two tests may be compared on the basis of their power functions. This basis of comparison is discussed again later in this section when relative efficiency is defined.
Computer assistance

- The power of a test is a function of the level of significance, the simple alternative hypothesis of interest, and the sample size.
- *SAS* concentrates on computing power of a test, given the level of significance, the range of alternatives of interest, and the sample size.
- It can also compute the sample size required for a given power.

Definition 2.4.1 *An unbiased test is a test in which the probability of rejecting H_0 when H_0 is false always greater than or equal to the probability of rejecting H_0 when H_0 is true.*

- An unbiased test is one where the power is always at least as large as the level of significance.
- A test that is not unbiased is called a *biased test*.

Definition 2.4.2 *A sequence of tests is consistent against all alternatives in the class H_1 if the power of the tests approaches 1.0 as the sample size approaches infinity, for each fixed alternative possible under H_1. The level of significance of each test in the sequence is assumed to be as close as possible to but not exceeding some constant $\alpha > 0$.*

Example 2.4.2 *We wish to determine whether human births tend to produce more babies of one sex, instead of both sexes being equally likely. We are testing*

H_0: *A human birth is equally likely to be male or female, ($p = 1/2$)*
H_1: *Male births are either more likely, or less likely, to occur than female births ($p \neq 1/2$)*

- The sampled population consists of births registered in a particular country.
- The sample consists of the last n births registered, for some selected value of n.

- It is also assumed that the probability p (say) of a male birth remains constant from birth to birth and that the births are mutually independent as far as the events "male" and "female" go.

- Then the hypotheses are equivalent to the following

$$H_0: \ p = 1/2 \quad \text{v.s.} \quad H_1: \ p \neq 1/2.$$

- T: Number of male births.

- Critical region: Symmetrically to the largest values and smallest values of T, of the largest size not exceeding 0.05.

- We have described an entire sequence of tests, one for each value of the sample size.

- Each test is two tailed and has a level of significance of 0.05 or smaller.

- $T \sim \text{Binomial}(n, p)$

- Critical region are given by Dixon (1953):

	Values of T corresponding			
n	to the critical region			α
5		None		0
6	$T = 0$	and	$T = 6$	0.03125
8	$T = 0$	and	$T = 8$	0.00781
10	$T \leq 1$	and	$T \geq 9$	0.02148
15	$T \leq 3$	and	$T \geq 12$	0.03516
20	$T \leq 5$	and	$T \geq 15$	0.04139
30	$T \leq 9$	and	$T \geq 21$	0.04277
60	$T \leq 21$	and	$T \geq 39$	0.02734
100	$T \leq 39$	and	$T \geq 61$	0.03520

- For $n \leq 20$ these same values can be obtained from Table A3. For $n > 20$ the normal approximation could be used.

- A comparison of several power functions (Fig. 5).

- We can see that as the sample size increases, the power at each fixed value of p (except $p = 0.5$) increases toward 1.0.

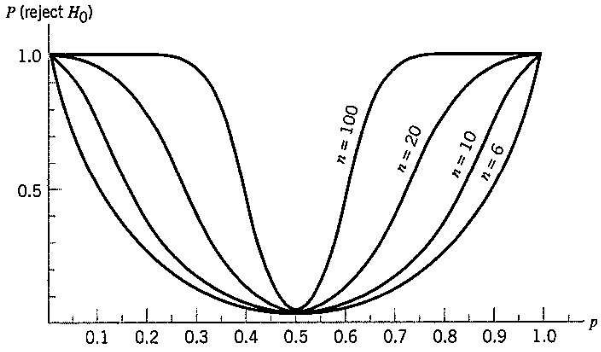


Fig. 5 A comparison of several power functions.

□

This example merely demonstrates the idea behind the term consistent as it applies to a sequence of tests.
This demonstration is not a proof that the sequence of tests is consistent.
A rigorous proof of consistency usually requires more mathematics than we care to use in this introductory book.
Relative efficiency

- Efficiency is a relative term and is used to compare the sample size of one test with that of another test under similar conditions.
- Suppose two tests may be used to test a particular H_0 against a particular H_1.
- Also suppose that the two tests have the same α or and the same β and therefore are "comparable" with respect to level of significance and power.
- Then the test requiring the smaller sample size is preferred over the other test, because a smaller sample size means less cost and effort is required in the experiment.
- The test with the smaller sample sue is said to be *more efficient* than the other test, and its *relative efficiency* is greater than one.

Definition 2.4.3 *Let T_1 and T_2 represent two tests that test the same H_0 against the same H_1, with the critical regions of the same size α and with the same value of β. The relative efficiency of T_1 to T_2 is the ratio n_2/n_1, where n_1 and n_2 are the sample sizes of the tests T_1 and T_2, respectively.*

If the alternative hypothesis is composite, the relative efficiency may be computed for each probability function defined by the alternative hypothesis.

Example 2.4.3 *Two tests are available for testing the same H_0 against the same H_1. Both tests have $\alpha = 0.01$ and $\beta = 0.14$.*

- $n_1 = 75$ and $n_2 = 50$
- Relative efficiency of the first test to the second test $50/75 = 0.67$.
- Relative efficiency of the second test to the first test $75/50 = 1.5$.
- If the efficiency of the first test relative to the second test at $\alpha = 0.05, \beta = 0.30, n_1 = 40$ is 0.75, then the sample size required by the second test: $0.75 = n_2/40 \Rightarrow n_2 = 30$.
□

Asymptotic relative efficiency

- The relative efficiency depends on the choice of α, the choice of β, and the particular alternative being considered if H_1 is composite.
- In order to provide an overall comparison of one test with another it is clear that relative efficiency leaves much to be desired.
- We would prefer a comparison that does not depend on our choice of α, β, or a particular alternative possible under H_0 if H_1 is composite, which it usually is.

Definition 2.4.4 *Let n_1 and n_2 be the sample sizes required for two tests T_1 and T_2 to have the same power under the same level of significance. If α and β remain fixed the limit of n_2/n_1, as n_1 approaches infinity, is called the asymptotic relative efficiency (A.R.E) if the first test to the second test, if that limit is independent of α and β.*

- In our quest to select the test with greatest power, we usually are forced to select the test with the greatest A.R.E. because the power depends on too many factors.
- The A.R.E. of two tests is usually difficult to calculate (Noether, 1967).
- Studies of the exact relative efficiency for very small sample sizes show that A.R.E. provides a good approximation to the relative efficiency in many situations of practical interest.
- The A.R.E. often provides a compact summary of the relative efficiency between two tests.

Definition 2.4.5 *A test is conservative if the actual level of significance is smaller than the stated level of significance.*

- At times it is difficult to compute the exact level of significance of a test, and then some methods of approximating α are used.
- The approximate value is then reported as being the level of significance.
- If the approximate level of significance is larger than the true (but unknown) level of significance, the test is conservative, and we know the risk of making a type I error is not as great as it is stated to be.

2.5 Some comments on nonparametric statistics

- We will attempt to distinguish between the terms parametric statistics and nonparametric statistics, although the distinction is not always clear even in the minds of professional statisticians.
- We will use the term nonparametric and the more descriptive term distribution-free interchangeably even though some statisticians distinguish between the two.
- Also we will provide some guidance for when to use nonparametric methods in the analysis of data, and when parametric methods should be preferred.

Nonparametric v.s. parametric statistic

- Good methods to use:
  - First discuss *hypothesis testing* and *confidence intervals*.
  - A test of hypothesis relies on a good test statistic, one that is sensitive to the difference between the null hypothesis and the alternative hypothesis, and one whose probability distribution under the null hypothesis is known.

  - ▷ A confidence interval is the *inversion of a hypothesis test* in that the confidence interval is the collection of the null hypotheses that are not rejected by the data, so a good (powerful) hypothesis test relates to a good (short) confidence interval.
  - ▷ The sample mean $\bar{X}$ is a good test statistic for testing hypotheses about the population mean μ because it is sensitive to differences in the population mean.
  - ▷ S and s are good statistics to use for inferences about the population standard deviation σ.
  - ▷ However, the probability distributions of $\bar{X}, S$ and s depend on the population probability distribution of Xs, which is usually unknown.

- ▶ *Parametric methods*
  - ▷ If the population probability distribution is normal distribution then $\bar{X}$ is normal.
  - ▷ Any hypothesis test or confidence interval that is based on the assumption that the population distribution function is known, or known except for some unknown parameters, is called a *parametric method*.
  - ▷ Most parametric methods are based on the normality assumption because the theory behind the test can be worked out with the normal population distribution.
  - ▷ The resulting procedures are efficient and powerful procedures for normally distributed data.
  - ▷ Other parametric procedures have been developed by assuming the population has other distributions, such as the exponential, Weibull, and so on.

- ▶ *Robust methods*
  - ▷ No population has exactly a normal distribution, or any other known distribution.
  - ▷ If the population distribution is approximately normal, then usually (but not always) it is safe to use a method based on the normal distribution.
  - ▷ However, if the data appear to come from a distinctly nonnormal distribution, or a distribution not covered by the parametric methods, then a nonparametric method should be considered.
  - ▷ A method of analysis that is approximately valid even if one of the assumptions behind the method is not true is considered to be *robust* against that assumption.
  - ▷ One-sample t test or two-sample t test are *robust* against the assumption of normality.
  - ▷ A method is robust is no assurance that the method is still powerful when the population is nonnormal.

- ▶ *Nonparametric methods*
  - ▷ Nonparametric methods are based on some of the same assumptions on which parametric methods are based, such as the assumption that the sample is a random sample.

  - ▷ However, nonparametric methods do not assume a population probability distribution, and are therefore valid for data from any population with any probability distribution, which can remain unknown.
  - ▷ Nonparametric methods are perfectly robust for distribution assumptions on the population, because they are equally valid for all assumptions.
  - ▷ If the population distribution function has *lighter tails* than the normal distribution, such as the uniform distribution, then the parametric procedures based on the normality assumption generally have good power, equal to or greater than the power of nonparametric methods based on ranks, presented in Chap. 5.
  - ▷ On the other hand, if the population distribution function has *heavier tails* than the normal distribution, such as with the exponential distribution (presented in Chap. 6), the lognormal distribution (where the logs of the data follow a normal distribution), the chi-squared distribution (or its parent family of gamma distributions), and many other distributions that appear to be reasonable population models, then the parametric methods based on the normality assumption may have low power compared to nonparametric methods based on ranks.
  - ▷ Data containing outliers are good examples for considering to use nonparametric methods.
  - ▷ In those cases it is important to consider using nonparametric methods, such as the rank methods introduced in Chap. 5, to analyze the data because of the superior power of those rank methods when compared with the parametric methods based on the normal assumption.

- ▶ *Asymptotically distribution-free*

  - ▷ Many parametric tests that are robust against the assumption of nonnormality are also *asymptotically distribution-free*.
  - ▷ As the sample size gets larger the method become more robust, approaching the point where for an infinite sample size the method becomes exact, no matter what the population distribution may be.
  - ▷ The central limit theorem is usually the basis for showing parametric methods based on the sample mean to be asymptotically distribution-free.
  - ▷ A statistical procedure should not be preferred over others simply because it is nonparametric, or robust, or asymptotically distribution-free.
  - ▷ The relative power of the parametric test, or the relative size of the confidence interval, as compared with its nonparametric alternative, usually remains good or remains bad regardless of the size of the sample, even if the procedure is asymptotically distribution-free.
  - ▷ Keep in mind that most methods we are considering are consistent, which means that larger sample sizes mean more absolute power.
  - ▷ Carefully selecting the more powerful procedure may be unnecessary if the sample size is large enough to reject the null hypothesis using a less-powerful test, or if the confidence interval is small enough for the experimenter's purposes using a less efficient method.

- ▶ *Methods for analyzing nominal data*

  - Most people think of nonparametric methods they think of methods based on ranks presented in Chap. 5 and 6.
  - Nonparametric methods also may be used on qualitative data, with nominal scale of measurement, or with ordinal scale data.
  - The concept of a population probability distribution for nominal or ordinal data is difficult to imagine without treating such data as if it were at least interval, so there are *no parametric methods for purely nominal or ordinal data*.
  - Chap. 3 and 4 present methods for analyzing qualitative data.
  - Most of the methods in Chap. 5 and 6 are valid for ordinal data.

Definition of nonparametric

Definition 2.5.1 *A statistical method is nonparametric if it satisfies at least one of the following criteria.*

1. *The method maybe used on data with a nominal scale of measurement.*
2. *The method may be used on data with an ordinal scale of measurement.*
3. *The method may be used on data with an interval or ratio scale of measurement, where the distribution function of the random variable producing the data is either unspecified or specified except for an infinite number of unknown parameters.*

- Nearly all nonparametric hypothesis tests satisfy one of these two criteria.
- This book primarily concerned with hypothesis testing and the forming of confidence intervals.
- Several types of problems are not covered in this book: bioassay, survival curves, longitudinal studies, multivariate methods, discrimination analysis. Robust methods are methods that depend to some extent on the population distribution function but are not very sensitive to departures from the assumed distributional form.
- Robust methods are discussed briefly in Section 5.12.

2.6 Summary

1. *Statistic:* Concepts of the basic science for data analysis. 32
2. *Population:* The collection of all elements under investigation. 33
3. *Sample:* A collection of some elements of a population. 33
4. *Target population:* The population about which information is wanted. 33
5. *Sampled population:* The population to be sampled. 33
6. *Random sample:* A sample from a finite population is a random sample if each of the possible samples was equally likely to be obtained. 33
7. A random sample of size n is a sequence of n independent and identically distributed random variables $X_1, X_2, \ldots, X_n$. 33

8. *Multivariate random variable:* k-variate

$$X_i = (Y_{i1}, Y_{i2}, \dots, Y_{ik})$$

. . . . 33

9. *Measurement scale:* An excellent paper by Stevens(1946). 34

10. *Nominal scale:* Use numbers merely as a means of separating the properties of elements into different classes or categories. 34

11. *Ordinal scale:* Refer to measurements where only the comparisons "greater", "less", or "equal" between measurements are relevant. 34

12. *Interval scale:* Consider as pertinent information not only the relative order of the measurements as in the ordinal scale but also the size of the interval between measurements. 34

13. *Ratio scale:* Not only the order and interval size are important, but also the ratio between two measurements is meaningful. 34

14. A statistic is a function which assigns real numbers to the points of a sample space, where the points of the sample space are possible values of some multivariate random variable. In other words, a *statistic* is a function of several random variables. 35

15. *Ordered observation:* $x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}$ 35

16. The *order* statistic of rank k, $X^{(k)}$, is the statistic that takes as its value the kth smallest element $x^{(k)}$ in each observation $(x_1, x_2, \dots, x_n)$ of $(X_1, X_2, \dots, X_n\ )$. 35

17. *Estimator:* A statistic that is used to estimate. 35

18. Let $X_1, X_2, \dots, X_n$ be a random sample. The *empirical distribution function* $S(x)$ is a function of x, which equals the fraction of X_is that are less than or equal to x, $-\infty < x < \infty$. 36

19. Let $X_1, X_2, \dots, X_n$ be a random sample. The *pth sample quantile* is that number Q_p satisfies the two conditions:

    1. The fraction of the X_is that are less than Q_p is $\leq p$.
    2. The fraction of the X_is that exceed Q_p is $\leq 1 - p$.

    36

20. Let $X_1, X_2, \dots, X_n$ be a random sample. The *sample mean* $\bar{X}$ is defined by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

The *sample variance* S^2 is

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2$$

Sample standard deviation S is the square root of the sample variance. 37

21. *Interval estimator:* An interval estimator consists of two statistics, one for each end of the interval. 37

22. *Confidence coefficient:* The probability that the interval estimator will contain the unknown population quantity. 37

23. An estimator $\hat{\theta}$ is an *unbiased estimator* of a population parameter θ if $E\left[\hat{\theta}\right] = \theta$.38

24. Let $X_1, X_2, \ldots, X_n$ be independent random variables from a population with mean μ and variance σ^2. Then $E\left[\bar{X}\right] = \mu$ and $\mathrm{Var}\left(\bar{X}\right) = \sigma^2/n$. 38

25. *Standard error:* The standard deviation of an estimator. 38

26. *Unbiased estimator for σ^2*

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

. . . . 38

27. *Survival function:* $P(x) = 1 - F(x)$ 40

28. *Kaplan-Meier estimator* of $P(x)$:

$$\hat{P}(x) = \begin{cases} 1 & \text{for } x < u_1 \\ \prod_{u_i \le x} \hat{p}_i & \text{for } x \ge u_1 \end{cases}$$

where $p_i = P(X > u_i | X > u_{i-1})$ and $\hat{p}_i = \frac{\text{Number of items known to survive past time } u_i}{\text{Number of items known to survive past time } u_{i-1}}$. . . 41

29. *Alternative (research) hypothesis:* The statement that the experimenter would like to prove. 42

30. *Null hypothesis:* The negation of the alternative hypothesis. 43

31. *Test statistic:* A good test statistic is a sensitive indicator of whether the data agree or disagree with the null hypothesis. 43

32. *Decision rule:* Decide whether to accept or reject the null hypothesis. 43

33. *Null hypothesis:* H_0 : The machine is operating properly 43

34. *Alternative hypothesis:* H_1 : The machine needs attention 43

35. The hypothesis is *simple* if the assumption that the hypothesis is true leads to only one probability function defined on the sample space. The hypothesis is *composite* if the assumption that the hypothesis is true leads to only two or more probability functions defined on the sample space. 43

36. A *test statistic* is a statistic used to make the decision in a hypothesis test. 43

37. The *critical (rejection) region* is the set of all points in the sample space that result in the decision to reject the null hypothesis. 44

38. The *type I error* is the error of rejecting a true null hypothesis. 44

39. The *type II error* is the error of accepting a false null hypothesis. 44

40. The *level of significance*, or α, is the maximum probability of rejecting a true null hypothesis. 44

41. The *null distribution* of the test statistic is its probability distribution when the null hypothesis is assumed to be true. 44

42. The *power*, denoted by $1 - \beta$, is the probability of rejecting a false null hypothesis. 44

43. The *p-value* is the smallest significance level at which the null hypothesis would be rejected for the given observation. 44

44. *Power function:* If H_1 is composite, the power may vary as the probability function varies. If H_1 is stated in terms of some unknown parameter, the power usually may be given as a function of that parameter. 46

45. An *unbiased test* is a test in which the probability of rejecting H_0 when H_0 is false always greater than or equal to the probability of rejecting H_0 when H_0 is true. . 47

46. A sequence of tests is *consistent against all alternatives in the class* H_1 if the power of the tests approaches 1.0 as the sample size approaches infinity, for each fixed alternative possible under H_1. The level of significance of each test in the sequence is assumed to be as close as possible to but not exceeding some constant $\alpha > 0$. . 47

47. *Relative efficiency:* Efficiency is a relative term and is used to compare the sample size of one test with that of another test under similar conditions. 49

48. Let T_1 and T_2 represent two tests that test the same H_0 against the same H_1, with the critical regions of the same size α and with the same value of β. The relative efficiency of T_1 to T_2 is the ratio n_2/n_1, where n_1 and n_2 are the sample sizes of the tests T_1 and T_2, respectively. 49

49. *Asymptotic relative efficiency:* The relative efficiency depends on the choice of α, the choice of β, and the particular alternative being considered if H_1 is composite. . . . 49

50. Let n_1 and n_2 be the sample sizes required for two tests T_1 and T_2 to have the same power under the same level of significance. If α and β remain fixed the limit of n_2/n_1, as n_1 approaches infinity, is called the *asymptotic relative efficiency* (A.R.E) if the first test to the second test, if that limit is independent of α and β. 50

51. A test is *conservative* if the actual level of significance is smaller than the stated level of significance. 50

52. A statistical method is nonparametric if it satisfies at least one of the following criteria.

    (a) The method maybe used on data with a nominal scale of measurement.

    (b) The method may be used on data with an ordinal scale of measurement.

    (c) The method may be used on data with an interval or ratio scale of measurement, where the distribution function of the random variable producing the data is either unspecified or specified except for an infinite number of unknown parameters.

    53

Chapter 3

Some Tests Based on the Binomial Distribution

Contents

Binomial probability distribution

- Probabilities associated with the number of heads when a coin is tossed n times.
- $P(\text{Head}) = p$ and $P(\text{tail}) = 1 - p = q$
- The *binomial distribution* describes the probability of obtaining exactly k heads.
- Table A3 presents some of the binomial distribution functions.
- Many experimental situations in the applied sciences may be modeled this way.
    - Several customers enter a store and independently decide to buy or not to buy a particular product.
    - Several animals are given a certain medicine and either the are cured or not cured.
    - Examples can be found in almost any field.
- Data obtained in these situations may be analyzed by methods based on the binomial distribution.
- In this chapter we present a few of the available methods. The literature abounds with other procedures based on the binomial distribution.
- After studying the variety of tests presented in this chapter, the reader should be able to invent variations to match a given experimental situation.

3.1 The binomial test and estimation of p

- In Example 2.3.1 the binomial test was applied to a quality control problem.
- This entire chapter (Chap. 3) is little more than an elaboration of Example 2.3.1, showing the many uses and amazing versatility of that simple little binomial test.
- The *binomial test* may be adapted to test almost any hypothesis, with almost any type of data amenable to statistical analysis.
- In some situations the binomial test is the most powerful test; in those situations the test is claimed by both parametric and nonparametric statistic.
- In other situations more powerful tests are available, and the binomial test is claimed only by nonparametric statistic.
- However, even in situations where more powerful tests are available, the binomial test is sometimes preferred because it is usually simple to perform, simple to explain, and sometimes powerful enough to reject the null hypothesis when it should be rejected.

The binomial test

Data The sample consists of the outcomes of n independent trials. Each outcome is in either "class 1" or "class 2", but not both. The number of observations in class 1 is O_1 and the number of observations in class 2 is $O_2 = n - O_1$.

Assumptions

1. The n trials are mutually independent.
2. Each trial has probability p of resulting in the outcome "class 1", where p is the same for all n trials.

Test statistic T: Number of times the outcome is "class 1".

Null distribution

- Table A3 for $n \leq 20$ and selected values of p.
- For other values of n and p the normal approximation is used.

$$x_q = np + z_q\sqrt{np(1-p)}$$

Hypothesis

A. *Two-tailed test*:

$$H_0: \ p = p^* \qquad H_1: \ p \neq p^*$$

- α_1: Size of the lower tail.
- α_2: Size of the upper tail.
- $\alpha = \alpha_1 + \alpha_2$: Size of test.
- Table A3 for t_1 and t_2 such that $P(Y \leq t_1) = \alpha_1$ and $P(Y \geq t_2) = \alpha_2$.
- If $n > 20$ use the normal approximation.
- Reject H_0 if T is less than or equal to t_1 or if T is greater than t_2. Otherwise accept the null hypothesis.

- The p-value is twice the smaller of the probabilities that Y is less than or equal to the observed value of T, or greater or equal to the observed value of T.
- For $n > 20$, using

$$P(Y \leq t_{\text{obs}}) \approx P\left(Z \leq \frac{t_{\text{obs}} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

$$P(Y \geq t_{\text{obs}}) \approx 1 - P\left(Z \leq \frac{t_{\text{obs}} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

B. Lower-tailed test:

$$H_0: \ p \geq p^* \qquad H_1: \ p < p^*$$

- Table A3 for t such that $P(Y \leq t) = \alpha$.
- If $n > 20$ use the normal approximation.
- Reject H_0 if T is less than or equal to t. Otherwise accept the null hypothesis.
- The p-value is the probability that Y is smaller than or equal to the observed value of T.
- For $n > 20$, using

$$P(Y \leq t_{\text{obs}}) \approx P\left(Z \leq \frac{t_{\text{obs}} - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

C. Upper-tailed test:

$$H_0: \ p \leq p^* \qquad H_1: \ p > p^*$$

- Table A3 for t such that $P(Y \leq t) = 1 - \alpha$.
- If $n > 20$ use the normal approximation.
- Reject H_0 if T is greater than t. Otherwise accept the null hypothesis.
- The p-value is the probability that Y is greater than or equal to the observed value of T.
- For $n > 20$, using

$$P(Y \geq t_{\text{obs}}) \approx 1 - P\left(Z \leq \frac{t_{\text{obs}} - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Computer assistance *Minitab*, *SAS*, *S-Plus*, and *StatXact* can perform binomial test.

Example 3.1.1 *It is estimated that at least half of the men who currently undergo an operation to remove prostate (前列腺) cancer suffer a particular undesirable side effect. In an effort to reduce the likelihood of this side effect the FDA studied a new method of performing the operation. Out of 19 operations only 3 men suffered the unpleasant side effect. Is it safe to conclude the new method of operating is effective in reducing the side effect?*

- ▶ p: Probability of the patient experiencing the side effect.
- ▶ $H_0: \ p \geq 0.5 \quad H_1: \ p < 0.5$ (Lower-tailed test)

- $\alpha = 0.05,\ n = 19$ and $p = 0.5$.
- $P(Y \leq 5) = 0.0318$ (Table A3)
- Reject H_0 since the observed value of $T = 3$ is smaller than 5 ($P(Y \leq 3) = 0.0022$).
- One should always use exact methods when exact methods are available.
- Normal approximation:
  - $x_{0.05} = 19(0.5) + (-1.6449)\sqrt{19(0.5)(0.5)} = 5.9$ resulting in the same rejection region as before.
  - $P(Y \leq 5) \approx P\left(Z \leq \frac{5-19(0.5)+0.5}{\sqrt{19(0.5)(0.5)}}\right) = 0.033$ is close to the exact α, 0.032.
  - $P(Y \leq 3) \approx P\left(Z \leq \frac{3-19(0.5)+0.5}{\sqrt{19(0.5)(0.5)}}\right) = 0.003$ is close to the exact p-value, 0.002.

  □

Example 3.1.2 *Under simple Mendelian inheritance (孟德爾遺傳) a cross between plants of two particular genotypes may be expected to produce progeny one-fourth of which are "dwarf" and three-fourths of which are "tall". In an experiment to determine if the assumption of simple Mendelian inheritance is reasonable in a certain situation, a cross results in progeny (後嗣) having* 243 *dwarf plants and* 682 *tall plants.*

- If "class 1" denotes "tall", then $p^* = 3/4$ and T equals the number of tall plants.
- $H_0:\ p = 3/4 \quad$ v.s. $\quad H_1:\ p \neq 3/4$
- $\alpha = 0.05,\ n = 925$ and $p^* = 3/4$.
- $t_1 = (925)(3/4) + (-1.960)\sqrt{(925)(3/4)(1/4)} = 667.94$
- $t_2 = (925)(3/4) + (1.960)\sqrt{(925)(3/4)(1/4)} = 719.56$
- Accept H_0 since $t_1 < 682 < t_2$.
- $P(Y \leq 682) \approx P\left(Z \leq \frac{682-693.75+0.5}{13.17}\right) = P(Z \leq -0.8542) = 0.196$
- p-value: $2P(Y \leq 682) = 0.392$ □

The previous example illustrates the two-tailed form of the binomial test. The one-tailed binomial test was also illustrated in Example 2.3.1.

Theory

- That the test statistic in the binomial test has a binomial distribution is easily seen by comparing the assumptions in the binomial test with the assumptions in Examples 1.3.5 and 1.2.8.
- If T equals the number of trials that result in the outcome "class 1," where the trials are mutually independent and where each trial has probability p of resulting in that outcome (as stated by the assumptions), then T has the binomial distribution with parameters p and n. The size of the critical region is a maximum when p equals p^*, under the null hypothesis, and so Table A3 is entered with n and p^* to determine the exact value of α.

Hypothesis testing is only one branch of statistical inference. We will now discuss another branch, interval estimation.

Confidence interval

- *Interval estimator* (L, U) where L and U are the lower and upper interval boundary respectively.
- *Confidence coefficient*: The probability that the unknown population parameter lies within its interval estimates.
- The interval estimator together with the confidence coefficient provide us with the *confidence interval*.

Confidence interval for a probability or population proportion

Data A sample consisting of observations on n independent trials is examined, and the number Y of times the specified event occurs is noted.

Assumptions

1. The n trial are mutually independent.
2. The probability p of the specified event occurring remains constant from one trial to the next.

Method A For n less than or equal to 30, and confidence coefficients of $0.90, 0.95$, or 0.99, use Table A4. Simply enter the table with sample size n and the observed Y. Reading across gives the exact lower and upper bounds in the columns for the desired confidence interval.

Method B For n greater than 30, use the normal approximation.

$$L = \frac{Y}{n} - z_{1-\alpha/2}\sqrt{Y(n-Y)/n^3}$$
$$U = \frac{Y}{n} + z_{1-\alpha/2}\sqrt{Y(n-Y)/n^3}$$

Example 3.1.3 *In a certain state 20 high schools were selected at random to see if they met the standards of excellence proposed by a national committee on education. It was found that 7 schools did qualify and accordingly were designated "excellent." What is a 95% confidence interval for p, the proportion of all high schools in the state that would qualify for the designation "excellent"?*

- The high schools are classified "excellent" or "not excellent" independently of one another.
- $n = 20$ and $Y = 7$.

Method A: $(L, U) = (.154, .592)$ and $P(.154 < p < .592) = .95$ by Table A4.

Method B:

$$\begin{aligned} L &= \frac{Y}{n} - z_{.975}\sqrt{Y(n-Y)/n^3} \\ &= .35 - (1.960)\sqrt{(7)(13)/(20)^3} \\ &= .141 \\ U &= .35 + .209 = .559 \end{aligned}$$

$$P(.141 < p < .559) = .95$$

The confidence interval furnished by the normal approximation is from 0.141 to 0.559, which is close to the exact confidence interval, but still different enough to show the clear advantage of using exact intervals when they are available. □

Theory

- *Exact Method A*: The confidence interval consists of all values of p^* such that the data obtained in the sample would result in acceptance of

$$H_0: \ p = p^* \quad \text{v.s.} \quad H_1: \ p \neq p^*$$

For the given values of Y, which values may we use for p^* in the hypothesis such that a two-tailed binomial test (at level α) would result in acceptance of H_0?

  - Since each tail of the binomial test has probability $\alpha/2$, the value of L is selected as the value of p^* that would barely result in rejection of H_0, for the given value of Y, say y, or a larger value. Thus p_1^* is selected so that

$$P(Y \geq y | p = p_1^*) = \frac{\alpha}{2} = \sum_{i=y}^{n} \binom{n}{i} (p_1^*)^i (1-p_1^*)^{n-i}$$

  and then $L = p_1^*$.

  - Another value of p^* is selected so the same value y is barely in the lower tail. That is, p_2^* is selected so

$$P(Y \leq y | p = p_2^*) = \frac{\alpha}{2} = \sum_{i=0}^{y} \binom{n}{i} (p_2^*)^i (1-p_2^*)^{n-i}$$

  and we set $U = p_2^*$.

More information on confidence intervals for the binomial parameter p in Clopper and Pearson (1934).

- *Large sample approximation*:

  - If Y is a binomially distributed random variable with parameters p and large n, then

$$Z = \frac{Y - np}{\sqrt{npq}} \approx N(0,1).$$

  - $1 - \alpha \approx P\left(-z_{1-\alpha/2} < \frac{Y-np}{\sqrt{n(Y/n)(1-Y/n)}} < z_{1-\alpha/2}\right)$

  - ▷ $1-\alpha \approx P\left(p \in \frac{Y}{n} \pm z_{1-\alpha/2}\sqrt{\frac{Y(n-Y)}{n^3}}\right)$

- Multiplication by the sample size n in the preceding procedures gives nL and nU as the lower and upper bounds of the confidence interval for np, used to test hypotheses involving the mean of a binomial random variable.

- Other methods of obtaining binomial confidence limits are given by Anderson and Burstein (1967 and 1968).

- Methods dealing with simultaneous confidence intervals for multinomial proportions are given by Quesenberry and Hurst (1964) and Goodman (1965).

3.2 The quantile test and estimation of x_p

The binomial test may be used to test hypotheses concerning the quantiles of a random variable.
Test whether the median of X is 17.

- If the median of X is 17, then about half of the observations should fall on either side of 17.

- If very few of the sample observations are less than 17, the median of X appears to be larger than 17.

- If by far most of the sample observations are less than 17, the median of X appears to be less than 17.

- The measurement scale is usually at least ordinal for the quantile test, although the binomial test only required the weaker nominal for its measurements.

- For continuous random variable:

  $$H_0 : \text{The } p^*\text{th quantile of } X \text{ is } x^* \equiv H_0 : P(X \leq x^*) = p^*$$

  Two-tailed binomial test may be used.

- In general (including the discrete random variable):

  $$H_0 : \text{The } p^*\text{th quantile of } X \text{ is } x^* \equiv H_0 : P(X \leq x^*) \geq p^* \text{ and } P(X < x^*) \leq p^*$$

The quantile test

Data Let $X_1, X_2, \ldots, X_n$ be a random sample. The data consist of observations on the X_i.

Assumptions

1. The X_is are a random sample.
2. The measurement scale of the X_is is at least ordinal.

Test statistic
Let T_1 equal the number of observations less than or equal to x^*, and let T_2 equal the number of observations less x^*. Then $T_1 = T_2$ if none of the numbers in the data exactly equals x^*. Otherwise, T_1 is greater than T_2.

Null distribution

The null distribution of the test statistic T_1 and T_2 is the binomial distribution, with parameters n = sample size, and $p = p^*$ as given in the null hypothesis.

- ▷ The null distribution is given in Table A3 for $n \leq 20$ and selected values of p.
- ▷ For other values of n and p the normal approximation is used:

$$x_q \approx np + z_q\sqrt{np(1-p)}$$

Hypotheses

A. *Two-tailed test*

$$H_0:\ x_p = x^* \qquad H_1:\ x_p \neq x^*$$

$$H_0:\ P(X \leq x^*) \geq p^* \text{ and } P(X < x^*) \leq p^*$$

- ∗ α_1: Size of the lower tail.
- ∗ α_2: Size of the upper tail.
- ∗ $\alpha = \alpha_1 + \alpha_2$: Size of test.
- ∗ Table A3 for t_1 and t_2 such that $P(Y \leq t_1) = \alpha_1$ and $P(Y \leq t_2) = 1 - \alpha_2$.
- ∗ If $n > 20$ use the normal approximation.
- ∗ Reject H_0 if T_1 is less than or equal to t_1 [indicating that perhaps $P(X \leq x^*)$ is less than p^*] or if T_2 is greater than t_2 [indicating that perhaps $P(X < x^*)$ is greater than p^*]. Otherwise accept the null hypothesis.
- ∗ The p-value is twice the smaller of the probabilities that Y is less than or equal to the observed value of T_1, or greater or equal to the observed value of T_2.
- ∗ For $n > 20$, using

$$P(Y \leq T_1) \approx P\left(Z \leq \frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right) \tag{4}$$

$$P(Y \geq T_2) \approx 1 - P\left(Z \leq \frac{T_2 - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right) \tag{5}$$

B. *Lower-tailed test*:

$$H_0:\ x_p \leq x^* \qquad H_1:\ x_p > x^*$$

$$H_0:\ P(X < x^*) \geq p^* \quad H_1:\ P(X < x^*) < p^*$$

- ∗ Table A3 for t_1 such that $P(Y \leq t_1) = \alpha$.
- ∗ If $n > 20$ use the normal approximation.
- ∗ Reject H_0 if T_1 is less than or equal to t_1. Otherwise accept the null hypothesis.
- ∗ The p-value is the probability that Y is smaller than or equal to the observed value of T_1.
- ∗ For $n > 20$, using

$$P(Y \leq T_1) \approx P\left(Z \leq \frac{T_1 - np^* + 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

C. *Upper-tailed test*:

$$H_0: \ x_p \geq x^* \quad \text{v.s.} \quad H_1: \ x_p < x^*$$

$$H_0: \ P(X < x^*) \leq p^* \quad \text{v.s.} \quad H_1: \ P(X < x^*) > p^*$$

- Table A3 for t_2 such that $P(Y \leq t_2) = 1 - \alpha$.
- If $n > 20$ use the normal approximation.
- Reject H_0 if T_2 is greater than t_2. Otherwise accept the null hypothesis.
- The p-value is the probability that Y is greater than or equal to the observed value of T_2.
- For $n > 20$, using

$$P(Y \geq T_2) \approx 1 - P\left(Z \leq \frac{T_2 - np^* - 0.5}{\sqrt{np^*(1-p^*)}}\right)$$

Example 3.2.1 (The two-tailed quantile test) *Entering college freshmen have taken a particular high school achievement examination for many years, and the upper quartile is well established at a score of 193. A particular high school sends 15 of its graduates to college, where they take the exam and get the following scores*

189	*233*	*195*	*160*	*212*
176	*231*	*185*	*199*	*213*
202	*193*	*174*	*166*	*248*

- H_0: The upper quantile is 193
  H_1: The upper quantile is not 193

- $\alpha = 0.05, n = 15$ and $p = 0.75$ from Table A3:

$$\begin{aligned} P(Y \leq 7) &= 0.0173 \\ P(Y \leq 14) &= .9866 = 1 - 0.0134 \\ \alpha &= 0.0173 + 0.0134 = 0.0307 \end{aligned}$$

- In this example $T_1 = 7$ the number of observations less than or equal to 193, and $T_2 = 6$, since one observation exactly equals 193. Reject H_0 since T_1 is too small.

- p-value is $2P(Y \leq 7) = 0.0346$. □

Example 3.2.2 (The one-tailed quantile test with the large sample approximation) *The time interval between of Old Faithful geyser is recorded 112 times to see whether the median interval is less than or equal to 60 minutes (null hypothesis) or whether the median interval is greater than 60 minutes (alternative hypothesis). If the median interval is 60, 60 is $x_{0.50}$, or the median. If the median interval is less than 60, 60 is a p quantile for some $p \geq 0.50$.*

- Let X be the time interval between eruptions.

$$H_0: P(X \leq 60) \geq .50 \quad \text{v.s.} \quad H_1: P(X \leq 60) < .50$$

- Use lower-tailed quantile test.
- T_1 = Number of intervals that are less than or equal to 60 minutes.
- Critical region of size 0.05 corresponds to values of T_1 less than or equal to

$$\begin{aligned} t_1 &= np^* + z_{.05}\sqrt{np^*(1-p^*)} \\ &= (112)(.50) - (1.645)\sqrt{(112)(.50)(.50)} \\ &= 47.3 \end{aligned}$$

- Reject H_0 since $T_1 = 8$.
- p-value by Equation 4:

$$P(Y \leq 8) \cong P\left[Z \leq \frac{8-(112)(.50)}{\sqrt{(112)(.50)(.50)}}\right] = P(Z \leq -8.977) \ll .0001 \qquad \square$$

Theory

- Why the hypotheses within the parentheses in A, B, and C are equivalent to the hypotheses not in parentheses. Refer to Fig. 1.
- $x_{p^*} \leq x^*$ implies $p^* \leq p_0$ if $x^* = x_{p_0}$.
- $P(X > x^*) \leq 1 - p_0$ is the same as $p_0 \leq 1 - P(X > x^*) = P(X \leq x^*)$.
- Since $p^* \leq p_0$, this implies

$$p^* \leq P(X \leq x^*) \tag{15}$$

  which is the equivalent form of H_0 in set B of hypotheses.
- The negation of H_0 is H_1, and the negation of Equation 15 is

$$p^* > P(X \leq x^*).$$

- Fig. 1 is used to visualize that $x_{p_0} \leq x_{p^*}$ (H_0 in C) implies that $p_0 \leq p^*$.
- If $x^* = x_{p_0}$ then by Definition 1.4.1

$$P(X < x^*) \leq p_0 \leq p^*$$

  is true, which furnishes the equivalent form of H_0.
- The binomial test is applied directly to test the hypothesis in parentheses. H_0 in C is tested by defining the "class 1" of the binomial test as those observations at least as great as x^*.
- H_0 in B is tested by considering "class 1" to represent those observations less than or equal to x^*.
- The two tests in B and C are combined to give the two-tailed test in A.

1. Previous section shows how to find a confidence interval for a probability p.
2. The same method is used to find a confidence interval for $F(x_0)$.

3. In the previous section we showed how to find a confidence interval for a probability p (vertical confidence interval).

4. This section shows how to find a confidence interval for a quantile by using the ordered statistics (horizontal confidence interval).

$$P\big(X^{(r)} \leq x_{p^*} \leq X^{(s)}\big) = 1 - \alpha$$

   where $1 - \alpha$ is a known *confidence coefficient* and where $X^{(r)})$ and $X^{(s)}$ are order statistics with rand s specified.

5. This statistical method may be applied freely to any random sample from any population.

Confidence interval for a quantile

Data The data consist of observations on $X_1, X_2, \ldots, X_n$, which are independent and identically distributed. We wish to find a confidence interval for p^* quantile, where p^* is some specified number between zero and one.

Assumptions

1. The sample $X_1, X_2, \ldots, X_n$ is a random sample.
2. The measurement scale of the X_is is at least ordinal.

Method A (small samples)

- For $n \leq 20$ Table A3 may be used to find y, $r = y + 1$ and $s = y + 1$.
- $P\big(X^{(r)} \leq x_{p^*} \leq X^{(s)}\big) \geq 1 - \alpha_1 - \alpha_2$ provides the confidence interval.
- If the unknown distribution is continuous, then $P\big(X^{(r)} \leq x_{p^*} \leq X^{(s)}\big) = 1 - \alpha_1 - \alpha_2$

Method B (large sample approximation)

- For n greater than 20 the approximation based on the central limit theorem may be used.
- $r^* = np^* + z_{\alpha/2}\sqrt{np^*(1-p^*)}$ and $s^* = np^* + z_{1-\alpha/2}\sqrt{np^*(1-p^*)}$
- Let r and s be the integers obtained by rounding r^* and s^* upward to the next higher integers.
- One-sided confidence intervals:

$$P\big(X^{(r)} \leq x_{p^*}\big) = 1 - \alpha_1 \quad \text{and} \quad P\big(x_{p^*} \leq X^{(s)}\big) = 1 - \alpha_2$$

  if the distribution function is continuous.

$$P\big(X^{(r)} \leq x_{p^*}\big) \geq 1 - \alpha_1 \quad \text{and} \quad P\big(X^{(s)} \leq x_{p^*}\big) \geq 1 - \alpha_2$$

  otherwise.

Example 3.2.3 *Sixteen transistors are selected at random from a large batch of transistors and are tested. The number of hours until failure is recorded for each one. We wish to find a confidence interval for the upper quartile, with a confidence coefficient close to 90%.*

- $\alpha_1 = 0.0271, y = 8, r = 9$ from Table A3 with $n = 16, p = 0.75$.
- The probability closest to 0.95 is $0.9365 = 1 - \alpha_2$ which has a corresponding y of 14. Thus $s = 15$.
- CI: $P\big(X^{(9)} = 63.3 \leq x_{.75} \leq X^{(15)} = 73.3\big) = 0.9094$
- Large sample approximation:

$$\begin{aligned} r^* &= (16)(.75) + (-1.645)\sqrt{(16)(.75)(.25)} \\ &= 12 - 2.86 \\ &= 9.14 \\ s^* &= 12 + 2.86 \\ &= 14.86 \end{aligned}$$

Therefore $r = 10$ and $s = 15$, so the 90% CI becomes $(63.4, 73.3)$. □

Theory: Consider first the simpler case where the distribution function is continuous. If x_{p^*} is the p^*th quantile, we have the exact relationship

$$P(X \geq x_{p^*}) = P(X > x_{p^*}) = 1 - p^*$$

where the distribution function of X is the same as that of the random sample.

- $P\big(x_{p^*} < X^{(1)}\big) = (1 - p^*)^n$
- $P\big(x_{p^*} < X^{(2)}\big) = \sum_{i=0}^{1} \binom{n}{i} (p^*)^i (1 - p^*)^{n-i}$
- $P\big(x_{p^*} < X^{(r)}\big) = \sum_{i=0}^{r-1} \binom{n}{i} (p^*)^i (1 - p^*)^{n-i}$
- Confidence coefficient:

$$1 - \alpha \approx P\big(X^{(r)} \leq x_{p^*} \leq X^{(s)}\big)$$

- $P\big(x_{p^*} \leq X^{(s)}\big) \approx 1 - \alpha/2$
- $P\big(x_{p^*} < X^{(r)}\big) \approx \alpha/2$

If X is not continuous: conservative, $=$ changes to $\geq$ or $\leq$.

- $P(X > x_{p^*}) \leq 1 - p^*$ and $P(X \geq x_{p^*}) \geq 1 - p^*$
- $P\big(x_{p^*} < X^{(r)}\big) \leq \sum_{i=0}^{r-1} \binom{n}{i} (p^*)^i (1 - p^*)^{n-i} \leq \alpha_1$
- $P\big(x_{p^*} \leq X^{(s)}\big) \geq \sum_{i=0}^{s-1} \binom{n}{i} (p^*)^i (1 - p^*)^{n-i} \geq 1 - \alpha_2$
-

$$\begin{aligned} P\big(X^{(r)} \leq x_{p^*} \leq X^{(s)}\big) &= P\big(x_{p^*} \leq X^{(s)}\big) - P\big(x_{p^*} < X^{(r)}\big) \\ &\geq P\big(x_{p^*} \leq X^{(s)}\big) - \alpha_1 \\ &\geq 1 - \alpha_1 - \alpha_2 \end{aligned}$$

- This method may be conservative with discrete random variables, or ordinal data with ties.
- The method of finding a confidence interval for a quantile has been justified for the case where exact tables of the binomial distribution function are available.
- The large sample method of obtaining r and s is based on the use of the standard normal distribution to approximate the binomial distribution.

3.3 Tolerance limits

Confidence intervals of Sec. 3.1 and 3.2 provide interval estimates for unknown population parameters p and x_p.
Tolerance limits provide an interval within which at least a proportion q of the population lies, with probability $1-\alpha$ or more that stated interval does indeed contain the proportion q of the population.

- A typical application: How large must the sample size n be so that at least a proportion q of the population is between $X^{(1)}$ and $X^{(n)}$ with probability $1-\alpha$ or more?
- In general: How large must the sample size n be so that at least a proportion q of the population is between $X^{(r)}$ and $X^{(n+1-m)}$ with probability $1-\alpha$ or more?
- The numbers q, r, m, and $1-\alpha$ are known (or selected) beforehand, and only n needs to be determined.
- Another typical situation: When a random sample is available, we want 95% confidence that the limits we choose will contain at least q of the population.
- What will the population proportion q be if we choose the two extreme values in the sample, $X^{(1)}$ and $X^{(n)}$, as our limits?
- In this version of the problem, q is the unknown quantity and is obtained after we know, or set, values for α, n, r, and m.
- One-sided tolerance limits: At least a proportion q of the population is greater than $X^{(r)}$, with probability $1-\alpha$.

Data The data consist of a random sample $X_1, X_2, \ldots, X_n$ from a large population. Choose a confidence coefficient $1-\alpha$ and a pair of positive integers r and m. Either we wish to *determine the required sample size n* after selecting a desired population proportion q (see Method A), or we wish to *determine the population proportion q* for a given sample size n (see Method B). We are trying to make the statement, "The probability is $1-\alpha$ that the random interval from $X^{(r)}$ to $X^{(n+1-m)}$ inclusive contains a proportion q or more of the population." Note that we are using the convention $X^{(0)} = -\infty$ and $X^{(n+1)} = \infty$, so that one-sided tolerance limits may be obtained by setting either r or m equal to zero.

Assumptions

1. The $X_1, X_2, \ldots, X_n$ constitute a random sample.
2. The measurement scale is at least ordinal.

Method A (to find n) If $r+m$ equals 1, that is, if either r or m equals zero as in one-sided tolerance limit, read n directly from Table A5 for the appropriate values of α and q. If $r+m$ equals 2, read n directly from Table A6 for the appropriate values of α and q. If Table A5 and A6 are not appropriate, use the approximation

$$n \approx \frac{1}{4} x_{1-\alpha} \frac{1+q}{1-q} + \frac{1}{2}(r+m-1)$$

where $x_{1-\alpha}$ is the $(1-\alpha)$ quantile of a chi-squared random variable with $2(r+m)$ degrees of freedom, obtained from A2.

Method B (to find q) For a given sample size n and selected values of α, r, and m, the approximate value of q, the proportion of the population, is given by

$$q = \frac{4n - 2(r+m-1) - x_{1-\alpha}}{4n - 2(r+m-1) + x_{1-\alpha}}$$

Tolerance limit With a sample of size n, there is probability at least $1-\alpha$ that q of the population is between $X^{(r)}$ and $X^{(n+1-m)}$ inclusive. For one-sided tolerance regions let either r or m equal zero.

Example 3.3.1 (Two-sided tolerance limit) *Probability the most widely used two-sided tolerance limits are those where $r = 1$ and $m = 1$. Electric seat adjusters are available on a popular luxury car. The manufacturer wants to know what range of vertical adjustment is necessary to be 90% certain that at least 80% of population of potential buyers will be able to adjust their seats to the desired height. What must n be so that $X^{(n)}$ and $X^{(1)}$ furnish our upper and lower limits?*

- $n = 18$ from Table A6 with $q = 0.80$ and $1-\alpha = 0.90$.
- Approximation by Equation 1:

$$\begin{aligned} n &\cong \frac{1}{4} x_{1-\alpha} \frac{1+q}{1-q} + \frac{1}{2}(r+m-1) \\ &= \frac{1}{4}(7.779)\frac{1.80}{0.20} + \frac{1}{2} \\ &= 18.003 \end{aligned}$$

- Largest and smallest value in sample:

$$X^{(18)} = 7.57, \qquad X^{(1)} = 1.21$$

- There is a probability 0.90 that at least 80% of the population requires a vertical seat adjustment to or between 1.21 and 7.57 inches. □

Example 3.3.2 (One-sided tolerance limit) *Along with each lot of steel bars, the manufacturer guarantees that at least 90% of the bars will have a breaking point above a number specified for each lot. Because of variable manufacturing conditions the guaranteed breaking points is established separately for each lot by breaking a random sample of bars from each lot and setting the guaranteed breaking point equal to the minimum breaking point in the sample. How large should the sample be so that the manufacture can be 95% sure the guarantee statement is correct?*

- $n = 29$ from Table A5 with $q = 0.90$ and $1 - \alpha = 0.95$.

- In each lot a sample of size 29 is selected at random, and the smallest breaking point of these bars in the sample is stated as the guaranteed breaking point, at which at least 90% of the bars in the lot will still be intact, with probability 0.95. □

Example 3.3.3 *A large population of drums (圓桶) containing radioactive waste is being stored for safe keeping. Each drum has marked on it the amount of radioactive waste contained in the drum. Periodic audits are made where randomly selected drums are scanned externally to estimate the amount of radioactive waste contained in the drum, and the estimate is compared with the label to obtain the discrepancy X. Over a period of three months 122 drums have been examined in this way, and the results are a random sample $X_1, \ldots, X_{122}$, where each X_i is the discrepancy between the amount marked on the drum and the amount estimated by the scan.*

- Select $r = 2, m = 2$ and $1 - \alpha = 0.95$.

- Approximate by Equation 2 with 0.95 quantile from $\chi^2_{2(r+m)=8}$:

$$q = \frac{4n - 2(r + m - 1) - x_{1-\alpha}}{4n - 2(r + m - 1) + x_{1-\alpha}} = \frac{488 - 6 - 15.51}{488 - 6 + 15.51} = 0.938$$

- We can be 95% confident that at least 93.8% of the drums have discrepancies between the second smallest and the second largest observed discrepancies in the 122 observed drums. □

Theory

- A careful examination of the statement furnished by the one-sided tolerance limit reveals the similarity it has with one-sided confidence interval for quantiles.

- One sided tolerance limit:

$$\begin{aligned} &P\big(\text{at least } q \text{ of the population is } \leq X^{(n+1-m)}\big) \\ =&P\big(\text{the } q \text{ quantile is } \leq X^{(n+1-m)}\big) \\ =&P\big(x_q \leq X^{(n+1-m)}\big) \geq 1 - \alpha \end{aligned} \tag{5}$$

From Equation 3.2.43 it gives

$$P\big(x_q \leq X^{(n+1-m)}\big) \geq \sum_{i=0}^{n-m} \binom{n}{i} q^i (1-q)^{n-i} \tag{6}$$

Rewrite the right side of Equation 6 as

$$\sum_{i=0}^{n-m} \binom{n}{i} q^i (1-q)^{n-i} = 1 - \sum_{i=n-m+1}^{n} \binom{n}{i} q^i (1-q)^{n-i} \tag{7}$$

A change of index, $j = n - i$, on the right side of Equation 7 results in

$$\sum_{i=0}^{n-m} \binom{n}{i} q^i (1-q)^{n-i} = 1 - \sum_{j=0}^{m-1} \binom{n}{j} (1-q)^j q^{n-j} \tag{8}$$

Equations 8 and 6 shows that we could find n by solving for the smallest value of n that satisfies

$$1 - \sum_{j=0}^{m-1} \binom{n}{j} (1-q)^j q^{n-j} \geq 1 - \alpha \tag{9}$$

$$\sum_{j=0}^{m-1} \binom{n}{j} (1-q)^j q^{n-j} \leq \alpha \tag{10}$$

- The other one-sided tolerance limit:

$$P\big(X^{(r)} \leq \text{at least } q \text{ of the population}\big) \geq 1 - \alpha$$

$$P\big(X^{(r)} \leq x_{1-q}\big) \geq 1 - \alpha \tag{12}$$

Equation 12 becomes

$$1 - P\big(X^{(r)} \leq x_{1-q}\big) = P\big(x_{1-q} < X^{(r)}\big) \leq \alpha \tag{13}$$

From Equation 3.2.41 we see that the solution to Equation 13 is the smallest value of n such that

$$\sum_{i=0}^{r-1} \binom{n}{i} (1-q)^i q^{n-i} \leq \alpha$$

- It can be shown, with the aid of calculus (see Noether, 1967a), that for the two-sided tolerance limits and for both two types of one-sided tolerance limits, the sample size n depends on the solution to

$$\sum_{i=0}^{r+m-1} \binom{n}{i} (1-q)^i q^{n-i} \leq \alpha.$$

which depends on $r + m$ only.

- The approximation in Equation 1 is furnished without proof by Scheffé and Tukey (1944).

- Equation 2 is obtained by solving Equation 1 for q.

3.4 The sign test

- The sign test is the oldest of all nonparametric tests.

- The sign test is just the binomial test with $p^* = 1/2$.

- Dating back to 1710.

- $p^* = 1 - p^* = 1/2$ makes it even simpler than the binomial test.
- Useful for testing whether one random variable in a pair (X, Y) tends to be larger than the other random variable in the pair.
- Test for trend in a series of ordinal measurements or as a test for correlation.
- In many situations where the sign test may be used, more powerful nonparametric tests are available for the same model.
- However, the sign test is usually simpler and easier to use, and special tables to find the critical region are sometimes not needed.

The sign test

Data The data consist of observations on a bivariate random sample $(X_1, Y_1), \ldots, (X_{n'}, Y_{n'})$, where there are n' pairs of observations. There should be some natural basis for pairing the observations; otherwise the Xs and Ys are independent, and the more powerful Mann-Whitney test of Chap. 5 is more appropriate. Within each pair (X_i, Y_i) a comparison is made, and the pair is classified as "+" or "plus" if $X_i < Y_i$ as "−" or "minus" if $X_i > Y_i$, or as "0" or "tie" if $X_i = Y_i$. Thus the measurement scale needs only to be ordinal.

Assumptions

1. The bivariate random variables X_i, Y_i are mutually independent.
2. The measurement scale is at least ordinal within each pair. That is, each pair (X_i, Y_i) may be determined to be a "plus", "minus", or "tie".
3. The pairs (X_i, Y_i) are internally consistent, in that if $P(+) > P(-)$ for one pair (X_i, Y_i), then $P(+) > P(-)$ for all pairs. The same is true for $P(+) < P(-)$, and $P(+) = P(-)$.

Test statistic T: Number of "plus" pairs.

Null distribution The null distribution of T is the binomial distribution with $p = 1/2$ and $n =$ the number of nontied pairs.

Hypothesis

A. *Two-tailed test*:

$$H_0: \ P(+) = P(-) \qquad H_1: \ P(+) \neq P(-)$$

∗ For $n \leq 20$, use Table A3 with the proper value of n and $p = 1/2$.

∗ $\alpha_1 = \alpha/2$: $P(Y \leq t) = \alpha_1$

∗ Reject H_0 if T is less than or equal to t or if T is greater than $n - t$. Otherwise accept the null hypothesis.

∗ If $n > 20$ use the normal approximation (Table A3).

$$t = (n + z_{\alpha/2}\sqrt{n})/2$$

∗ The p-value is twice the smaller of the probabilities that Y is less than or equal to the observed value of T, or greater or equal to the observed value of T.

- $*$ For $n > 20$, using

$$P(Y \le t_{\text{obs}}) \approx P\left(Z \le \frac{2t_{\text{obs}} - n + 1}{\sqrt{n}}\right)$$
$$P(Y \ge t_{\text{obs}}) \approx 1 - P\left(Z \le \frac{2t_{\text{obs}} - n - 1}{\sqrt{n}}\right)$$

B. *Lower-tailed test*:

$$H_0: \ P(+) \ge P(-) \qquad H_1: \ P(+) < P(-)$$

- ▷ Table A3 with $p = 1/2$ and n for t such that $P(Y \le t) = \alpha$.
- ▷ If $n > 20$ use the normal approximation.

$$t = (n + z_\alpha \sqrt{n})/2$$

- ▷ Reject H_0 if T is less than or equal to t. Otherwise accept the null hypothesis.
- ▷ The p-value is the probability that Y is smaller than or equal to the observed value of T.
- ▷ For $n > 20$, using

$$P(Y \le t_{\text{obs}}) \approx P\left(Z \le \frac{2t_{\text{obs}} - n + 1}{\sqrt{n}}\right)$$

C. *Upper-tailed test*:

$$H_0: \ P(+) \le P(-) \qquad H_1: \ P(+) > P(-)$$

- ▷ Table A3 for t such that $P(Y \le t) = \alpha$.
- ▷ If $n > 20$ use the normal approximation.
- ▷ Reject H_0 if T is greater than $n - t$. Otherwise accept the null hypothesis.
- ▷ The p-value is the probability that Y is greater than or equal to the observed value of T.
- ▷ For $n > 20$, using

$$P(Y \ge t_{\text{obs}}) \approx 1 - P\left(Z \le \frac{2t_{\text{obs}} - n - 1}{\sqrt{n}}\right)$$

Remarks: The sign test is unbiased and consistent when testing these hypotheses. The sign test is also used for testing the following counterparts of the hypotheses, in which case it is neither unbiased nor consistent unless additional assumptions concerning the distributions of (X_i, Y_i) are made.

A. *Two-tailed test*: X_i and Y_i have the same location parameter,

$$H_0: \ E(X_i) = E(Y_i) \qquad H_1: \ E(X_i) \ne E(Y_i)$$

or

H_0: The median of X_i equals the median of Y_i for all i.
H_1: X_i and Y_i have different medians for all i.

B. Lower-tailed test:

$$H_0: \; E(X_i) \leq E(Y_i) \qquad H_1: \; E(X_i) > E(Y_i)$$

C. Upper-tailed test:

$$H_0: \; E(X_i) \geq E(Y_i) \qquad H_1: \; E(X_i) < E(Y_i)$$

Example 3.4.1 *An item A is manufactured using a certain process. Item B serves the same function as A but is manufactured using a new process. The manufacturer wishes to determine whether B is preferred to A by the consumer, so she selects a random sample consisting of 10 consumers, gives each of them one A and one B, and asks them to use the items for some period of time.*

The sign test (one-tailed) will be used to test

$$H_0: P(+) \leq P(-) \quad H_1: P(+) > P(-)$$

"+": Item B is preferred over item A
"−": Item A is preferred over item B

- T = Number of + signs.
- Critical region corresponds to values of T greater than or equal to $n - t$.
- Consumers report:

$$8 = \text{number of } +'\text{s}$$
$$1 = \text{number of } -'\text{s}$$
$$1 = \text{number of ties}$$

$$\begin{aligned} n &= \text{number of } +\text{'s and } -\text{'s} \\ &= 8 + 1 = 9 \\ T &= \text{number of } +\text{'s} = 8 \end{aligned}$$

- $P(Y \geq 8 = 9 - 1) = P(Y \leq 1) = 0.0195$ from Table A3 with $n = 9$ and $p = 1/2$.
- Reject H_0. □

Example 3.4.2 (Use of the large sample approximation) *In what was perhaps the first published report of a nonparametric test, Arbuthnott (1710) examined the available London birth records of 82 years and for each year compared the number of males born with the number of female born. If for each year we denote the event "more males than females were born" by "+" and the opposite event by "−", (there were no ties), we may consider the hypotheses to be*

$$H_0: P(+) = P(-) \qquad H_1: P(+) \neq P(-)$$

- T = number of + signs.

- Critical region of size $\alpha = 0.05$ corresponds to values of T less than

$$t = .5(82 - (1.960)\sqrt{82}) = 32.1$$

  and values of T greater than $n - t = 82 - 32.1 = 49.9$.

- From the records there were 82 plus signs and no ties.

- H_0 could be have been rejected at an α as small as

$$\begin{aligned}\hat{\alpha} &= P(T = 0) + P(T = 82) \\ &= \left(\frac{1}{2}\right)^{82} + \left(\frac{1}{2}\right)^{82} \\ &= \left(\frac{1}{2}\right)^{81}\end{aligned}$$

□

Example 3.4.3 *Ten homing (歸巢的) pigeons were taken to a point* 25 *kilometers west of their loft and released singly to see whether they dispersed at random in all directions (the null hypothesis) or whether they tended to proceed eastward toward their loft (鴿房). Field glasses were used to observed the birds until they disappeared from view, at which time the angle of the vanishing point was noted. These* 10 *angles are* $20, 35, 350, 120, 85, 345, 80, 320, 280,$ *and* 85 *degrees.*

- $+$: directions more eastward than westward
  $-$: directions away from the loft

$$H_0 : P(+) \leq P(-) \qquad H_1 : P(+) > P(-)$$

- Critical region consists of large values of T, the number of "+" signs.

- From Table A3, for $n = 10$ and $p = 1/2$, the critical region of size $\alpha = 0.0547$ corresponds to values of T greater than or equal to $10 - 2 = 8$.

- Reject H_0 since $T = 9$.

- p-value $P(T \geq 9) = 0.0107$. □

Theory Omitting ties:

- $H_0 : P(+) = P(-) \equiv H_0 : P(+) = 1/2$

- The binomial test procedure is used with $p^* = 1/2$.

- When the sign test is used with the original sets A, B, and C of hypotheses, it is unbiased and consistent (Hemelrijk, 1952).

- Power functions graphed in Fig. 2.4.4 in that example are power functions for the sign test.

- If, in addition to the assumptions in the sign test, we can also assume legitimately that the differences $Y_i - X_i$ are random variables with a symmetric distribution, the is more appropriate.

- Furthermore, if the differences $Y_i - X_i$ are independent and identically distributed normal random variables, the appropriate parametric test is called the paired t test.

- Data that occur naturally in pairs are usually analyzed by reducing the sequence of pairs to a sequence of single values, and then the data are analyzed as if only one sample were involved.

3.5 Some variations of the sign test

- Suppose that the data are not ordinal as in the sign test but nominal, with two categories "0" and "1."

- Data: $(X_i, Y_i),\ X_i, Y_i = 0, 1$

- Can we detect a difference between the probability of $(0,1)$ and the probability of $(1,0)$?

- Such a question arises when X_i in the pair (X_i, Y_i) represents the condition of the subject before the experiment and Y_i represents the condition of the same subject after the experiment.

The McNemar test for significance of changes

Data The data consist of observations on n' independent bivariate random sample $(X_1, Y_1), \ldots, (X_{n'}, Y_{n'})$. The measurement scale for X_i and Y_i is nominal with two categories, which we call "0" and "1"; that is, the possible values of (X_i, Y_i) are $(0,0), (0,1), (1,0)$, and $(1,1)$. In the McNemar test the data are usually summarized in a 2×2 contingency table, as follows.

	$Y_i = 0$	$Y_i = 1$
$X_i = 0$	$a = \#$ of $(X_i, Y_i) = (0,0)$	$b = \#$ of $(X_i, Y_i) = (0,1)$
$X_i = 1$	$c = \#$ of $(X_i, Y_i) = (1,0)$	$d = \#$ of $(X_i, Y_i) = (1,1)$

Assumptions

1. The pairs (X_i, Y_i) are mutually independent.
2. The measurement scale is nominal with two categories for all X_i and Y_i.
3. The difference $P(X_i = 0, Y_i = 1) - P(X_i = 1, Y_i = 0)$ is negative for all i, or zero for all i, or positive for all i.

Test statistic $T_1 = \frac{(b-c)^2}{b+c}$ and $T_2 = b$ if $b + c \leq 20$.

- ▷ Neither T_1 nor T_2 depends on a or d.

Null distribution The null distribution of T_1 is approximately the chi-squared distribution with 1 degree of freedom when $(b + c)$ is large. The exact distribution of T_2 is the binomial distribution with $p = 1/2$ and $n = b + c$.

Hypotheses

Two-tailed test:

$$\begin{array}{lll} H_0: & P(X_i = 0, Y_i = 1) = P(X_i = 1, Y_i = 0) & \text{for all } i \\ H_1: & P(X_i = 0, Y_i = 1) \neq P(X_i = 1, Y_i = 0) & \text{for all } i \end{array}$$

- For $n = b + c \leq 20$, use Table A3 with the proper value of n and $p = 1/2$.
- $\alpha_1 = \alpha/2$: $P(Y \leq t) = \alpha_1$
- Reject H_0 if T_2 is less than or equal to t or if T_2 is greater than $n - t$. Otherwise accept the null hypothesis.
- If $n > 20$ uses T_1 and Table A2.
- The p-value is twice the smaller of the probabilities that Y is less than or equal to the observed value of $T_1(T_2)$, or greater or equal to the observed value of $T_1(T_2)$.

Add $P(X_i = 0, Y_i = 0)$ to both sides of H_0 and H_1 to get

$$\begin{array}{lll} H_0: & P(X_i = 0) = P(Y_i = 0) & \text{for all } i \\ H_1: & P(X_i = 0) \neq P(Y_i = 0) & \text{for all } i \end{array}$$

Add $P(X_i = 1, Y_i = 1)$ to both sides of H_0 and H_1 to get

$$\begin{array}{lll} H_0: & P(X_i = 1) = P(Y_i = 1) & \text{for all i} \\ H_1: & P(X_i = 1) \neq P(Y_i = 1) & \text{for all i} \end{array}$$

The latter sets of hypotheses are usually easier to interpret in terms of the experiment.

- ▶ Let $n = b + c$. If $n \leq 20$, use Table A3.
- ▶ If α is the desired level of significance, enter Table A3 with $n = b + c$ and $p = 1/2$ to find the table entry approximately equal to $\alpha/2$. Call this entry α_1, and the corresponding value of y is called t.
- ▶ Reject H_0 if $T_2 \leq t$, or if $T_2 \geq n - t$, at a level of significance of $2\alpha_1$.
- ▶ Otherwise accept H_0. The p-value is twice the probability of T_2 being less than or equal to the observed value, or greater than or equal to the observed value, whichever is smaller. The probabilities are found from Table A3 using $p = 1/2$ and $n = b + c$.
- ▶ If n exceeds 20, use T_1, and Table A2. Reject H_0 at a level of significance a if T_1 exceeds the $(1 - \alpha)$ quantile of a chi-squared random variable with 1 degree of freedom.
- ▶ Otherwise accept H_0. The p-value is the probability of T_1 exceeding the observed value, as found in Table A2 for the chi-squared distribution with 1 degree of freedom.
- ▶ A more precise p-value can be found by comparing the negative square root of T_1 with Table A1, and doubling the lower-tailed probability.

Example 3.5.1 *Prior to nationally televised debate between the two presidential candidates, a random sample of* 100 *persons stated their choice of candidates as follows. Eighty-four persons favored the Democratic candidate, and the remaining* 16 *favored the Republican. After the debate the same* 100 *people expressed their preference again. Of the persons who formerly favored the Democrat, exact one-fourth of them changed their minds, and also one-fourth of the people formerly favoring the Republican switched to the Democratic side.*

		After		Total
		Democrat	Republican	before
Before	Democrat	63	21	84
	Republican	4	12	16
				100

H_0: The population voting alignment was not altered by the debate
H_1: There has been a change in the proportion of all voters who favor the Democrat.

- X_i is 0 if the ith person favored the Democrat before or 1 if the Republican was favored before.
- Y_i represents the choice of the ith person after the debate.
- $T_1 = \frac{(b-c)^2}{b+c} = \frac{(21-4)^2}{21+4} = 11.56$
- The critical region of size $\alpha = 0.05$ corresponds to all values of T_1 greater than 3.814, the 0.95 quantile of a chi-squared random variable with 1 degree of freedom, obtained from Table A2.
- Reject H_0 since $11.56 > 3.841$. □

Theory

- Variation of the sign test.
- $(0,1) =$ “ + ”, $(1,0) =$ “ − ” and $(1,1), (0,0)$ ties.
- $H_0 : P(+) = P(-)$
- Critical region for T_2 is just as in the sign test for $n \leq 20$.
- For $n > 20$,

$$\begin{aligned} Z &= \frac{T_2 - n/2}{\sqrt{n(1/2)(1/2)}} = \frac{b - n/2}{\sqrt{n}/2} \\ &= \frac{b-c}{\sqrt{b+c}} \\ T_1 &= Z^2 \end{aligned}$$

- Another modification of the sign test is one introduced by Cox and Stuart (1955), it is used to test for the presence of trend.
- A sequence of numbers is said to have trend if the later numbers in the sequence tend to be greater than the earlier numbers (upward trend) or less than the earlier numbers (downward trend).

Cox and Stuart test for trend

Data $X_1, X_2, \ldots, X_{n'}$ are grouped into pairs $(X_1, X_{1+c}), (X_2, X_{2+c}), \ldots,$ $(X_{n'-c}, X_{n'})$ where $c = n'/2$ if n' is even, and $c = (n'+1)/2$ if n' is odd. Replace each pair (X_i, X_{i+c}) with a ”+” if $X_i < X_{i+c}$, or a ”−” if $X_i > X_{i+c}$, eliminating ties. The total number of untied pairs is called n.

Assumptions

1. The random variables $X_1, X_2, \ldots, X_{n'}$ are mutually independent.
2. The measurement scale of the X_i is at least ordinal.
3. Either the X_i's are identically distributed or there is a trend; that is, the later random variables are likely to be greater than instead of less than the earlier random variables.

Test statistic

$$T = \text{total number of } +\text{'s}.$$

Null distribution $T \sim B(n, 1/2)$

Example 3.5.2 *The total annual precipitation is recorded each year for 19 years, and this record is examined to see if the amount of precipitation is tending to increase or decrease. The precipitation in inches was* $45.25, 45.83, 41.77, 36.26, 45.37, 52.25, 35.37, 57.16, 35.37, 58.32, 41.05, 33.72, 45.73, 37.90, 41.72, 36.07, 49.83, 36.24,$ *and* 39.90. *Because* $n' = 19$ *is odd, the middle number* 58.32 *is omitted.*

- $n = 9$ and T equals the number of pairs in which the second number exceeds the first number.
- The critical region of size 0.039 corresponds to values of T less than or equal to 1 and values of T greater than or equal to 8.
- Accept H_0 since the observed $T = 4$. □

Sufficient assumptions:

- The bivariate random variables (X_i, X_{i+c}) are mutually independent.
- The probabilities $P(X_i < X_{i+c})$ and $P(X_i > X_{i+c})$ have the same relative size for all pairs.
- Each pair (X_i, X_{i+c}) may be judged to be a $+$, a $-$, or a tie.

Example 3.5.3 *On a certain stream (溪流) the average rate of water discharge is recorded each month (in cubic feet per second) for a period of 24 months.*

H_0: *The rate of discharge is not decreasing*
H_1: *The rate of discharge is decreasing*

The rate of discharge is known to follow a yearly cycle, so that nothing is learned by pairing stream discharges for two different months. However, by pairing the same months in two successive years the existence of a trend can be investigated. The following data were collected.

Month	First year	Second year	Month	First year	Second year
Jan	14.6	14.2	Jul	92.8	88.1
Feb	12.2	10.5	Aug	74.4	80.0
Mar	104.0	123.0	Sep	75.4	75.6
Apr	220.0	190.0	Oct	51.7	48.8
May	110.0	138.0	Nov	29.3	27.1
Jun	86.0	98.1	Dec	16.0	15.7

- $T = 5$: Number of pairs where the second year had a higher discharge than the first year.
- Critical region of size 0.073 corresponds to all values of $T \leq 3$ (Table A3, $n = 12, p = 1/2$).
- p-value $P(T \leq 5 | H_0$ is true$) = 0.3872$ which is too large to be an acceptable α. □

Sign test is used to detect correlation:

- The test involves arranging the pairs (the pairs remain intact) so that one member of the pair (usually the variable with the fewer ties, which may be either the first member or second) is arranged in increasing order.
- If there is correlation the other member of the pair will exhibit a trend, upward if the correlation is positive, and downward if the correlation is negative.
- The Cox and Stuart test for trend may be used on the sequence formed by the other member of the pair.

Example 3.5.4 *Cochran (1937) compares the reactions of several patients with each of two drugs, to see if there is a positive correlation between the two reactions for each patient.*

Patient	Drug 1	Drug 2	Patient	Drug 1	Drug 2
1	$+0.7$	$+1.9$	6	$+3.4$	$+4.4$
2	-1.6	$+0.8$	7	$+3.7$	$+5.5$
3	-0.2	$+1.1$	8	$+0.8$	$+1.6$
4	-1.2	$+0.1$	9	0.0	$+4.6$
5	-0.1	-0.1	10	$+2.0$	$+3.4$

Ordering the pairs according to the reaction from drug 1 gives

Patient	Drug 1	Drug 2	Patient	Drug 1	Drug 2
2	-1.6	$+0.8$	1	$+0.7$	$+1.9$
4	-1.2	$+0.1$	8	$+0.8$	$+1.6$
3	-0.2	$+1.1$	10	$+2.0$	$+3.4$
5	-0.1	-0.1	6	$+3.4$	$+4.4$
9	0.0	$+4.6$	7	$+3.7$	$+5.5$

- One-tailed Cox and Stuart test for trend is applied to the newly arranged sequence of observations on drug 2.
- Resulting pairs: $(+0.8, +1.9), (+0.1, +1.6)$, $(+1.1, +3.4), (-0.1, +4.4)$, and $(+4.6, +5.5)$.

$$\begin{array}{ll} H_0: & \text{There is no positive correlation} \\ H_1: & \text{There is positive correlation} \end{array}$$

- $T = 5$
- Critical region of size 0.0312 (Table A3 for $n = 5, p = 1/2$, and hence $t = 0$).
- Reject H_0. □

Sign test is used to test for the presence of a predicted pattern:

Example 3.5.5 *The number of eggs laid by a group of insects in a laboratory is counted on an hourly basis during a 24-hour experiment, to test*

> H_0: *The 24 eggs counts constitute observations on 24 identically distributed random variables*
>
> H_1: *The number of eggs laid tends to be a minimum at 2:15 pm increasing to a maximum at 2:15 am and decreasing again until 2:15 pm.* □

Time	Number	Time	Number	Time	Number
9 A.M.	151	5 P.M.	83	1 A.M.	286
10 A.M.	119	6 P.M.	166	2 A.M.	235
11 A.M.	146	7 P.M.	143	3 A.M.	223
Noon	111	8 P.M.	116	4 A.M.	176
1 P.M.	63	9 P.M.	163	5 A.M.	176
2 P.M.	84	10 P.M.	208	6 A.M.	174
3 P.M.	60	11 P.M.	283	7 A.M.	139
4 P.M.	109	Midnight	296	8 A.M.	137

If the alternative hypothesis is true, the egg counts nearest 2:15 P.M. should tend to be the smallest and those nearest 2:15 A.M. should tend to be largest. Therefore the number of eggs is rearranged according to times, from the time nearest 2:15 P.M. to the times nearest 2:15 A.M.

Time	Number	Time	Number
2 P.M.	84	8 A.M.	137
3 P.M.	60	9 A.M.	163
1 P.M.	63	7 A.M.	139
4 P.M.	109	10 A.M.	208
Noon	111	6 A.M.	174
5 P.M.	83	11 P.M.	283
11 A.M.	146	5 A.M.	176
6 P.M.	166	Midnight	296
10 A.M.	119	4 A.M.	176
7 P.M.	143	1 A.M.	286
9 A.M.	151	3 A.M.	223
8 P.M.	116	2 A.M.	235

- If H_1 is true these numbers should exhibit an upward trend.
- Cox and Stuart one-tailed test for trend is used.
- $T = 12$
- Critical region of size 0.0193 for $T \geq 12 - 2 = 10$ (Table A3, $n = 12, p = 1/2$).
- Reject H_0.
- p-value: $P(T \geq 12) = 0.0002$.

Theory

- The Cox and Stuart test for trend is an obvious modification of the sign test and the distribution of the test statistic when H_0 is true is binomial.

- The test is unbiased and consistent when the first sets A, B, and C of hypotheses are being used.

3.6 Summary

1. *Binomial distribution:* the probability of obtaining exactly k heads. 57

2. *Binomial test:* The binomial test may be adapted to test almost any hypothesis, with almost any type of data amenable to statistical analysis. 58

3. *Interval estimator:* (L, U) where L and U are the lower and upper interval boundary respectively. 61

4. *Confidence coefficient:* The probability that the unknown population parameter lies within its interval estimates. 61

5. *Confidence interval:* the interval estimator together with the confidence coefficient. 61

Chapter 4

Contingency Tables

Contents

Preliminary remarks

Contingency table: An array of natural numbers in matrix form where those natural numbers represent counts, or frequencies.

An entomologist observing insects:

1×3 *contingency table*:

Moths	Grasshoppers	Others	Total
12	22	3	37

The entomologist may wish to be more specific and use a 2×3 contingency table, as follows.

2×3 *contingency table*:

	Moths	Grasshoppers	Others	Total
Alive	3	21	3	27
Dead	9	1	0	10
Total	12	22	3	37

- *Two row (three column) totals* and *grand total*, are optional and are usually included only for the reader's convenience.
- may be extended to $r \times c$ contingency table.
- Contingency table with three or more dimensions also may occur.

4.1 The 2×2 contingency table

- $r = 2$ and $c = 2$ (*fourfold* contingency table)
- Arise when N objects (persons), possibly selected at random from some population, are classified into one of two categories before a treatment is applied or an event takes place. After the treatment is applied is applied the same N objects are examined and classified into two categories.
- The question to be answered is, "Does the treatment significantly alter the proportion of objects in each of the two categories?"
- *McNemar test* (one random sample, Sec. 3.5)
  - Able to detect subtle differences, primarily because the same sample is used in the two situations.
- *Chi-squared test* (two random samples)
  - Is the proportion of the population with characteristic A the same for both populations?

The chi-squared test for differences in probabilities, 2×2

Data A random sample of n_1 observations is drawn from one population and each observation is classified into either class 1 or class 2, the total numbers in the two classes being O_{11} and O_{12}, respectively, where $O_{11} + O_{12} = n_1$. A second random sample of n_2 observations is drawn from a second population and the number of observations in class 1 or class 2 is O_{11} or O_{12}, respectively, where $O_{21} + O_{22} = n_2$.

	Class 1	Class 2	Total
Population 1	O_{11}	O_{12}	n_1
Population 2	O_{21}	O_{22}	n_2
Total	C_1	C_2	$N = n_1 + n_2$

Assumptions

1. Each sample is a random sample.
2. The two samples are mutually independent.
3. Each observation may be categorized either into class 1 or class 2.

Test statistic If any column total is zero, the test statistic is defined as $T_1 = 0$. Otherwise,

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}}$$

Null distribution The exact distribution of T_1 is difficult to tabulate because of all the different combinations of values possible for O_{11}, O_{12}, O_{21}, and O_{22}. Therefore the large sample approximation is used, which is the standard normal distribution whose quantiles are given in Table A1.

Hypotheses Let the probability that a randomly selected element will be in class 1 be denoted by p_1 in population 1 and p_2 in population 2. Note that it is not necessary for p_1 and p_2 to be known. The hypotheses merely specify a relationship between them.

A. Two-tailed test: $H_0 : p_1 = p_2$ v.s. $H_1 : p_1 \neq p_2$. Reject H_0 at the approximate level α if T_1 is less than the $z_{\alpha/2}$ or greater than $z_{1-\alpha/2}$. The p-value is twice the smaller of the probabilities that Z is less than the observed value of T_1 or greater than the observed value of T_1.

B. Lower-tailed test: $H_0 : p_1 \geq p_2$ v.s. $H_1 : p_1 < p_2$. Reject H_0 at the approximate level α if T_1 is less than the $z_{\alpha/2}$. The p-value is the probability that Z is less than the observed value of T_1.

C. Upper-tailed test: $H_0 : p_1 \leq p_2$ v.s. $H_1 : p_1 > p_2$. Reject H_0 at the approximate level α if T_1 is greater than the $z_{1-\alpha/2}$. The p-value is the probability that Z is greater than the observed value of T_1.

Example 4.1.1 (Two-tailed Chi-squares test) *Two carloads of manufactured items are samples randomly to determine if the proportion of defective items is different for the two carloads. From the first carload 13 of the 86 items were defective. From the second carload 17 of the 74 items were considered defective.*

	Defective	*Nondefective*	*Totals*
Carload 1	*13*	*73*	*86*
Carload 2	*17*	*57*	*74*
Totals	*30*	*130*	*160*

- Two-tailed test is used.
- H_0 : The proportion of defectives is equal in the two carloads using the test statistic:

$$\begin{aligned} T_1 &= \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}} \\ &= \frac{\sqrt{160}((13)(57) - (73)(17))}{\sqrt{(86)(74)(30)(130)}} \\ &= -1.2695 \end{aligned}$$

- Accept H_0 since $z_{0.975} = 1.96$.
- p-value: $2P(Z < -1.2695) = 0.102$
- Therefore the decision to accept H_0 seems to be a fairly safe one. □

Example 4.1.2 (One-tailed Chi-squares test) *At the U.S. Naval Academy a new lighting system was installed throughout the midshipmen's (見習軍官) living quarters. It was claimed that the new light system resulted in poor eyesight due to a continual strain on the eyes of the midshipmen. Consider a study to test the null hypothesis,*

> H_0 : *The probability of a graduating midshipman having 20-20 (good) vision is the same or greater under the new lighting system, than it was under the old lighting system.*
>
> H_1 : *The probability of (good) vision is less now than it was.*

- p_1 (p_2): Probability that a randomly selected graduating midshipman had good vision under the old (new) lighting system.
- $H_0 : p_1 \leq p_2$ v.s. $H_1 : p_1 > p_2$

	Good vision	Poor vision	Totals
Old lights	$O_{11} = 714$	$O_{12} = 111$	$n_1 = 825$
New lights	$O_{21} = 662$	$O_{22} = 154$	$n_2 = 816$
Totals	1376	265	$N = 1641$

$$\begin{aligned} T_1 &= \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}} \\ &= 2.982 \end{aligned}$$

- Reject H_0 since $z_{0.95} = 1.645$.
- p-value: $P(Z > 2.982) = 0.002$
- Conclude that the populations represented by the two graduation classes do differ with respect to the proportions having poor eyesight, and in the direction predicted. □

Theory

- The 2×2 contingency table just presented is actually a special case of the $r \times c$ contingency table. The exact distribution of the test statistic is difficult to find unless r and c are very small.
- Exact probability distribution of T_1 when $H_0 : p_1 = p_2 = p$.

$$\begin{aligned} P((x_1, n_1 - x_1)|\text{Population 1}) &= \binom{n_1}{x_1} p^{x_1}(1-p)^{n_1 - x_1} \\ P((x_2, n_2 - x_2)|\text{Population 2}) &= \binom{n_2}{x_2} p^{x_2}(1-p)^{n_2 - x_2} \end{aligned}$$

- Two samples are independent

$$P\left(\begin{pmatrix} x_1, & n_1 - x_1 \\ x_2, & n_2 - x_2 \end{pmatrix} \middle| \begin{pmatrix} \text{Population 1} \\ \text{Population 2} \end{pmatrix}\right) = \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1+x_2}(1-p)^{N-x_1-x_2}$$

- $n_1 = 2$ and $n_2 = 2$ there are nine different points in the sample space:

Tables		$(p = 1/2)$	$(p = 1)$	T_1
$\begin{pmatrix} 2 & 0 \\ 2 & 0 \end{pmatrix}$	p^4	1/16	1	Undefined
$\begin{pmatrix} 2 & 0 \\ 1 & 1 \end{pmatrix}$	$2p^3(1-p)$	1/8	0	1.1547
$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$	$p^2(1-p)^2$	1/16	0	2.0000

$\begin{pmatrix}1 & 1\\ 2 & 0\end{pmatrix}$	$2p^3(1-p)$	$1/8$	0	-1.1547
$\begin{pmatrix}1 & 1\\ 1 & 1\end{pmatrix}$	$4p^2(1-p)^2$	$1/4$	0	0
$\begin{pmatrix}1 & 1\\ 0 & 2\end{pmatrix}$	$2p(1-p)^3$	$1/8$	0	1.1547
$\begin{pmatrix}0 & 2\\ 2 & 0\end{pmatrix}$	$p^2(1-p)^2$	$1/16$	0	-2.0000
$\begin{pmatrix}0 & 2\\ 1 & 1\end{pmatrix}$	$2p(1-p)^3$	$1/8$	0	-1.1547
$\begin{pmatrix}0 & 2\\ 0 & 2\end{pmatrix}$	$(1-p)^4$	$1/16$	1	Undefined

- The undefined values for T_1 arise from the indeterminate form $0/0$.

- However, since the two outcomes that result in undefined values for T_1 are strongly indicative that H_0 is true, just as the fifth outcome is strongly indicative that T_1 is true, we may arbitrarily define T_1 to be 0 for the first and last outcomes in agreement with the fifth outcome.

- Then T_1 has the following probability distribution.

$p = 1/2$	$p = 1$
$P(T_1 = -2) = 1/16$	$P(T_1 = 0) = 1$
$P(T_1 = -1.1547) = 1/4$	
$P(T_1 = 0) = 3/8$	
$P(T_1 = 1.1547) = 1/4$	
$P(T_1 = 2) = 1/16$	

- Similarly for any sample sizes n_1, and n_2, the exact probability distributions may be found after the appropriate defining of the undefined values of T_1.

- Normal approximation for large sample
  - $E(O_{11}/n_1 - O_{21}/n_2) = p_1 - p_2$
  - $\mathrm{Var}\left(\frac{O_{11}}{n_1} - \frac{O_{21}}{n_2}\right) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$
  - $\hat{p} = C_1/N$ and $\hat{q} = C_2/N$
  - O_{11}/n_1 and $O_{21}/n_2 \approx$ independent normal distribution

$$T_1 = \frac{O_{11}/n_1 - O_{21}/n_2}{\sqrt{\frac{C_1 C_2}{N^2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \approx N(0,1) \qquad \square$$

- Another use for 2×2 contingency table appears when each observation in a single sample of size N is classified according to two properties, where each property may take one of two forms.

- This use of 2×2 contingency table is a special case of $r \times c$ contingency table and does not have any special variation (such as the one-sided test of this section, see Section 2).

- The primary difference between this type of contingency table and the first one is that in this contingency table the row totals are random variables whose values are unknown until after the data are examined.

- In the first table the row totals represented sample sizes for the two samples, which are known prior to the examination of the data and are therefore not random. In both tables the column totals are random variables.

- The third type of contingency table is one with nonrandom row and column totals.

- That is, both row totals and both column totals are known prior to an examination of the data.

- This situation does not occur as often as the first two types of contingency tables, but the following statistical procedure is often employed, no matter which of the three types of contingency tables actually occurs, because the exact p-value can be determined fairly easily.

- The procedure was developed almost simultaneously in the mid 1930's by R. A. Fisher (1935), J. 0. Irwin (1935), and F. Yates (1934). It is widely known as Fisher's exact test.

  Data types:

  1. Row totals are fixed and column totals are random variables.
  2. Row totals and column totals are random variables.
  3. Row totals and column totals are known prior to an examination of the data.

- No matter which of the three types of contingency tables actually occurs, the Fisher's exact test is often employed.

  - Simultaneously developed in the mid 1930's by Fisher (1935), Irwin (1935) and Yates (1934).

Fisher's exact test

Data The N observations in the data are summarized in a 2×2 contingency table as in the previous test, except both of the row totals, r and $N - r$, and both of the column totals, c and $N - c$, are determined beforehand, and are therefore fixed, not random.

	Col 1	Col 2	
Row 1	x	$r - x$	r
Row 2	$c - x$	$N - r - c + x$	$N - r$
	c	$N - c$	N

Assumptions

1. Each observation is classified into exactly one cell.
2. The row and column totals are fixed, not random.

Test statistic The test statistic T_2 is the number of observations in the cell in row 1, column 1.

Null distribution The exact distribution of T_2 when H_0 is true is given by the hypergeometric distribution:

$$P(T_2 = x) = \frac{\binom{r}{x}\binom{N-r}{c-x}}{\binom{N}{c}}, \quad x = 0, 1, \ldots, \min(r, c) \tag{6}$$

For a large sample approximation use

$$T_3 = \frac{T_2 - \frac{rc}{N}}{\sqrt{\frac{rc(N-r)(N-c)}{N^2(N-1)}}} \approx N(0,1)$$

Hypotheses Let p_1 be the probability of an observation in row 1 being classified into column 1, and let p_2 be the corresponding probability for row 2.

A. *Two-tailed test*: $H_0 : p_1 = p_2$ v.s. $H_1 : p_1 \neq p_2$. First find the p-value using Equation 6. The p-value is twice the smaller of $P(T_2 \leq t_{\text{obs}})$ or $P(T_2 \geq t_{\text{obs}})$. Reject H_0 at the level of significance α if the p-value is $\leq \alpha$.

B. *Lower-tailed test*: $H_0 : p_1 \geq p_2$ v.s. $H_1 : p_1 < p_2$. Find the p-value $P(T_2 \leq t_{\text{obs}})$ using Equation 6. Reject H_0 at the level of significance α if $P(T_2 \leq t_{\text{obs}})$ is $\leq \alpha$.

C. *Upper-tailed test*: $H_0 : p_1 \leq p_2$ v.s. $H_1 : p_1 > p_2$. Find the p-value $P(T_2 \geq t_{\text{obs}})$ using Equation 6. Reject H_0 at the level of significance α if $P(T_2 \geq t_{\text{obs}})$ is $\leq \alpha$.

Comment

- Valid for contingency tables with random row totals, random column totals, or both.
  - This exact test finds the p-value for one subset of the sample space, the one with the given row and column totals.
  - Each different set of row and column totals represents another mutually exclusive subset, thus partitioning the entire sample space into mutually exclusive subsets.
  - If the critical region in each regions has an unconditional probability $\leq \alpha$ under H_0, then the union of all critical regions has an unconditional probability $\leq \alpha$ under H_0, and the test is valid.
- The power of this exact test is usually less than the power of a more appropriate, approximate, test in those cases where row totals, or column totals, or both, are random.

Comment (continuity correction)

The large sample approximation for T_3 is improved by using a continuity correction. That is, for lower-tailed probabilities, add 0.5 to the numerator of T_3 before looking up the p-value in Table 1. For upper-tailed probabilities subtract 0.5 from the numerator.

Example 4.1.3 *Fourteen newly hired business majors, 10 males and 4 females, all equally qualified, are being assigned by the bank president to their new jobs. Ten of the new jobs are as tellers, and four are as account representatives. The null hypothesis is that males and females have equal chances at getting more desirable account representative jobs. The one-sided alternative of interest is that females are more likely than males to get the account representative jobs.*

Only one female is assigned a teller position. Can the null hypothesis be rejected?

	Account Representative	Teller	
Males	1	9	10
Females	3	1	4
	4	10	14

- $H_0 : p_1 \geq p_2$ v.s. $H_1 : p_1 < p_2$
- Exact lower-tailed p-value by Equation 6:

$$\begin{aligned} P(T_2 \leq 1) &= P(T_2 = 0) + P(T_2 = 1) \\ &= \frac{\binom{10}{0}\binom{4}{4}}{\binom{14}{4}} + \frac{\binom{10}{1}\binom{4}{3}}{\binom{14}{4}} \\ &= 0.041 \end{aligned}$$

- Reject H_0 at $\alpha = 0.05$.

Comment
Compare the exact p-value in Example 3 with the exact p-value if the column totals were random.

- The exact p-value using Eq. 4 ($p = 0.3$) is $0.012 < 0.041$ (Problem 4.1.3).
- Normal approximation for $T_1 = -2.4321$ is about 0.008 which is close to the true p-value.
- This illustrates that Fisher's exact test is exact only if the row and column totals are nonrandom.

Theory Show that T_2 has the hypergeometric distribution:
Proof.

- A contingency table with fixed row totals, whose probability is given

$$\binom{r}{x}\binom{N-r}{c-x} p^c (1-p)^{N-c}. \tag{8}$$

- The probability of getting the column totals c and $N - c$:

$$P((c, N-c)) = \binom{N}{c} p^c (1-p)^{N-c} \tag{9}$$

- Conditional probability of getting the table results: (8)/(9).
- The large sample normal approximation is obtained by subtracting the mean and dividing by the standard deviation of the hypergeometric distribution to obtain T_3. □

- Combine results of several 2 × 2 contingency tables into one overall analysis.

- This situation occurs when the overall experiment consists of several smaller experiments conducted in various environments, the common probability p under H_0 may be different from one environment to another, and each sub-experiment results in its own 2×2 contingency table.

- One method for combining several 2×2 contingency tables was presented by Mantel and Haenszel (1959).

The Mantel-Haenszel test (1959)

Data The data are summarized in several 2×2 contingency tables, each with nonrandom row and column totals. Let the number of tables be $k \geq 2$, and let the ith table be represented with the following notation.

	Col 1	Col 2	Total
Row 1	x_i	$r_i - x_i$	r_i
Row 2	$c_i - x_i$	$N_i - r_i - c_i + x_i$	$N_i - r_i$
Total	c_i	$N_i - c_i$	N_i

Assumptions

1. The assumptions for each contingency table are the same as for the Fisher exact test.
2. The several contingency tables are obtained from independent experiments.

Test statistic

$$T_4 = \frac{\sum x_i - \sum \frac{r_i c_i}{N_i}}{\sqrt{\sum \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{N_i^2 (N_i - 1)}}}$$

Null distribution

- $T_4 \approx N(0,1)$ when H_0 is true.
- The exact probabilities are improved by using a continuity correction (±0.5).
- The resulting probabilities will be more accurate in most cases.

Hypotheses Let p_{1i} be the probability of an observation in row 1 being classified into column 1, in the ith contingency table, and let p_{2i} be the corresponding probability for row 2.

A. *Two-tailed test*:
$H_0 : p_{1i} = p_{2i}$ v.s. $H_1 : p_{1i} \neq p_{2i}$
Reject H_0 at the level of significance α if $T_4 > z_{1-\alpha/2}$ or $T_4 < z_{\alpha/2}$.
The p-value is $2\min(P(Z \leq T_4), P(Z \geq T_4))$.

B. *Lower-tailed test*:
$H_0 : p_{1i} \geq p_{2i}$ v.s. $H_1 : p_{1i} < p_{2i}$
Reject H_0 at the level of significance α if $T_4 < z_\alpha$.
The p-value is $P(Z \leq T_4)$.

C. *Upper-tailed test*:
$H_0 : p_{1i} \leq p_{2i}$ v.s. $H_1 : p_{1i} > p_{2i}$
Reject H_0 at the level of significance α if $T_4 > z_{1-\alpha}$.
The p-value is $P(Z \geq T_4)$.

Comment

- This test is valid even though the row totals or column totals are random.

- However, in that case it is more accurate to use the test statistic

$$T_5 = \frac{\sum x_i - \sum \frac{r_i c_i}{N_i}}{\sqrt{\sum \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{N_i^3}}}$$

  instead of T_4.

- *The continuity correction should not be used to find p-values when T_5 is used.*

Example 4.1.4 *From Li, Simon, and Gart (1979), three groups of cancer patients are given an experimental treatment to see if the success rate is improved. The numbers of successes and failures are summarized as follows.*

	Group 1		*Group 2*		*Group 3*	
	Success	*Failure*	*Success*	*Failure*	*Success*	*Failure*
Treatment	*10*	*1*	*9*	*0*	*8*	*0*
Control	*12*	*1*	*11*	*1*	*7*	*3*

- The upper-tailed test is used:

$$\begin{aligned} T_4 &= \frac{(10+9+8) - \left(\frac{11 \cdot 22}{24} + \frac{9 \cdot 20}{21} + \frac{8 \cdot 15}{18}\right)}{\sqrt{\frac{11 \cdot 22 \cdot 13 \cdot 2}{24^2 \cdot 23} + \frac{9 \cdot 20 \cdot 12 \cdot 1}{21^2 \cdot 20} + \frac{8 \cdot 15 \cdot 10 \cdot 3}{18^2 \cdot 17}}} \\ &= 1.4323 \end{aligned}$$

- Accept H_0 at size 5% since $T_4 = 1.4323 < z_{0.95} = 1.6449$.

- The corrected upper-tailed p-value is

$$P(T_4 \geq 1.0057 \text{ (corrected)}) = 0.157.$$

- Treat as random column totals.
  - $T_5 = 1.4690$ and the upper-tailed p-value 0.071.
  - The greater power associated with using the more appropriate T_5 in the case of random row totals or column totals, or both.
  - The null hypothesis is still accepted at $\alpha = .05$. □

Theory

- Numerator of T_4 is the sum of Fisher exact test statistic T_2 in each table.
- As in T_3, the mean is subtracted, and the result is divided by the standard deviation.
- By CLT, $T_4 \approx N(0,1)$.
- The statistic T_5 is obtained by rearranging T_1 to look like T_3, and noting that the only difference between T_1 and T_3 is in the term N instead of $N-1$ in the denominator.
- $T_5 \approx N(0,1)$. □

Confidence intervals may be performed for any unknown probabilities associated with the 2×2 contingency table or any contingency table, for that matter, by applying the procedure described in Section 3.1.

4.2 The $r \times c$ contingency table

Three applications:

- Present a tabulation of the data contained in several samples, where the data represent at least a nominal scale of measurement, and to test the hypothesis that the probabilities do not differ from sample to sample.
- Another use for the $r \times c$ contingency table is with the single sample, each element may be classified into one of r different categories according to one criterion and, at the same time, into one of c different categories according to a second criterion.
- A third application will also be discussed.

r samples with c categories, random column sums:

- Because of the many rows and columns, the one-sided hypotheses of the previous section are no longer appropriate.
- Two-sided hypothesis will be considered.
- Test statistic is the square of T_1 generalized to the $r \times c$ case.

The Chi-squared test for differences in probabilities, $r \times c$

Data There are r populations in all, and one random sample is drawn from each population. Let n_i represent the number of observations in the ith sample. Each observation in each sample may be classified into one of c different categories. Let O_{ij} be the number of observations from ith sample that fall into category j, so

$$n_i = \sum_j O_{ij}$$

	Class 1	Class 2	$\cdots$	Class c	Totals
Population 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Population 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Population r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
Totals	C_1	C_2	$\cdots$	C_c	N

$$
\begin{aligned}
N &= n_1 + n_2 + \cdots + n_r \\
C_j &= O_{1j} + O_{2j} + \cdots + O_{rj}
\end{aligned}
$$

Assumptions

1. Each sample is a random sample.
2. The outcomes of the various samples are mutually independent.
3. Each observation may be categorized into exactly one of the c categories or classes.

Test statistic

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \sum_j \frac{O_{ij}^2}{E_{ij}} - N$$

$E_{ij} = n_i C_j / N$ represents the expected number of observations in cell (i, j), if H_0 is true.

Null distribution

- $T \approx \chi^2_{(r-1)(c-1)}$
- Very difficult to find the exact distribution of T.
- Chi-squared approximation is satisfactory if the E_{ij}s in the test statistic are not too small.
- Chi-squared approximation appears to be satisfactory in most cases if all E_{ij}s are greater than 0.5 and at least half are greater than 1.0.

Hypotheses p_{ij}: Probability of a randomly selected value from ith population being classified in the jth class.

H_0: $p_{1j} = p_{2j} = \cdots = p_{rj}$ for all j
H_1: At least two of the probabilities in the same column are not equal to each other.

Comment

- The approximate value of α is a good approximation to the true value of α if the E_{ij} are fairly large.
- If some of E_{ij}s are small, the approximation may be poor.
- Cochran (1952) states that if any E_{ij} is less than 1 or if more than 20% of the E_{ij}s are less than 5, the approximation may be poor.

- Cochran's conclusion seems to be overly conservative according to unpublished studies by various researchers.
- If some of the E_{ij}s are too small, several categories should be combined to eliminate the E_{ij}s that are too small.

Example 4.2.1 *A sample of students randomly selected from private high schools and a sample of students randomly selected from public high schools were given standardized achievement tests with the following results.*

	Test scores				
	0-275	*276-350*	*351-425*	*426-500*	*Totals*
Private	*6*	*14*	*17*	*9*	*46*
Public	*30*	*32*	*17*	*3*	*82*
Totals	*36*	*46*	*34*	*12*	*128*

- To test the null hypothesis that the distribution of test scores is the same for private and public high school students.
- $(r-1)(c-1) = (2-1)(4-1) = 3$
- E_{ij}:

	Column			
	1	2	3	4
Row 1	12.9	16.5	12.2	4.3
Row 2	23.1	29.5	21.8	7.7

- $\frac{(O_{11}-E_{11})^2}{E_{11}} = \frac{(6-12.9)^2}{12.9} = 3.69$
- $T = 3.69 + 0.38 + 1.89 + 5.14 + 2.06 + 0.21 + 1.06 + 2.87 = 17.3$
- Reject H_0 at size 0.05 since $T > \chi^2_{3,.95} = 7.815$.
- p-value is approximately .001.
- Test scores are distributed differently among public and private high school students. □

- In Example 1 the data possessed at least an ordinal scale of measurement.
- If the alternative hypothesis of interest was that students from the private schools tended to score higher (or lower) than that students from the public schools, then a more powerful test based on ranks could have been used, such as the Mann-Whitney test.
- The alternative hypothesis in this example included differences of all types, such as higher scores, lower scores, a smaller variance in scores, a larger variance in scores, and so forth, so this Chi-squared test is more appropriate.

Theory

- ▶ The exact distribution of T in the $r \times c$ case may be found in exactly the same way as it was found in the previous section for the 2×2 case.
- ▶ The row totals are held constant, and then all possible contingency tables having those same row totals are listed, and their probabilities are calculated using the multinomial distribution for each row.
- ▶ In all three applications of contingency tables in this section, the asymptotic distribution of T is the same, $\chi^2_{(r-1)(c-1)}$. □

The second application of the $r \times c$ contingency table involves a single random sample of size N, where each observation may be classified according to two criteria.

The Chi-squared test for independence

Data A random sample of size N is obtained. The observations in the random sample may be classified according to two criteria. Using the first criterion each observation is associated with one of the r rows, and using the second criterion each observation is associated with one of the c columns. Let O_{ij} be the number of observations associated with row i and column j simultaneously.

	1	2	$\cdots$	c	Totals
Row 1	O_{11}	O_{12}	$\cdots$	O_{1c}	R_1
Row 2	O_{21}	O_{22}	$\cdots$	O_{2c}	R_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Row r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	R_r
Totals	C_1	C_2	$\cdots$	C_c	N

Assumptions

1. The sample of N observations is a random sample.
2. Each observation may be categorized into exactly one of the r different categories according to one criterion and into exactly one of c different categories according to a second criterion.

Test statistic

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \sum_j \frac{O_{ij}^2}{E_{ij}} - N$$

$E_{ij} = R_i C_j / N$ represents the expected number of observations in cell (i, j), if H_0 is true.

Null distribution

- ▷ $T \approx \chi^2_{(r-1)(c-1)}$
- ▷ Very difficult to find the exact distribution of T.
- ▷ Chi-squared approximation is satisfactory if the E_{ij}s in the test statistic are not too small.
- ▷ Chi-squared approximation appears to be satisfactory in most cases if all E_{ij}s are greater than 0.5 and at least half are greater than 1.0.

Hypotheses

$$\begin{aligned} H_0: &\quad P(\text{row } i, \text{ column } j) = P(\text{row } i)P(\text{column } j) \text{ for all } i, j \\ H_1: &\quad P(\text{row } i, \text{ column } j) \neq P(\text{row } i)P(\text{column } j) \text{ for some } i, j \end{aligned}$$

Reject H_0 if T exceeds $\chi^2_{(r-1)(c-1),1-\alpha}$.

Example 4.2.2 *A random sample of students at a certain university were classified according to the college in which they were enrolled and also according to whether they graduated from a high school in the state or out of state. The results were put into a* 2×4 *contingency table.*

		Arts and	*Home*		
	Engineering	*Sciences*	*Economics*	*Other*	*Totals*
In state	*16*	*14*	*13*	*13*	*56*
Out of state	*14*	*6*	*10*	*8*	*38*
Totals	*30*	*20*	*23*	*21*	*94*

- To test the null hypothesis that the college in which each student is enrolled is independent of whether high school training was is in state or out state, the Chi-squared test for independence is selected.
- $(r-1)(c-1) = 3$
- $T = 1.52$
- Accept H_0 at size 0.05 since $T < \chi^2_{3,.95} = 7.815$.
- p-value > 0.25. □

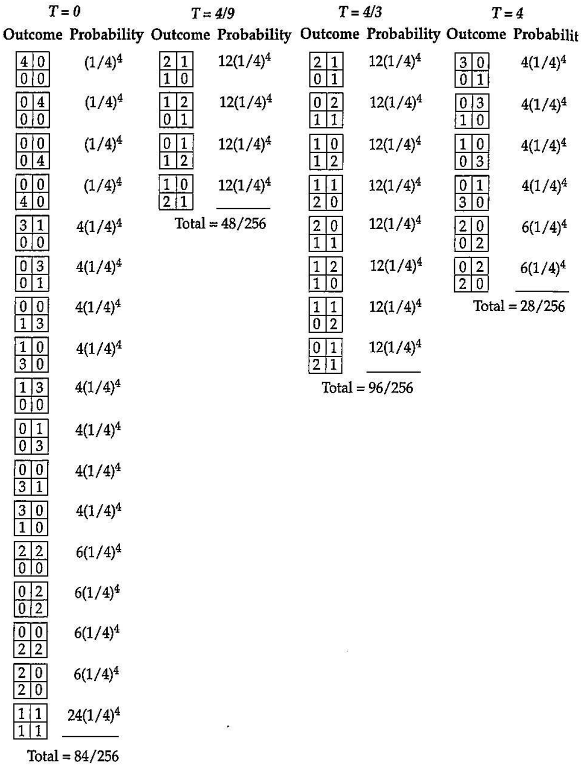
Theory Exact distribution of T with $N = 4$

- Let p_{ij} be the probability of an observation being classified in row i and column j (cell i, j).
- Then the probability of the particular outcome

$$\begin{array}{r|c|c|l} & \multicolumn{2}{c}{\text{Column}} & \\ & \multicolumn{1}{c}{1} & \multicolumn{1}{c}{2} & \\ \cline{2-3} \text{Row } 1 & a & b & \\ \cline{2-3} \text{Row } 2 & c & d & \\ \cline{2-3} \multicolumn{3}{c}{} & N \end{array}$$

- Multinomial distribution: $\frac{N!}{a!b!c!d!} \, (p_{11})^a (p_{12})^b (p_{21})^c (p_{22})^d$
- The maximum size of the upper tail of T when H_0 is true, is found by setting all of the p_{ij}s equal to each other: $\frac{N!}{a!b!c!d!} \left(\frac{1}{4}\right)^N$
- There are $\binom{7}{3} = 35$ $(a + b + c + d = 4)$ different contingency tables (Fig. 1).
- $P(T = 0) = 84/256 = 0.33$

- $P(T = 4/9) = 48/256 = 0.19$
- $P(T = 4/3) = 96/256 = 0.37$
- $P(T = 4) = 28/256 = 0.11$
- The distribution of T is more complicated to obtain than the row totals are fixed. □



$T=0$		$T=4/9$		$T=4/3$		$T=4$	
Outcome	Probability	Outcome	Probability	Outcome	Probability	Outcome	Probabilit
4 0 0 0	$(1/4)^4$	2 1 1 0	$12(1/4)^4$	2 1 0 1	$12(1/4)^4$	3 0 0 1	$4(1/4)^4$
0 4 0 0	$(1/4)^4$	1 2 0 1	$12(1/4)^4$	0 2 1 1	$12(1/4)^4$	0 3 1 0	$4(1/4)^4$
0 0 0 4	$(1/4)^4$	0 1 1 2	$12(1/4)^4$	1 0 1 2	$12(1/4)^4$	1 0 0 3	$4(1/4)^4$
0 0 4 0	$(1/4)^4$	1 0 2 1	$12(1/4)^4$	1 1 2 0	$12(1/4)^4$	0 1 3 0	$4(1/4)^4$
3 1 0 0	$4(1/4)^4$	Total = 48/256		2 0 1 1	$12(1/4)^4$	2 0 0 2	$6(1/4)^4$
0 3 0 1	$4(1/4)^4$			1 2 1 0	$12(1/4)^4$	0 2 2 0	$6(1/4)^4$
0 0 1 3	$4(1/4)^4$			1 1 0 2	$12(1/4)^4$	Total = 28/256	
1 0 3 0	$4(1/4)^4$			0 1 2 1	$12(1/4)^4$		
1 3 0 0	$4(1/4)^4$			Total = 96/256			
0 1 0 3	$4(1/4)^4$						
0 0 3 1	$4(1/4)^4$						
3 0 1 0	$4(1/4)^4$						
2 2 0 0	$6(1/4)^4$						
0 2 0 2	$6(1/4)^4$						
0 0 2 2	$6(1/4)^4$						
2 0 2 0	$6(1/4)^4$						
1 1 1 1	$24(1/4)^4$						
Total = 84/256							

Figure 4.1: *The exact distribution of T, when all p_is equal $1/4$*

The third application of the contingency table is the row totals and column totals are fixed.

- The exact distribution of T is easier to find than in both applications previously introduced.
- The exact distribution is still too complicated for practical purposes.
- The chi-squared approximation is recommended for finding the critical region and α.

The Chi-squared test with fixed marginal totals

Data The data are summarized in an $r \times c$ contingency table:

	1	2	$\cdots$	c	Totals
Row 1	O_{11}	O_{12}	$\cdots$	O_{1c}	n_1
Row 2	O_{21}	O_{22}	$\cdots$	O_{2c}	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Row r	O_{r1}	O_{r2}	$\cdots$	O_{rc}	n_r
Totals	c_1	c_2	$\cdots$	c_c	N

Assumptions

1. Each observation is classified into exactly one cell.
2. The observations are observations on a random sample. Each observation has the same probability of being classified into cell (i, j) as any other observation.
3. The row and column totals are given, not random.

Test statistic

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \sum_j \frac{O_{ij}^2}{E_{ij}} - N$$

$E_{ij} = n_i c_j / N$ represents the expected number of observations in cell (i, j), if H_0 is true.

Null distribution

- ▷ $T \approx \chi^2_{(r-1)(c-1)}$
- ▷ Very difficult to find the exact distribution of T.
- ▷ Chi-squared approximation is satisfactory if the E_{ij}s in the test statistic are not too small.
- ▷ Chi-squared approximation appears to be satisfactory in most cases if all E_{ij}s are greater than 0.5 and at least half are greater than 1.0.

Hypotheses Variations of the independence hypotheses of the previous test. Reject H_0 if T exceeds $\chi^2_{(r-1)(c-1),1-\alpha}$.

Example 4.2.3 *The chi-squared test with fixed marginal totals may be used to test the hypothesis that two random variables X and Y are independent. Starting with a scatter diagram of 24 points, which represent independent observations on the bivariate random variable (X, Y), a contingency table may be constructed. The x-coordinate of each point is the observed value of X and the y-coordinate is the observed value of Y in each observation (X, Y). Assume the observed pairs (X, Y) are mutually independent. We wish to test*

$$H_0 : X \textit{ and } Y \textit{ are independent of each other}$$

against the alternative hypothesis of dependence.

- ▶ To form the contingency table so that all E_{ij} are equal, we note that 3 and 4 both are factors of the sample size 24.
- ▶ Divide the points into 3 rows of 8 points each, and 4 columns of 6 points each, using dotted lines as in Figure 2.
- ▶ One way of accomplishing this is by having equal row totals and equal column totals.)
- ▶ $(r - 1)(c - 1) = 6$

	1	2	3	4	Totals
Row 1	0	4	4	0	8
Row 2	2	1	2	3	8
Row 3	4	1	0	3	8
Totals	6	6	6	6	24

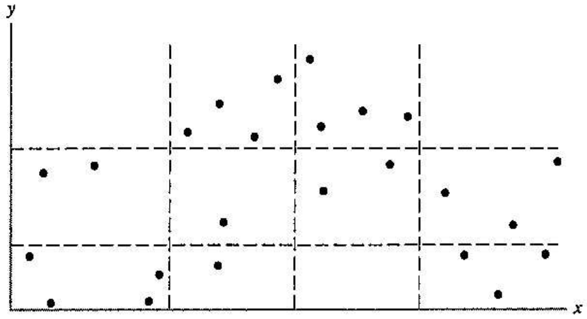


Figure 4.2: *Random points*

- $E_{ij} = (6)(8)/24 = 2$ and $T = 14$
- Reject H_0 at size 0.05 since $T > \chi^2_{6,.95} = 12.59$.
- p-value is 0.03.
- Conclude that X and Y are not independent. □

Example 4.2.4 *A psychologist asks a subject to learn 25 words. The subject is given 25 blue cards, each with one word on it. Five of the words are nouns, 5 are adjectives, 5 are adverbs, 5 are verbs, and 5 are prepositions. She must pair these blue cards with 25 white cards, each with one word on it and also containing the different parts of speech, 5 words each. The subject is allowed 5 minutes to pair the cards and 5 minutes to study the pairs thus formed. Then she is asked to close her eyes, and the words on the white cards are read to her one by one. When each word is read to her, she tries to furnish the word on the blue card associated with the word read.*

The psychologist is not interested in the number of correct words but, instead, in examining the pairing structure to see if it represents an ordering of some sort.

H_0 : There is no organization of pairs according to parts of speech

H_1 : The subjects tends to pair particular parts of speech on the blue cards with particular parts of speech on the white cards.

The pairing are summarized in a 5×5 contingency table.

	Noun	Adjective	Adverb	Verb	Preposition	Totals
Noun		3			2	5
Adjective	4	1				5
Adverb				5		5
Verb			5			5
Preposition	1	1			3	5
Totals	5	5	5	5	5	25

- $(r-1)(c-1) = 16$ and $E_{ij} = \frac{(5)(5)}{25} = 1$
- $T = 66$
- Reject H_0 at size 0.05 since $T > \chi^2_{16,.95} = 26.3$.

- p-value < 0.001. □

Theory

- The exact distribution of O_{11} is the hypergeometric distribution for 2×2 case.
- Row totals and column totals are all equal 2.

Table	Probability	T
$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$	$\frac{\binom{2}{2}\binom{2}{0}}{\binom{4}{2}} = 1/6$	4
$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$	$\frac{\binom{2}{1}\binom{2}{1}}{\binom{4}{2}} = 2/3$	0
$\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$	$\frac{\binom{2}{0}\binom{2}{2}}{\binom{4}{2}} = 1/6$	4

- Fixed row totals and fixed column totals greatly reduce the contingency tables possible. When $r = c = 2$, the test is known as "Fisher's exact test".
- For r and c in general, the exact probability of the table with fixed marginal totals is given by

$$\text{probability} = \frac{\binom{n_1}{O_{1i}} \cdots \binom{n_r}{O_{ri}}}{\binom{N}{c_i}}$$

where the multinomial coefficients $\binom{n_m}{O_{mi}} = \frac{n_m!}{O_{m1}!O_{m2}! \cdot O_{mc}!}$. □

Two-way contingency table can be written as

$$T = \sum \frac{\left(O_{ij} - N\frac{R_i}{N}\frac{C_j}{N}\right)^2}{N\frac{R_i}{N}\frac{C_j}{N}}$$

Three (or more) way contingency table test:

- $R_i = \sum O_{ijk}$
- $C_j = \sum O_{ijk}$
- $B_k = \sum O_{ijk}$
- $E_{ijk} = N\frac{R_i}{N}\frac{C_j}{N}\frac{B_k}{N}$
- $T = \sum \frac{(O_{ijk}-E_{ijk})^2}{E_{ijk}} \approx \chi^2_{rct-r-c-t+2}$

So-called "*loglinear models*" have been used successfully to analyze multidimensional contingency tables and are discussed in the final section of this book.
An excellent and readable survey article on contingency tables is one by Mosteller (1968).

4.3 The median test

- Designed to examine whether several samples came from population having the same median.

- A special application of the chi-squared test with fixed marginal totals in the previous section.

- To test whether c populations have the same median, a random sample is drawn from each population.

- A $2 \times c$ contingency table is constructed and the two entities in the ith column are the numbers of observations in the ith sample that are above and below the grand median.

The median test

Data From each of c populations a random sample of size n_i is obtained. The combined sample median is determined; that is, the number that is exceeded by about half of the observations in the entire array of $N = \sum n_i$ sample values is determined. This is called the grand median.

Sample	1	2	$\cdots$	c	Totals
$>$ Median	O_{11}	O_{12}	$\cdots$	O_{1c}	a
$\leq$ Median	O_{21}	O_{22}	$\cdots$	O_{2c}	b
Totals	n_1	n_2	$\cdots$	n_c	N

Assumptions

1. Each sample is a random sample.
2. The samples are independent of each other.
3. The measurement scale is at least ordinal.
4. If all populations have the same median, all populations have the same probability p of an observation exceeding the grand median.

Test statistic Rearrange the test statistic given in Sec. 4.2:

$$T = \frac{N^2}{ab} \sum \frac{(O_{1i} - n_i a/N)^2}{n_i} = \frac{N^2}{ab} \sum \frac{O_{1i}^2}{n_i} - \frac{Na}{b}$$

Note that $(O_{2i} - n_i b/N)^2 = (n_i - O_{1i} - n_i b/N)^2 = (n_i a/N - O_{1i})^2$.

▷ If $a = b$, then $T = \sum \frac{(O_{1i}-O_{2i})^2}{n_i}$.

$$
\begin{aligned}
T &= \sum_{i=1}^{c} \frac{(O_{1i} - an_i/N)^2}{an_i/N} + \frac{(O_{2i} - bn_i/N)^2}{bn_i/N} \\
&= \sum_{i=1}^{c} \frac{(O_{1i} - n_i/2)^2}{n_i/2} + \frac{(O_{2i} - n_i/2)^2}{n_i/2} && (a = b = N/2) \\
&= \sum_{i=1}^{c} \frac{2O_{1i}^2 + 2O_{1i}^2 - n_i^2}{n_i} \\
&= \sum_{i=1}^{c} \frac{(O_{1i} - O_{2i})^2}{n_i} && (n_i = O_{1i} + O_{2i})
\end{aligned}
$$

▷ If $a \approx b$, then $T \approx \sum \frac{(O_{1i}-O_{2i})^2}{n_i}$.

Null distribution $T \approx \chi^2_{c-1}$

Hypotheses

H_0: All c populations have the same median
H_1: At least two of the populations have different medians

▷ Reject H_0 at size α if $T > \chi^2_{c-1,1-\alpha}$.

Multiple comparisons If the null hypothesis is rejected, pairwise multiple comparisons may be made between populations by using median test repeatedly on 2×2 contingency tables.

Example 4.3.1 *Four different methods of growing corn were randomly assigned to a large number of different plots of land and the yield per acre was computed for each plot.*

Method			
1	2	3	4
83	91	101	78
91	90	100	82
94	81	91	81
89	83	93	77
89	84	96	79
96	83	95	81
91	88	94	80
92	91		81
90	89		
	84		

In order to determine whether there is a difference in yields as a result of the method used, the median test was employed because it was felt that a difference in population medians could be interpreted as a difference in the value of the method used. The hypotheses may be stated as follows.

H_0 : All methods have the same median yield per acre.

H_1 : At least two of the methods differ with respect to the median yield per acre.

- The average of 17th and 18th smallest observations is the grand median 89.

	Method				
	1	2	3	4	Totals
> 89	6	3	7	0	16
≤ 89	3	7	0	8	18
Totals	9	10	7	8	34

- $T = 4.01(0.34 + 0.29 + 1.97 + 1.78) = 17.6$ which can be also approximated by

$$T = \sum_{i=1}^{c} \frac{(O_{1i} - O_{2i})^2}{n_i} = 9/9 + 16/10 + 49/7 + 64/8 = 17.6$$

since $a \approx b$.

- Reject H_0 at size .05 since $T > \chi^2_{3,.95} = 7.815$.
- p-value is slightly less than 0.001.
- Multiple comparisons for 2×2 contingency tables: $\chi^2_{1,.95} = 3.841$

Methods	Median	T
1 and 2	89	2.55*
1 and 3	92.5	6.35
1 and 4	83	13.43
2 and 3	91	13.25
2 and 4	82.5	14.40
3 and 4	82	15.00

□

Multiple comparisons of the populations by repeatedly using the same test on subgroups of the original data always distorts the true level of significance of all tests but the first.
Median test v.s. one-way ANOVA

- In Example 1 the experiment has been arranged in a so-called *complete randomized design*.
- The usual parametric method of analyzing the data is called a one-way analysis of variance.
- For normal distributions the A.R.E. is only $2/\pi = 64\%$.
- For double exponential distributions the A.R.E. is only 200%.

The median test may be extended to become a “quantile test” for testing the null hypothesis that several populations have the same quantile.

Theory

$$P\left(\begin{pmatrix} O_{11} & O_{12} & \cdots & O_{1c} \\ O_{21} & O_{22} & \cdots & O_{2c} \end{pmatrix}\right) = \binom{n_1}{O_{11}}\binom{n_2}{O_{12}}\cdots\binom{n_c}{O_{1c}} p^a(1-p)^b \qquad (A)$$

$$P((a,b)|N) = \binom{N}{a} p^a (1-p)^b \qquad (B)$$

$(A) \div (B)$

$$P\left(\begin{pmatrix} O_{11} & O_{12} & \cdots & O_{1c} & a \\ O_{21} & O_{22} & \cdots & O_{2c} & b \\ n_1 & n_2 & \cdots & n_c & N \end{pmatrix}\right) = \frac{\binom{n_1}{O_{11}}\binom{n_2}{O_{12}}\cdots\binom{n_c}{O_{1c}}}{\binom{N}{a}}$$

which is the same as

$$\text{Probability} = \frac{\left[\binom{a}{O_{1i}}\right]\left[\binom{b}{O_{2i}}\right]}{\left[\binom{N}{n_i}\right]}$$

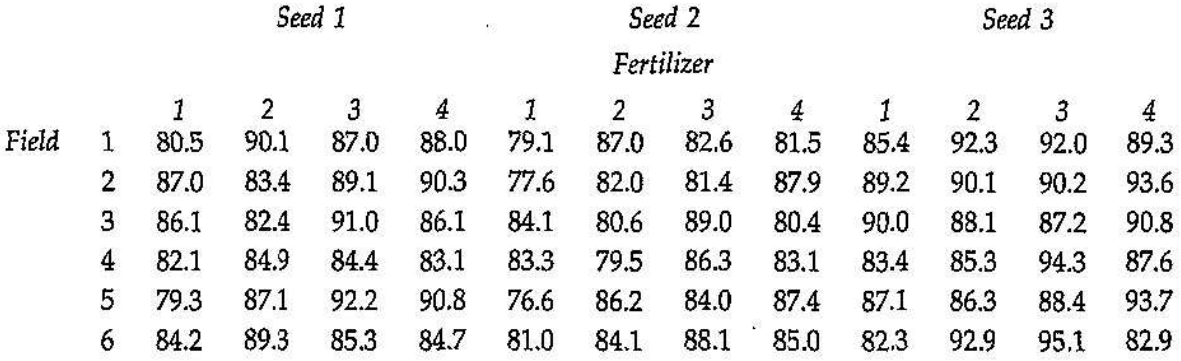
which is the same as

$$\text{Probability} = \frac{\left[\binom{a}{O_{1i}}\right]\left[\binom{b}{O_{2i}}\right]}{\left[\binom{N}{n_i}\right]} \qquad \square$$

An extension of the median test

Example 4.3.2 *Four different fertilizers are used on each of six different fields, and the entire experiment is replicated using three different types of seed. The yield per acre is calculated at the conclusion of the experiment under each of 72 different conditions with the following results. To test the null hypothesis*

H_0 : *There is no difference in median yields due to the different fertilizers*



		Seed 1				Seed 2				Seed 3			
		Fertilizer											
		1	2	3	4	1	2	3	4	1	2	3	4
Field	1	80.5	90.1	87.0	88.0	79.1	87.0	82.6	81.5	85.4	92.3	92.0	89.3
	2	87.0	83.4	89.1	90.3	77.6	82.0	81.4	87.9	89.2	90.1	90.2	93.6
	3	86.1	82.4	91.0	86.1	84.1	80.6	89.0	80.4	90.0	88.1	87.2	90.8
	4	82.1	84.9	84.4	83.1	83.3	79.5	86.3	83.1	83.4	85.3	94.3	87.6
	5	79.3	87.1	92.2	90.8	76.6	86.2	84.0	87.4	87.1	86.3	88.4	93.7
	6	84.2	89.3	85.3	84.7	81.0	84.1	88.1	85.0	82.3	92.9	95.1	82.9

Figure 4.3: Yield per acre for different fertilizers, fields and seeds

- $x_{i_1 i_2 i_3}$: observed yield using fertilizer i_1 in field i_2 with seed i_3.
- For example, x_{213} is the yield using fertilizer 2 in field 1 with seed 3, which is 92.3.

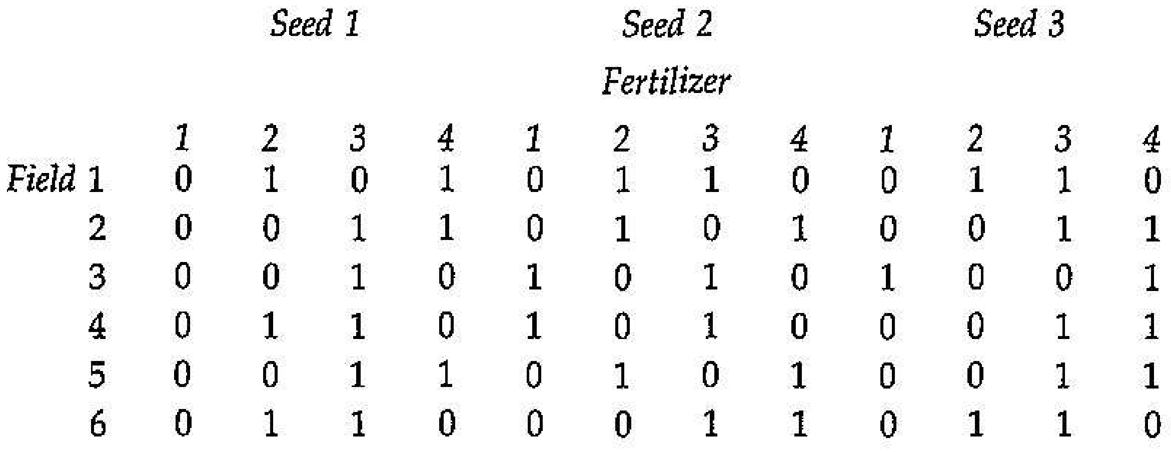
- Then x_{213} s compared with the median of $x_{113}, x_{213}, x_{313}$ and x_{413}, the four yields obtained under identical circumstances except for fertilizers.

- Thus x_{213} is compared with the median of $85.4, 92.3, 92.0$, and 89.3, which is $(89.3 + 92.0)/2 = 90.65$.

- If x_{213} exceeds 90.65, it is replaced in the table by a one; otherwise it is replaced by a zero.

- Each yield is compared with the median of the yields in the same row (field) and same block (seed) and replaced by one or zero according to whether it exceeds or does not exceed its respective median.

- The results are as follows.



	Seed 1				Seed 2				Seed 3			
	Fertilizer											
	1	2	3	4	1	2	3	4	1	2	3	4
Field 1	0	1	0	1	0	1	1	0	0	1	1	0
2	0	0	1	1	0	1	0	1	0	0	1	1
3	0	0	1	0	1	0	1	0	1	0	0	1
4	0	1	1	0	1	0	1	0	0	0	1	1
5	0	0	1	1	0	1	0	1	0	0	1	1
6	0	1	1	0	0	0	1	1	0	1	1	0

Figure 4.4: *Results for different fertilizers, fields and seeds*

- O_j: number of fields in which fertilizer j was used and where the yield exceeded ("ones") its respective median.

	Fertilizer				Total
	1	2	3	4	Totals
$O_j = \#$ of 1	3	8	14	10	35
$O_j = \#$ of 0	15	10	4	8	37
	18	18	18	18	72

- Using Eq. (3) $T = \frac{(144+4+100+4)}{18} = 14$

- Rejection H_0 at size .05 since $T > \chi^2_{3,.95} = 7.815$.

- p-value is about .004. □

4.4 Measures of dependence

- The *contingency table* is a convenient form for examining data to see if there is some sort of dependence inherent in the data.

- Particular type of dependence: Row-column dependence.

- Express the degree of dependence in some simple form that easily conveys to other people the exact degree of dependence exhibited by the table.
- Statistic: $T = \sum\sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
- If it is good enough to test for dependence, it is good enough to measure dependence.

Example 4.4.1 *In Example 4.2.1*

	Scores				
	0-275	*276-350*	*351-425*	*426-500*	*Totals*
Private	*6*	*14*	*17*	*9*	*46*
Public	*30*	*32*	*17*	*3*	*82*
Totals	*36*	*46*	*34*	*12*	*128*

- $T = 17.3$ corresponds to the 0.999 quantile of a chi-squared random variable with 3 degrees of freedom.
- p-value $= 1 - p = .001$
- Strongly reject H_0.
- Do little toward measuring the level of dependence. □

Cramér's contingency coefficient

- Provide an easily interpreted measure of dependence.
- Divide T by the maximum value of T.
- $\max T = N(q-1)$ when there are zeros in every cell except for one cell in each row and each column where $q = \min(r, c)$.

	1	2	$\cdots$	q	$\cdots$	c	Totals
Row 1	n_1	0	$\cdots$	0	$\cdots$	0	n_1
Row 2	0	n_2	$\cdots$	0	$\cdots$	0	n_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
Row r	0	0	$\cdots$	n_q	$\cdots$	0	n_q
Totals	n_1	n_2	$\cdots$	n_q	$\cdots$	0	N

- $R_1 = \frac{T}{\max T} = \frac{T}{N(q-1)}$
- R_1 is close to 1 if the table indicates a strong row-column dependence and close to 1 if the numbers across each row are in the same proportions to each other as the column totals are to each other.
- *Cramér's coefficient*

$$\sqrt{R_1} = \sqrt{\frac{T}{N(q-1)}}$$

  (*SAS, StatXact*)

  The most widely used measure of dependence for $r \times c$ contingency tables.

Example 4.4.2 *In the previous example the* 2×4 *contingency table,* $T = 17.3$.

- $N = 128, q = 2$ give that

$$R_1 = \frac{T}{N(q-1)} = \frac{17.3}{128} = .135$$

- *Cramér's coefficient* is $\sqrt{0.135} = 0.368$ □

- *Cramér's coefficient*, like all good measures of dependence, is "*scale invariant*."

	Scores				
	0-275	276-350	351-425	426-500	Totals
Private	60	140	170	90	460
Public	300	320	170	30	820
Totals	360	460	340	120	1280

Pearson's contingency coefficient

- *Pearson's coefficient* of mean square contingency (Yule and Kendall, 1950):

$$R_2 = \sqrt{\frac{T}{N+T}}$$

- $\max R_2 = \sqrt{(q-1)/q}$ when $T = N(q-1)$.
- $0 \leq R_2 \leq \sqrt{(q-1)/q} < 1.0$
- R_2 is also called the *contingency coefficient* by McNemar and Siegel.

Example 4.4.3 *In the contingency table of the two previous examples we have* $T = 17.3$ *and* $N = 128$,

$$R_2 = \sqrt{\frac{T}{N+T}} = \sqrt{\frac{17.3}{128+17.3}} = 0.345$$ □

Pearson's mean-square contingency coefficient (Yule and Kendall, 1950)

- Mean-square contingency:

$$R_3 = \frac{T}{N}$$

- $0 \leq R_3 \leq q-1$

Example 4.4.4 *For the same contingency table used in the previous example we have*

$$R_3 = \frac{17.3}{128} = 0.135$$ □

Tschuprow's coefficient

$$R_4 = \sqrt{\frac{T}{N\sqrt{(r-1)(c-1)}}}$$

The choice of a measure of dependence is largely a personal decision.

For 2×2 contingency table

	Column		
	1	2	
Row 1	a	b	r_1
Row 2	c	d	r_2
	c_1	c_2	N

We know from Problem 4.2.2 that

$$T = \frac{N(ad-bc)^2}{r_1 r_2 c_1 c_2}$$

R_1 and R_3 reduce to

$$R_1 = R_3 = \frac{T}{N} = \frac{(ad-bc)^2}{r_1 r_2 c_1 c_2}$$

Cramér's coefficient

$$\sqrt{R_1} = \sqrt{\frac{(ad-bc)^2}{r_1 r_2 c_1 c_2}}$$

$$R_2 = \sqrt{\frac{T}{N+T}} = \sqrt{\frac{(ad-bc)^2}{r_1 r_2 c_1 c_2 + (ad-bc)^2}}$$

In a four-fold contingency table, unlike the general $r \times c$ contingency table, it is meaningful to distinguish between a positive association and a negative association.

Example 4.4.5 *Forty children are classified according to whether their mothers have dark hair or light hair and as to whether their fathers have dark or light hair.*

The results may show a positive association

		Father		
		Dark	Light	
Mother	Dark	28	0	28
	Light	5	7	12
		33	7	40

or a negative association

		Father		
		Dark	Light	
Mother	Dark	21	7	28
	Light	12	0	12
		33	7	40

according to whether $ad - bc$ is positive or negative.

A lack of association (zero correlation)

		Father		
		Dark	Light	
Mother	Dark	23	5	28
	Light	10	2	12
		33	7	40

□

One measure of association that preserves direction:
The phi coefficient

$$R_5 = \frac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}}$$

- Set up the table so that a and d represent the number of similar classifications while b and c represent the number of unlike classifications.
- One measure of association that preserve direction is the *phi coefficient R_5*.
- $-1 \leq R_5 \leq 1$.
- One special case of *Pearson product moment correlation coefficient*.
- $R_5 = T_1/\sqrt{N}$

Example 4.4.6 *For the first table in Example 5 we have*

$$\begin{array}{ll} a = 28 & r_1 = 28 \\ b = 0 & r_2 = 12 \\ c = 5 & c_1 = 33 \\ d = 7 & c_2 = 7 \end{array}$$

$$\begin{aligned} R_5 &= \frac{ad - bc}{\sqrt{r_1 r_2 c_1 c_2}} = \frac{(28)(7) - 0}{\sqrt{(28)(12)(33)(7)}} \\ &= 0.703 \end{aligned}$$

- For the second table in Example 5.

$$\begin{aligned} R_5 &= \frac{(21)(0) - (7)(12)}{\sqrt{(28)(12)(33)(7)}} \\ &= -0.302 \end{aligned}$$

which reflects the negative association of hairtypes. □

Other measures of association for 2 × 2 contingency table:

- $R_6 = \frac{ad - bc}{ad + bc}$ (Yule and Kendall, 1950)
- $R_7 = \frac{(a+d)-(b+c)}{a+b+c+d}$ (Gibbons, 1967)

Test the null hypothesis of independence

- R_1, R_2, R_3, R_4 and Cramér's coefficient are not appropriate for testing null hypothesis of independence since their values will all be too large whenever T is too large.
- A one-tailed test, appropriate only for the 2×2 contingency table may be based on R_5.
- $\sqrt{N}R_5 \approx N(\mu, \sigma^2)$
- Reject H_0: There is no positive (negative) correlation if $\sqrt{N}R_5$ too large (small).

Example 4.4.7 *In order to see if seat belts help prevent fatalities, records of the last* 100 *automobile accidents to occur along a high way were examined. These* 100 *accidents involved* 242 *persons. Each person was classified as using or not using seat belts when the accident occurred and as injured fatally or a survivor.*

		Injured Fatally?		
		Yes	No	Totals
Wearing	Yes	7	89	96
Seat Belts?	No	24	122	146
	Totals	31	211	242

The statement we wish to prove is, "Seat belts help prevent fatalities." However, a test for correlation does not automatically imply a cause and effect relationship. While a cause and effect relationship between two variables usually results in correlation, a significant correlation may be the result of both variables being influenced by a third variable, which might be the reckless nature of the driver in this case. Therefore the null hypothesis is

H_0 : There is no negative correlation between wearing seat belts and being killed in an automobile accident

H_1 : There is a negative correlation between wearing seat belts and being killed in an automobile accident

- Reject H_0 since $\sqrt{N}R_5 = -2.0829 < -1.645$.
- p-value is about 0.019. □

4.5 The chi-squared goodness-of-fit test

- Often the hypotheses being tested are statements concerning the unknown probability distribution of the random variable being observed.
- Examples:
  - The median is 4.0.
  - The probability of being in class 1 is the same for both populations.
  - The unknown distribution function is the normal distribution function with mean 3 and variance 1.
  - The distribution function of this random variable is the binomial, with parameters $n = 10$ and $p = 0.2$.

- *Goodness-of-fit test*: A test designed to compare the sample obtained with the hypothesized distribution to see if the hypothesized distribution function fits the data in the sample.

- The oldest and best-known goodness-of-fit test is the chi-squared test for goodness of fit, first presented by Pearson (1900).

The Chi-squared test for goodness of fit

Data The data consist of N independent observations of a random variable X. These N observations are grouped into c classes,

	Class				
	1	2	$\cdots$	c	Total
Frequencies	O_1	O_2	$\cdots$	O_c	N

Assumptions

1. The sample is a random sample.
2. The measurement scale is at least nominal.

Test statistic

$$T = \sum \frac{(O_j - E_j)^2}{E_j} = \sum \frac{O_j^2}{E_j} - N$$

- p_j^*: Probability of a random observation on X being in class j.
- $E_j = p_j^* N$

Null distribution $T \approx \chi^2_{c-1}$

Hypotheses

H_0: $P(X \text{ is in class } j) = p_j^*$

H_1: $P(X \text{ is in class } j) \neq p_j^*$ for at least one class

Reject H_0 if $T > \chi^2_{c-1,1-\alpha}$.

Comment

- If some of the E_js are small, the asymptotic chi-squared distribution may not be appropriate.

- *Cochran (1952)* suggests that none of the E_js should be less than 1 and no more 20% should be smaller than 5.

- *Yarnold (1970)* says, "If the number of classes s is 3 or more, and if r denotes the number of expectations less than 5, then the minimum expectation may be as small as $5r/s$."

- *Slakter (1973)* feels that the number of classes can exceed the number of observations, which means the average expected value can be less than 1.

- *Koehler and Larntz (1980)* finds the chi-squared approximation to be adequate as long as $N \geq 10, c \geq 3, N^2/c \geq 10$, and all $E_j \geq 0.25$.

- The user may wish to combine some cells with this discussion in mind if many of the E_js are small.

Example 4.5.1 *A certain computer program is supposed to furnish random digits. If the program is accomplishing its purpose, the computer prints out digits* $(2, 3, 7, 4,$ *etc.) that seem to be observations on independent and identically distributed random variables, where each digit* $0, 1, 2, \ldots, 8, 9$ *is equally likely to be obtained. One way of testing*

H_0 : *The number appear to be random digits*

H_1 : *Some digits are more likely than others*

Three hundred digits are generated with the following results.

Digit:	0	1	2	3	4	5	6	7	8	9	Total
O_j	22	28	41	35	19	25	25	40	30	35	300
E_j	30	30	30	30	30	30	30	30	30	30	300

- Test statistic: $T = \sum_{i=1}^{10} \frac{O_i^2}{E_i} - N = 17$
- Reject H_0 at size 0.05 since $T > \chi^2_{9,0.95} = 16.92$. □

Comment

- If the probability distribution of X is completely specified except for a number k of parameters, it is first necessary to estimate the parameters and then to proceed with the test as just outlined, $T \approx \chi^2_{c-1-k}$.
- The *minimum chi-squared method* (Cramér, 1946) involves using the value of the parameter that results in the smallest value of the test statistic.
- This procedure is impractical.
- The modified minimum chi-squared method (Cramér, 1946) consists of estimating the k unknown parameters by computing the first k sample moments of the grouped data.

Example 4.5.2 *Efron and Morris (1975) presented data on the first* 18 *major baseball players to have* 45 *times at bat in* 1970. *The players' names and the number of hits they got in their* 45 *times at bat are given as follows.*

Clemente	*18*	*Kessinger*	*13*	*Scott*	*10*
F. Robinson	*17*	*L.Alvarado*	*12*	*Petrocelli*	*10*
F. Howard	*16*	*Santo*	*11*	*E. Rodriguez*	*10*
Johnstone	*15*	*Swoboda*	*11*	*Campaneris*	*9*
Berry	*14*	*Unser*	*10*	*Munson*	*8*
Spencer	*14*	*Williams*	*10*	*Alvis*	*7*

- We will test the null hypothesis that these data follow a binomial distribution with $n = 45$. But we need to estimate $p = P(\text{hit})$ for each time at bat.

$$\hat{p} = \frac{\text{total number of hits}}{\text{total number of at-bats}} = \frac{215}{810} = 0.2654$$

- $n = 45$ and $p = 0.2654$,

$$P(X = i) = \binom{45}{i}(0.2654)^i(0.7346)^{45-i}, \quad i = 0, \dots, 45$$

- Expected cell counts: $E_i = 18P(X = i)$
- Cells with expected values less than 0.5 are combined to avoid problems of having a poor approximation by the chi-squared distribution.

	No. of hits								
	≤ 7	8	9	10	11	12	13	14	15
Observed	1	1	1	5	2	1	1	2	1
Expected	1.10	1.06	1.57	2.04	2.35	2.40	2.20	1.82	1.36

- $T = \sum_{i=1}^{12} \frac{O_i^2}{E_i} - N = 24.73 - 18 = 6.73$
- Accept H_0 since $T < \chi^2_{10,0.95} = 18.31$
- p-value is much larger than 0.25. □

Comment

- $\hat{p}$ in Example 2 is a good estimator, but it may not be one that minimizes the value of T.
- The p-value is already much greater than 0.25. There is no need to find the minimum value of T, which will further increase the p-value.
- Asymptotic theory holds as sample size goes to infinity and the expected values in each cell also go to infinity.
- This is no guarantee that the minimum chi-squared method results in a more accurate approximation for small sample sizes.
- The chi-squared goodness-of-fit test can also be used to test whether the data come from a specified continuous distribution.

The chi-squared goodness-of-fit test to a continuous distribution where two parameters are estimated from the data.

Example 4.5.3 *Fifty two-digit numbers were drawn from a telephone book, and the chi-squared test for goodness of fit is used to see if they could have been observations on a normally distributed random variable. The numbers, after being arranged in order from the smallest to the largest, are as follows.*

23	23	24	27	29	31	32	33	33	35
36	37	40	42	43	43	44	45	48	48
54	54	56	57	57	58	58	58	58	59
61	61	62	63	64	65	66	68	68	70
73	73	74	75	77	81	87	89	93	97

- H_0 : These numbers are observations on a normally distributed random variable.

  Step 1.
  Divide the observations into intervals of finite length.

	20-40	40-60	60-80	80-100	Total
O_j	12	18	15	5	50

  Step 2.
  Estimate μ and σ with the sample mean $\bar{X}$ and the sample deviation S of the grouped data.

  $$\bar{X} = 55.2 \qquad S = 18.7$$

  Step 3.
  Using the estimated parameters from Step 2, compute the E_js for the groups in Step 1 and for the "tails".

	< 20	20-40	40-60	60-80	80-100	≥ 100
p_j^*	.03	.18	.39	.31	.08	.01
E_j	1.5	9.0	19.5	15.5	4	0.5
O_j	0	12	18	15	5	0

- After combined small E_js:

	< 40	40-60	60-80	≥ 80
E_j	10.5	19.5	15.5	4.5
O_j	12	18	15	5

  Step 4.
  Compute $T = 0.401$: Accept H_0 since $T < \chi^2_{4-1-2,0.95} = 3.841$.

- Usually a modification called *Sheppard's correction* is used when the variance is being estimated from grouped data and when the interior intervals are of equal width, say h.

- Sheppard's correction consists of subtracting $h^2/12$ from S^2 in order to obtain a better estimate of variance.

- In this example $h = 20$ (the width of each interval), so $(20)^2/12 = 33.33$ could have been subtracted in Step 2 before extracting the square root.

- The result is $S = 17.8$, a smaller estimate for σ.

- This smaller estimate of σ results in a larger value of T in this example and, since our objective is to obtain estimates that give the smallest possible value for T, the correction was not used.

- In most situations we can expect a smaller T when the correction is used.

- Another peculiarity of this example is the fact that a smaller value of $T(0.279)$ may be obtained by using $\overline{X} = 55.04$ and $s = 19.0$ as estimates of μ and σ.

- These estimates are the sample moments obtained from the original observations, before grouping.

- No matter how they are obtained, the estimates to use are the estimates that result in the smallest value of T.
- The procedure described in this example can be relied on to provide a value of T not far from its minimum value in most cases. □

Theory

- The probability of any particular arrangement:

$$P(O_1, \ldots, O_c|N) = \frac{N!}{O_1! \cdots O_c!} \, p_1^{O_1} \cdots p_c^{O_c}$$

- There seems to be no theory developed to find the exact distribution of T when several parameters are first estimated from the sample.
- The large sample approximation is both practical and necessary in order to apply this goodness-of-fit test. □

4.6 Cochran's test for related observations

- Sometimes the use of a treatment, or condition, results in one of two possible outcomes, e.g. "sale" or "no sale", "success" or "failure".
- $2 \times c$ contingency table: several treatments where one row represents the number of successes and the other row represents the number of failures, and the null hypothesis of no treatment differences may be tested using a chi-squared contingency table test, as described in Section 4.2.
- However, it is possible to detect more subtle differences between treatments, that is, increase the power of the test by *applying all c treatments independently to the same blocks*, such as by trying all c sales techniques on each of several persons in an experimental situation and then recording for each person the results of each technique.
- Thus each block, or person, acts as its own control, and the treatments are more effectively compared with each other.
- Such an experimental technique is called ”*blocking*,” and the experimental design is called a *randomized complete block design*.

The Cochran test

Data Each of c treatments is applied independently to each of r blocks, or subjects, and the result of each treatment application is recorded as either 1 or 0, to represent "success" or "failure", or any other dichotomization of the possible treatment results.

	Treatments				
Blocks	1	2	$\cdots$	c	Row totals
1	X_{11}	X_{12}	$\cdots$	X_{1c}	R_1
2	X_{21}	X_{22}	$\cdots$	X_{2c}	R_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
r	X_{r1}	X_{r2}	$\cdots$	X_{rc}	R_r
Column totals	C_1	C_2	$\cdots$	C_c	N

Assumptions

1. The blocks were randomly selected from the population of all possible blocks.
2. The outcomes of the treatments may be dichotomized in a manner common to all treatments within each block, so the outcomes are listed as either "0" or "1."

Test statistic

$$T = c(c-1)\frac{\sum \left(C_j - \frac{N}{c}\right)^2}{\sum R_i(c-R_i)} = \frac{c(c-1)\sum C_j^2 - (c-1)N^2}{cN - \sum R_i^2}$$

Null distribution $T \approx \chi^2_{c-1}$

Hypotheses

H_0: The treatments are equally effective
H_1: There is a difference in effectiveness among treatments

Multiple comparison If H_0 is rejected, pairwise comparisons may be made between treatments using the McNemar test, which is the two-tailed sign test, as described in Sec. 3.5.

Example 4.6.1 *Each of three basketball had devised his own system for predicting the outcomes of collegiate basketball games. Twelve games were selected at random, and each sportsman presented a prediction of the outcome of each game. After the games were played, the results were tabulated, using* 1 *for successful prediction and* 0 *for unsuccessful prediction.*

	Sportsman			
Game	1	2	3	Totals
1	1	1	1	3
2	1	1	1	3
3	0	1	0	1
4	1	1	0	2
5	0	0	0	0
6	1	1	1	3
7	1	1	1	3
8	1	1	0	2
9	0	0	1	1
10	0	1	0	1
11	1	1	1	3
12	1	1	1	3
Totals	8	10	7	25

The Cochran test was used to test:

H_0 : Each sportsman is equally effective in his ability to predict the outcomes of the basketball games.

- Test statistic:

$$\begin{aligned} T &= c(c-1)\frac{\sum_{j=1}^{c}(C_j - \frac{N}{c})^2}{\sum_{i=1}^{r} R_i(c-R_i)} \\ &= \frac{(3)(2)[(-\frac{1}{3})^2 + (\frac{5}{3})^2 + (-\frac{4}{3})^2]}{2+2+2+2+2} \\ &= 2.8 \end{aligned}$$

- Accept H_0 at size 0.05 since $T < \chi^2_{2,0.95} = 5.99$.
- p-value is about 0.25. □

Theory

- $X_{ij} \sim B(1,p)$ where p is the same within each row under H_0, but can be different from block to block.
- From CLT: $\frac{C_j - E(C_j)}{\sqrt{\text{Var}(C_j)}} \approx N(0,1)$ where $C_j = \sum_{i=1}^{r} X_{ij}$.
- $\sum \left[\frac{C_j - E(C_j)}{\sqrt{\text{Var}(C_j)}}\right]^2 \approx \chi^2_c$
- $\hat{E}(C_j) = \frac{1}{c}\sum C_j = \frac{N}{c}$
- $\text{Var}(C_j) = \sum \text{Var}(X_{ij}) = \sum p(1-p)$
- $\hat{p}$ in row $i = R_i/c$

- Estimate of $\text{Var}(X_{ij}) = \frac{R_i}{c}\left(1 - \frac{R_i}{c}\right)$
- Improved estimate of $\text{Var}(X_{ij}) = \frac{R_i(c-R_i)}{c(c-1)}$
- Improved estimate of

$$\text{Var}(C_j) = \frac{1}{c(c-1)}\sum_{i=1}^{r} R_i(c-R_i)$$

- Substitution of the estimates for $E(C_j)$ and $\text{Var}(C_j)$ into Eq. (5) gives

$$T = c(c-1)\frac{\sum_{j=1}^{c}(C_j - \frac{N}{c})^2}{\sum_{i=1}^{r} R_i(c-R_i)}$$ □

Comment

If $c = 2$: Cochran test $\equiv$ McNemar test

- If only two treatments are being considered, such as "before" and "after" observations on the same block, with r blocks, the experimental situation is the same as that analyzed by the McNemar test for significance of changes.
- In each situation the null hypothesis is that the proportion of the population in class 1 is the same using treatment 1 (before) as it is using treatment 2 (after).

- For $c = 2$ the Cochran test statistic reduces to

$$\begin{aligned} T &= 2\frac{\left(C_1 - \frac{C_1+C_2}{2}\right)^2 + \left(C_2 - \frac{C_1+C_2}{2}\right)^2}{\sum_{i=1}^{r} R_i(2-R_i)} \\ &= 2\frac{\left(\frac{C_1-C_2}{2}\right)^2 + \left(\frac{C_2-C_1}{2}\right)^2}{\sum_{i=1}^{r} R_i(2-R_i)} \\ &= \frac{(C_1-C_2)^2}{\sum R_i(2-R_i)} \qquad (14) \end{aligned}$$

- If a block has ones in both columns, then $R_i = 2$ and $R_i(2 - R_i) = 0$.
- If both columns have zeros, then $R_i = 0$.
- If there is a change from zero to one or one to zero in a given row, then $R_i = R_i(2 - R_i) = 1$, and $\sum R_i(2 - R_i) = b + c$ is the total number of rows that go from 0 to 1 and 1 to 0.
- $C_1 = c + d$, the total number of ones in column one, or “before”.
- $C_2 = b + d$, the total number of ones in column two, or “after”.
- $C_1 - C_2 = c - b$
- $T = (c - b)^2/(b + c)$ is the McNemar’s test given in Equation 3.5.1.
- Both the McNemar test statistic and the Cochran test statistic with $c = 2$ are approximated by a chi-squared random variable with 1 degree of freedom.

4.7 Some comments on alternative methods of analysis

The likelihood ratio statistic

- Methods in this chapter:

$$T_1 = \sum \frac{(O_i - E_i)^2}{E_i}$$

(*Pearson chi-squared statistic*, 1900, 1922)

- A different method of analysis, called the *likelihood ratio test* and mentioned in Problem 4.2.3, employs the statistic

$$T_2 = 2\sum O_i \ln\left(\frac{O_i}{E_i}\right)$$

(*likelihood ratio chi-squared statistic*, Wilks, 1935, 1938)

- $T_1, T_2 \approx \chi^2_{c-1}$
- The choice of whether to use T_1 or T_2 depends largely on the user’s preference.
- A serious disadvantage in using T_2 is that the chi-squared approximation is usually poor if $N/rc < 5$, while the chi-squared approximation for T_1 holds up for much smaller values of N.

Loglinear models

- ▶ This method works well in analyzing contingency tables with three or more dimensions.
- ▶ The same statistic T_1 and T_2 just given are used with loglinear models; the difference is in the method used for obtaining the E_is. Usually iterative methods are used.
- ▶ The name of loglinear model arises from a two-way contingency table:

$$H_0 : p_{ij} = p_{i+} \cdot p_{+j}, \quad \text{all } i \text{ and } j$$
$$H_0 : \log p_{ij} = \log p_{i+} + \log p_{+j}$$

  where p_{ij} is the probability of an observation being classified in cell (i, j) and where p_{i+} and p_{+j} are the row and column marginal probabilities.

- ▶ The test then amounts to a test of whether or not the model for the logarithms of the cell probabilities is a linear function of the logarithms of the marginal probabilities.

4.8 Summary

1. *Contingency table:* An array of natural numbers in matrix form where those natural numbers represent counts, or frequencies. 84

2. *Chi-squared test for differences in probabilities T_1*: Two independent populations (rows fixed, columns random)

$$T_1 = \frac{\sqrt{N}(O_{11}O_{22} - O_{12}O_{21})}{\sqrt{n_1 n_2 C_1 C_2}} \approx N(0, 1)$$

   85

3. *Three data types of 2×2 contingency table*:

   (a) Row totals are fixed and column totals are random variables.
   (b) Row totals and column totals are random variables.
   (c) Row totals and column totals are known prior to an examination of the data.

   89

4. Fisher's exact test *T_2*: The number of observations in the cell in row 1, column 1 (rows and columns fixed).

$$P(T_2 = x) = \frac{\binom{r}{x}\binom{N-r}{c-x}}{\binom{N}{c}}, \quad x = 0, 1, \ldots, \min(r, c)$$

   90

5. *T_3*: Large sample approximation for T_2

$$T_3 = \frac{T_2 - \frac{rc}{N}}{\sqrt{\frac{rc(N-r)(N-c)}{N^2(N-1)}}} \approx N(0, 1)$$

   90

6. *Fisher's exact test*: Exact only if the row and column totals are nonrandom......91

7. T_2: has the hypergeometric distribution..91

8. *Mantel-Haenszel test*: Method for combining several 2 × 2 contingency tables....92

9. T_4: Test for combining several 2 × 2 contingency tables (rows and columns fixed)

$$T_4 = \frac{\sum x_i - \sum \frac{r_i c_i}{N_i}}{\sqrt{\sum \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{N_i^2 (N_i - 1)}}} \approx N(0,1)$$

....92

10. T_5: Test for 2 × 2 contingency tables with row totals or column totals are random.

$$T_5 = \frac{\sum x_i - \sum \frac{r_i c_i}{N_i}}{\sqrt{\sum \frac{r_i c_i (N_i - r_i)(N_i - c_i)}{N_i^3}}} \approx N(0,1)$$

....93

11. T_5 test is valid even though the row totals or column totals are random..........93

12. The continuity correction should not be used to find p-values when T_5 is used....93

13. *$r \times c$ contingency table*: The one-sided hypotheses of the previous section are no longer appropriate...94

14. T: Test for differences in probabilities, $r \times c$.

$$T = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_i \sum_j \frac{O_{ij}^2}{E_{ij}} - N \approx \chi^2_{(r-1)(c-1)}$$

$E_{ij} = n_i C_j / N$ represents the expected number of observations in cell (i, j), if H_0 is true...95

15.
- ▶ The exact distribution of T in the $r \times c$ case may be found in exactly the same way as it was found in the previous section for the 2 × 2 case.
- ▶ The row totals are held constant, and then all possible contingency tables having those same row totals are listed, and their probabilities are calculated using the multinomial distribution for each row.
- ▶ In all three applications of contingency tables in this section, the asymptotic distribution of T is the same, $\chi^2_{(r-1)(c-1)}$. □

....97

16. Exact distribution of T with $N = 4$

- ▶ $\frac{N!}{a!b!c!d!}(p_{11})^a (p_{12})^b (p_{21})^c (p_{22})^d$
- ▶ The maximum size of the upper tail of T when H_0 is true, is found by setting all of the p_{ij}s equal to each other: $\frac{N!}{a!b!c!d!}\left(\frac{1}{4}\right)^N$
- ▶ There are $\binom{7}{3} = 35$ different contingency tables (Fig. 1).

- $P(T=0) = 84/256 = 0.33$
- $P(T=4/9) = 48/256 = 0.19$
- $P(T=4/3) = 96/256 = 0.37$
- $P(T=4) = 28/256 = 0.11$
- The distribution of T is more complicated to obtain than the row totals are fixed. □

. . . . 99

17.
- The exact distribution of O_{11} is the hypergeometric distribution for 2 × 2 case.
- Row totals and column totals are all equal 2.

Table	Probability	T
$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$	$\frac{\binom{2}{2}\binom{2}{0}}{\binom{4}{2}} = 1/6$	4
$\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$	$\frac{\binom{2}{1}\binom{2}{1}}{\binom{4}{2}} = 2/3$	0
$\begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$	$\frac{\binom{2}{0}\binom{2}{2}}{\binom{4}{2}} = 1/6$	4

. . . . 102

18.
- Fixed row totals and fixed column totals greatly reduce the contingency tables possible. When $r = c = 2$, the test is known as "Fisher's exact test".
- For r and c in general, the exact probability of the table with fixed marginal totals is given by

$$\text{probability} = \frac{\binom{n_1}{O_{1i}}\cdots\binom{n_r}{O_{ri}}}{\binom{N}{c_i}}$$

where the multinomial coefficients $\binom{n_m}{O_{mi}} = \frac{n_m!}{O_{m1}!O_{m2}!\cdot O_{mc}!}$.

. . . . 102

19.

$$P\left(\begin{pmatrix} O_{11} & O_{12} & \cdots & O_{1c} \\ O_{21} & O_{22} & \cdots & O_{2c} \end{pmatrix}\right) = \binom{n_1}{O_{11}}\binom{n_2}{O_{12}}\cdots\binom{n_c}{O_{1c}}p^a(1-p)^b \qquad (A)$$

$$P((a,b)|N) = \binom{N}{a}p^a(1-p)^b \qquad (B)$$

$(A) \div (B)$

$$P\left(\begin{pmatrix} O_{11} & O_{12} & \cdots & O_{1c} & a \\ O_{21} & O_{22} & \cdots & O_{2c} & b \\ n_1 & n_2 & \cdots & n_c & N \end{pmatrix}\right) = \frac{\binom{n_1}{O_{11}}\binom{n_2}{O_{12}}\cdots\binom{n_c}{O_{1c}}}{\binom{N}{a}}$$

. . . . 106

20. *Contingency table:* The contingency table is a convenient form for examining data to see if there is some sort of dependence inherent in the data. 107

21. $R_1 = \frac{T}{\max T} = \frac{T}{N(q-1)}$ 108

22. *Cramér's coefficient:* $\sqrt{R_1} = \sqrt{\frac{T}{N(q-1)}}$ 108

23. *Pearson's coefficient (contingency coefficient):* Pearson's coefficient of mean square contingency (Yule and Kendall, 1950):

$$R_2 = \sqrt{\frac{T}{N+T}}$$

. 109

24. *Mean-square contingency*:

$$R_3 = \frac{T}{N}$$

. 109

25. *Tschuprow's coefficient*:

$$R_4 = \sqrt{\frac{T}{N\sqrt{(r-1)(c-1)}}}$$

. 110

26. *2 × 2 contingency table*:

$$\begin{aligned}
R_1 = R_3 = \frac{T}{N} &= \frac{(ad-bc)^2}{r_1 r_2 c_1 c_2} \\
R_2 = \sqrt{\frac{T}{N+T}} &= \sqrt{\frac{(ad-bc)^2}{r_1 r_2 c_1 c_2 + (ad-bc)^2}} \\
R_5 &= \frac{ad-bc}{\sqrt{r_1 r_2 c_1 c_2}} \\
R_6 &= \frac{ad-bc}{ad+bc} \\
R_7 &= \frac{(a+d)-(b+c)}{a+b+c+d}
\end{aligned}$$

. 110

27. *Goodness-of-fit test:* A test designed to compare the sample obtained with the hypothesized distribution to see if the hypothesized distribution function fits the data in the sample. 113

28. *Chi-squared test for goodness of fit*: $T = \sum \frac{(O_j-E_j)^2}{E_j} = \sum \frac{O_j^2}{E_j} - N \approx \chi_{c-1}$ 113

29. *Goodness-of-fit test*: The probability of any particular arrangement:

$$P(O_1,\ldots,O_c|N) = \frac{N!}{O_1!\cdots O_c!}\, p_1^{O_1}\cdots p_c^{O_c}$$

. 117

30. *Cochran's test*:

- $X_{ij} \sim B(1, p)$ where p is the same within each row under H_0, but can be different from block to block.
- From CLT: $\frac{C_j - E(C_j)}{\sqrt{\mathrm{Var}(C_j)}} \approx N(0,1)$ where $C_j = \sum_{i=1}^{r} X_{ij}$.
- $\sum_{j=1}^{c} \left[\frac{C_j - E(C_j)}{\sqrt{\mathrm{Var}(C_j)}} \right]^2 \approx \chi^2_c$
- $\hat{E}(C_j) = \frac{1}{c} \sum C_j = \frac{N}{c}$
- $\mathrm{Var}(C_j) = \sum \mathrm{Var}(X_{ij}) = \sum p(1-p)$
- $\hat{p}$ in row $i = R_i/c$
- Estimate of $\mathrm{Var}(X_{ij}) = \frac{R_i}{c}\left(1 - \frac{R_i}{c}\right)$
- Improved estimate of $\mathrm{Var}(X_{ij}) = \frac{R_i(c-R_i)}{c(c-1)}$
- Improved estimate of

$$\mathrm{Var}(C_j) = \frac{1}{c(c-1)} \sum_{i=1}^{r} R_i(c - R_i)$$

- Substitution of the estimates for $E(C_j)$ and $\mathrm{Var}(C_j)$ into Eq. (5) gives

$$T = c(c-1) \frac{\sum_{j=1}^{c} (C_j - \frac{N}{c})^2}{\sum_{i=1}^{r} R_i(c - R_i)}$$

□

....119

31. If $c = 2$, the Cochran test is identical with the McNemar test (Section 3.5)......119

32. *Likelihood ratio chi-squared statistic*: $T_2 = 2 \sum O_i \ln \left(\frac{O_i}{E_i} \right)$ 120

Chapter 5

Some Methods Based on Ranks

Preliminary remarks

- Most of the statistical procedures introduced in the previous chapters can be used on the data that have a nominal scale of measurement.
- Statistical procedures for dichotomous data in Chap. 3.
- Statistical procedures for data classified according to two or more different criteria and into two or more separate classes by each criterion in Chap. 4.
- Nominal scale of measurement in Chap. 3-4.
- All of those procedures may also be used where more than nominal information concerning the data is available.
- Some of the information contained in the data is disregarded and the data are reduced to nominal-type data for analysis.
- Such a loss of information usually results in a corresponding loss of power.
- In this chapter several statistical methods are presented that utilize more of the information contained in the data, if the data have at least an ordinal scale of measurement.
- If data are nonnumeric but are ranked as in ordinal-type data, the methods of this chapter are often the most powerful ones available.
- If data are numeric and, furthermore, are observations on random variables that have the normal distribution so that all of the assumptions of the usual parametric tests are met, the loss of efficiency caused by using the methods of this chapter is surprisingly small.
- The rank tests of this chapter are valid for all types of populations, whether continuous, discrete, or mixtures of the two.
- Data with many ties may be analyzed using rank tests if the data are ordinal.
- If there are extensive ties in the data, the large sample approximation and not the small sample tables in this book should be used.

Stem-and-leaf method (Tukey, 1977)

- A convenient method of arranging observations in increasing order.
- Suppose a class of 28 students obtained the following scores on an exam.

74	63	88	69	81	91	75
82	91	87	77	86	86	87
96	84	93	73	74	93	78
70	84	90	97	79	89	93

Stem-and-leaf:

9	0	1	1	3	3	3	6	7		
8	1	2	4	4	6	6	7	7	8	9
7	0	3	4	4	5	7	8	9		
6	3	9								

- The scores may be arranged from smallest to largest quite easily now.
- In this way the ranks may be assigned to the observations.

5.1 Two independent samples

- The test presented in this section is known as the *Mann-Whitney test* and also as the *Wilcoxon test*.
- The usual two-sample situation is one in which the experimenter has obtained two samples from possibly different populations and wishes to use a statistical test to see if the null hypothesis that the two populations are identical can be rejected.
- An equivalent situation is where one random sample is obtained, but it is randomly subdivided into two samples.
- If the samples consist of ordinal-type data, the most interesting difference is a difference in the locations of the two populations.
  - Does one population tend to yield larger values than the other population?
  - Are the two medians equal?
  - Are the two means equal?
- An intuitive approach to the two-sample problem is to combine both samples into a single ordered sample and assign ranks to the sample values from the smallest value to the largest, without regard to which population each value came from.
- Then the test statistic might be the sum of the ranks assigned to those values from one of the populations.
- If the sum is too small (or too large), there is some indication that the values from that population tend to be smaller (or larger, as the case may be) than the values from the other population.
- The null hypothesis of no differences between populations may be rejected if the ranks associated with one sample tend to be larger than those of the other sample.

▶ Ranks are considered preferable to the actual data:

1. If the numbers assigned to the observations have no meaning by themselves but attaining only in an ordinal comparison with the other observations.
2. If the numbers have meaning but the distribution function is not a normal distribution function, the probability theory is usually beyond our reach. The probability theory of statistics based on ranks is relative simple.
3. The A.R.E. of the Mann-Whitney test is never too bad when compared with the two-sample t test, the usual parametric counterpart.

The Mann-Whitney (Wilcoxon sign ranks) Test

Data The data consist of two random samples. Let $X_1, X_2, \ldots, X_n$ denote the random sample of size n from population 1 and let $Y_1, Y_2, \ldots, Y_m$ denote the random sample of size m from population 2. Assign the rank 1 to $n+m$ to the observations from smallest to largest. Let $R(X_i)$ and $R(Y_j)$ denote the rank assigned to X_i and Y_j for all i and j. Let $N = n+m$.

If several sample values are exactly equal to each other, assign to each the average of the ranks that would have been assigned to them had there been no ties.

Assumptions

1. Both samples are random samples from their respective populations,
2. In addition to independence within each sample, there is mutual independence between the two samples.
3. The measurement scale is at least ordinal.

Test statistic

▷ No ties or a few ties:

$$T = \sum R(X_i)$$

▷ Many ties:

$$T_1 = \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)}\sum R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}}$$

where $\sum R_i^2$ refers to the sum of the squares of all N of the ranks or average ranks actually used in both samples.

Null distribution

▷ Lower quantiles of the null distribution of T given in Table A7 for $n, m \leq 20$.

▷ Upper quantile w_p:

$$w_p = n(n+m+1) - w_{1-p}$$

▷ The quantiles in Table A7 are exact only if there are no ties in the data and therefore no average ranks are used.

▷ Approximate quantiles in the case of no ties, and n or m greater than 20:

$$w_p \approx \frac{n(N+1)}{2} + z_p\sqrt{\frac{nm(N+1)}{12}}$$

- ▷ If there are many ties in the data, then T_1 is used instead of T, and T_1 is approximately a standard normal random variable whose quantiles are given in Table A1.

Hypotheses Let $F(x)$ and $G(x)$ be the distribution functions corresponding to X and Y, respectively.

A. *Two-tailed test*:

$$\begin{aligned} &H_0 : F(x) = G(x) && \text{for all } x \\ &H_1 : F(x) \neq G(x) && \text{for some } x \end{aligned}$$

$$p\text{-value} = 2P\left(Z \leq \frac{T + \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right)$$

B. *Lower-tailed test*:

$$\begin{aligned} &H_0 : F(x) = G(x) && \text{for all } x \\ &H_1 : F(x) > G(x) && \text{for some } x \end{aligned}$$

$$p\text{-value} = P\left(Z \leq \frac{T + \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right)$$

C. *Upper-tailed test*:

$$\begin{aligned} &H_0 : F(x) = G(x) && \text{for all } x \\ &H_1 : F(x) < G(x) && \text{for some } x \end{aligned}$$

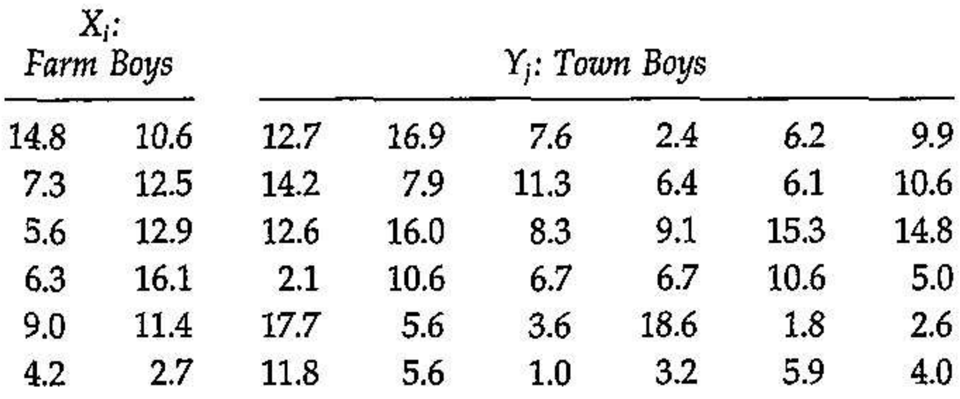
$$p\text{-value} = P\left(Z \geq \frac{T - \frac{1}{2} - n\frac{N+1}{2}}{\sqrt{\frac{nm(N+1)}{12}}}\right)$$

- ▶ The *Mann-Whitney test* is unbiased and consistent when testing the preceding hypotheses involving $P(X > Y)$.
- ▶ The same is not always true for the hypotheses involving $E(X)$ and $E(Y)$.
- ▶ To insure that the test remains consistent and unbiased for hypotheses involving $E(X)$:

  Assumption 4: If there is a difference between population distribution functions, that differences in the location of the distributions ($F(x) \equiv G(x + c)$).

Example 5.1.1 *The senior class in a particular high school had 48 boys. Twelve boys lived on farms and the other lived in town. A test was devised to see if farm boys in general were more physically fit than town boys. Each boy in the class was given a physical fitness test in which a low score indicates poor physical condition. The scores of the farm boys* (X_i) *and the town boys* Y_j.

- ▶ Neither group of boys is a random sample from any population.
- ▶ However, it is reasonable to assume that these scores resemble hypothetical random samples from the populations of farm and town boys in that age group, at least for similar localities.



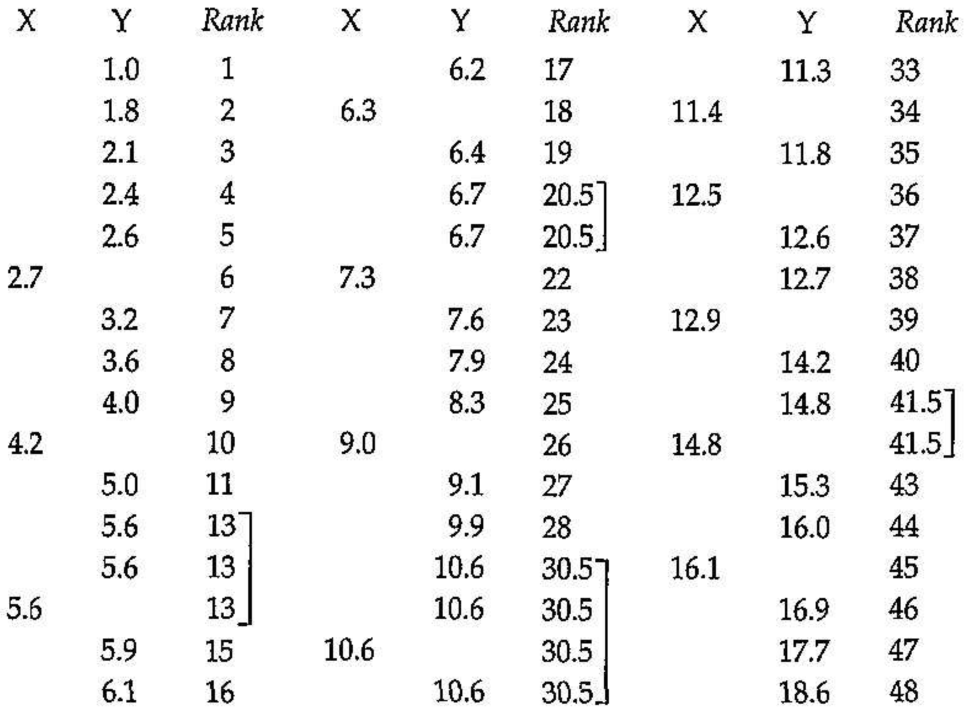
X_i: Farm Boys		Y_j: Town Boys					
14.8	10.6	12.7	16.9	7.6	2.4	6.2	9.9
7.3	12.5	14.2	7.9	11.3	6.4	6.1	10.6
5.6	12.9	12.6	16.0	8.3	9.1	15.3	14.8
6.3	16.1	2.1	10.6	6.7	6.7	10.6	5.0
9.0	11.4	17.7	5.6	3.6	18.6	1.8	2.6
4.2	2.7	11.8	5.6	1.0	3.2	5.9	4.0

Figure 5.1: *Scores for physically fit*

- The other assumptions of the model seem to be reasonable, such as independence between groups.
- The *Mann-Whitney test* is selected to test:

H_0 : Farm boys do not tend to be more fit, physically, than town boys
H_1 : Farms boys tend to be more fit than town boys

- The test is one tailed.
- The scores are ranked as follows.



X	Y	*Rank*	X	Y	*Rank*	X	Y	*Rank*
	1.0	1		6.2	17		11.3	33
	1.8	2	6.3		18	11.4		34
	2.1	3		6.4	19		11.8	35
	2.4	4		6.7	20.5	12.5		36
	2.6	5		6.7	20.5		12.6	37
2.7		6	7.3		22		12.7	38
	3.2	7		7.6	23	12.9		39
	3.6	8		7.9	24		14.2	40
	4.0	9		8.3	25		14.8	41.5
4.2		10	9.0		26	14.8		41.5
	5.0	11		9.1	27		15.3	43
	5.6	13		9.9	28		16.0	44
	5.6	13		10.6	30.5	16.1		45
5.6		13		10.6	30.5		16.9	46
	5.9	15	10.6		30.5		17.7	47
	6.1	16		10.6	30.5		18.6	48

Figure 5.2: *Ranks of scores for physically fit*

- This is not a large number of ties, so it is probably acceptable to use T instead of T_1.
- Both methods will be compared in the example.

- $n = 12, m = 36$ and $N = m + n = 48$.

$$\begin{aligned} T &= \sum_{i=1}^{n} R(X_i) \\ &= 6 + 10 + 13 + 18 + 22 + 26 + 30.5 + 34 + 36 + 39 + 41.5 + 45 = 321 \end{aligned}$$

$$\sum_{i=1}^{N} R_i^2 = 38,016$$

$$\begin{aligned} T_1 &= \frac{T - n\frac{N+1}{2}}{\sqrt{\frac{nm}{N(N-1)} \sum_{i=1}^{N} R_i^2 - \frac{nm(N+1)^2}{4(N-1)}}} \\ &= \frac{321 - 12\frac{49}{2}}{\sqrt{\frac{(12)(36)}{(48)(47)}(38016) - \frac{(12)(36)(49)^2}{4(47)}}} \\ &= 0.6431 \end{aligned}$$

$$\begin{aligned} w_{0.95} &= n\frac{N+1}{2} + (1.6449)\sqrt{nm(N+1)/12} \\ &= 294 + (1.6449)(42) \\ &= 363.1 \end{aligned}$$

- Accept H_0. □

- The next example illustrates a situation in which no random variables are defined explicitly.

Example 5.1.2 *A simple experiment was designed to see if flint (打火石) in area A tended to have the same degree of hardness as flint in area B. Four sample pieces of flint were collected in area A and five sample pieces of flint were collected in area B. To determine which of two pieces of flint was harder, the two pieces were rubbed against. The piece sustaining less damage was judged the harder of the two. In this manner all nine pieces of flint were ordered according to hardness. The rank 1 was assigned to the softest piece, rank 2 to the next softest, and so on.*

Rank	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Origin of piece	*A*	*A*	*A*	*B*	*A*	*B*	*B*	*B*	*B*

- H_0 : The flints from areas A and B are of equal hardness
  H_1 : The flints are not of equal hardness

- *Mann-Whitney two-tailed test* is used:

$$\begin{aligned} T &= \text{sum of the ranks of pieces from area A} \\ &= 1 + 2 + 3 + 5 \\ &= 11 \end{aligned}$$

- Reject H_0 at size .05 since the two-tailed critical region is either $T < 12$ or $T > 40 - 12 = 28$.

$$\begin{aligned} p\text{-value} &\cong 2 \cdot P\left(Z \le \frac{11 + \frac{1}{2} - 4\frac{10}{2}}{\sqrt{\frac{4 \cdot 5 \cdot 10}{12}}}\right) \\ &= 2 \cdot P(z \le -2.0821) \\ &= 0.038 \end{aligned}$$

Theory

- $T = \sum R(X_i)$
- Assume that X_i and Y_j are identically and independently distributed.
- Every arrangement of the Xs and Ys in the ordered combined sample is equally likely.
- The number of ways of selecting n integers from a total number of $n+m$ integers is $\binom{n+m}{n}$, and each way has equal probability according to the basic premise just stated.
- The probability $T = k$ may be found by counting the number of different sets of n integers from 1 to $n+m$ that add up to the value k and then dividing that number by $\binom{n+m}{n}$.
- Example: $n = 4$ and $m = 5$
- Number of ways of selecting four out of nine ranks:

$$\binom{n+m}{n} = \frac{9!}{4!5!} = 126$$

- $\min T = 10$, $P(T = 10) = P(T = 11) = 1/126$ and $P(T = 12) = 2/126$.
- $P(T \le 11) = 2/126 = 1/63 = 0.0159$
- For large n and m, use CLT to approximate the distribution of T.
- $E(T) = \frac{n(n+m+1)}{2}$ (Theorem 1.4.5)
- $\text{Var}(T) = n\frac{(N+1)(N-1)}{12} + n(n-1)\left(-\frac{N+1}{12}\right) = \frac{n(n+m+1)m}{12}$ (Theorem 1.4.5, $N = n+m$)
- Quantile: $w_p = E(T) + z_p\sqrt{\text{Var}(T)}$

  Mann-Whitney test may be used for testing:

$$H_0: \; E(X) = E(Y) + d$$

  By collecting all the values of d that would result in acceptance of the preceding H_0, we have a confidence interval for $E(X) - E(Y)$.

Confidence interval for the difference between two means

Data The data consist of two random samples $X_1, X_2, \ldots, X_n$ and $Y_1, \ldots, Y_m$ of size n and m, respectively. Let X and Y denote random variables with the same distributions as the X_i and Y_j, respectively.

Assumptions

1. Both samples are random samples from their respective populations.
2. In addition to independence within each sample, there is mutual independence between the two samples.
3. The two population distribution functions are identical except for a possible difference in location parameters. That is, there is a constant d such that X has the same distribution function as $Y + d$.

Method

- Determine $w_{\alpha/2}$ from Table A7 or Eq. 5 if n and m are large, where $(1-\alpha)$ is the desired confidence coefficient.
- $k = w_{\alpha/2} - n(n+1)/2$ (lower quartile)
- From all of possible pairs (X_i, Y_j), find the k largest differences $X_i - Y_j$ (U) and find the k smallest differences (L).

$$P[L \leq E(X) - E(Y) \leq U] \geq 1 - \alpha$$

Example 5.1.3 *A cake batter is to be mixed until it reaches a specified level of consistency. Five batches of the batter are mixed using mixer A, and another five batches are mixed using mixer B. The required times for mixing are given as follows*

Mixer A	*Mixer B*
7.3	*7.4*
6.9	*6.8*
7.2	*6.9*
7.8	*6.7*
7.2	*7.1*

- A 95% confidence interval is sought for the mean difference in mixing times, $E(X) - E(Y)$, where X refers to mixer A and Y refers to mixer B.
- $n = m = 5, \alpha = .05$, Table A7 yields $w_{.025} = 18$ and $k = 18 - 5 \cdot 6/2 = 3$.
- Two samples are ordered from smallest to largest and Xs are used as rows and Ys are used as columns:

$X_i \backslash Y_j$	6.7	6.8	6.9	7.1	7.4
6.9	0.2	0.1	0.0	-0.2	-0.5
7.2	0.5	0.4	0.3	0.1	-0.2
7.2	0.5	0.4	0.3	0.1	-0.2
7.3	0.6	0.5	0.4	0.2	-0.1
7.8	1.1	1.0	0.9	0.7	0.4

Smallest Differences	Largest Differences
$6.9-7.4=-0.5$	$7.8-6.7=1.1$
$6.9-7.1=-0.2$	$7.8-6.8=1.0$
$7.2-7.4=-0.2=L$	$7.8-6.9=0.9=U$

- 95% CI $(L,U)=(-0.2,0.9)$ □

Theory

- # of $(X_i,Y_j)=mn$
- If $k=\#$ of $X_i-Y_j>0$, then $T=k+n(n+1)/2$.
- $T=n(n+1)/2$ if no Ys are smaller than any of Xs.
- The borderline value of T, where H_0 is barely accepted, is given in Table A7 as $w_{\alpha/2}$.
- By subtracting $n(n+1)/2$ from $w_{\alpha/2}$, we find the borderline value of k.
- Want to find the value of d that we can add to the Ys to achieve barely this borderline value of k.
- If we add the maximum of all the differences X_i-Y_j to each of the Ys, then none of the Xs will be greater than the adjusted Ys.
- Add the kth largest difference X_i-Y_j to each of Ys, we achieve the borderline case: fewer than k pairs satisfy $X_i>Y_j+d$, and at least k pairs satisfy $X_i>Y_j+d$. In this way we obtain the largest value of d that results in acceptance of $H_0:\ E(X)=E(Y)+d$.
- By reversing the procedure and working from the lower end, we obtain the smallest value of d that results in acceptance of the same hypothesis.

Comparison with other procedures
Mann-Whitney test v.s. two-sample t test

- $t=\frac{(\overline{X}-\overline{Y})\sqrt{mn(N-2)/N}}{\sqrt{\sum(X_i-\overline{X})^2+\sum(Y_j-\overline{Y})^2}}$
- If the population has normal distribution, then the t test is the most powerful test.
- If the population is not normal distribution, then the t test when compare with the Mann-Whitney test results very little power. This is especially true when one or both samples contain unusually large or small observations, called “outliers.”
- The A.R.E. of the Mann-Whitney test as compared with the t test is 0.955 for normal population, 1.0 for uniform population, 1.5 for double exponential distribution.
- If the two populations differ only in their location parameters the A.R.E. is never lower than 0.864 but may be as high as infinity.
- The median test also may be used for data of this type. The A.R.E. of the Mann-Whitney test relative to the median test is 1.5 for normal populations, 3.0 for uniform distributions, but only 0.75 in the double exponential case.

- The Mann-Whitney test was first introduced for the case $n = m$ by Wilcoxon (1945). Wilcoxon's test was extended to the case of unequal samples sizes by White (1952).
- Mann and Whitney (1947) seem to be the first to consider unequal sample sizes and to furnish tables suitable for use with small sample.
- It is largely the work of Mann and Whitney that led to widespread use of the test.

5.2 Several independent samples

- The Mann-Whitney test for two independent samples was extended to the problem of analyzing k independent samples by Kruskal and Wallis (1952).
- The experimental situation is one where k random samples have been obtained, one from each of k possibly different populations, and we want to test the null hypothesis that all of the populations are identical against the alternative that some of the populations tend to furnish greater observed values than other populations.

The Kruskal-Wallis test

Data The data consist of k random samples of possibly different sizes. Denote the ith random sample of size n_i by $X_{i1}, X_{i2}, \dots, X_{in_i}$. Then the data may be arranged into columns.

Sample 1	Sample 2	$\cdots$	Sample k
$X_{1,1}$	$X_{2,1}$	$\cdots$	$X_{k,1}$
$X_{1,2}$	$X_{2,2}$	$\cdots$	$X_{k,2}$
$\cdots$	$\cdots$		$\cdots$
X_{1,n_1}	X_{2,n_2}	$\cdots$	X_{k,n_k}

Let N denote the total number of observations $N = \sum n_i$. Assign rank 1 to the smallest of the totality of N observations, rank 2 to the second smallest, and so on to the largest of all N observations, which receives rank N. Let $R(X_{ij})$ represent the rank assigned to X_{ij}. Let R_i be the sum of the ranks assigned to the ith sample. $R_i = \sum R(X_{ij})$. Compute R_i for each sample.

If the ranks may be assigned in several different ways because several observations equal to each other, assign the average rank to each of the tied observations, as in the previous test of this chapter.

Assumptions

1. All samples are random samples from their respective populations.
2. In addition to independence within each sample, there is mutual independence among the various samples.
3. The measurement scale is at least ordinal.
4. Either the k population distribution functions are identical, or else some of the populations tend to yield larger values than other populations do.

Test statistic

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right)$$

where $S^2 = \frac{1}{N-1}\left(\sum R(X_{ij})^2 - N\frac{(N+1)^2}{4}\right)$.

- ▷ If there are no ties $S^2 = N(N+1)/12$ and $T = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$.

Null distribution

- ▷ Exact distribution of T is given by Table A8 for $k = 3$ and all $n_i \leq 5$.
- ▷ Chi-squared approximation used: χ^2_{k-1}

Hypothesis

H_0: All of the k population distribution functions are identical

H_1: At least one of the populations tends to yield larger observations than at least one of the other populations.

- ▷ Reject H_0 at the level α if T is greater than its $1-\alpha$ quantile from the null distribution.
- ▷ The p-value is approximately the probability of a chi-squared random variable with $k-1$ degrees of freedom exceeding the observed value of T.

Multiple comparisons

After rejecting H_0, we can test that populations i and j seem to be different if the following inequality is satisfied:

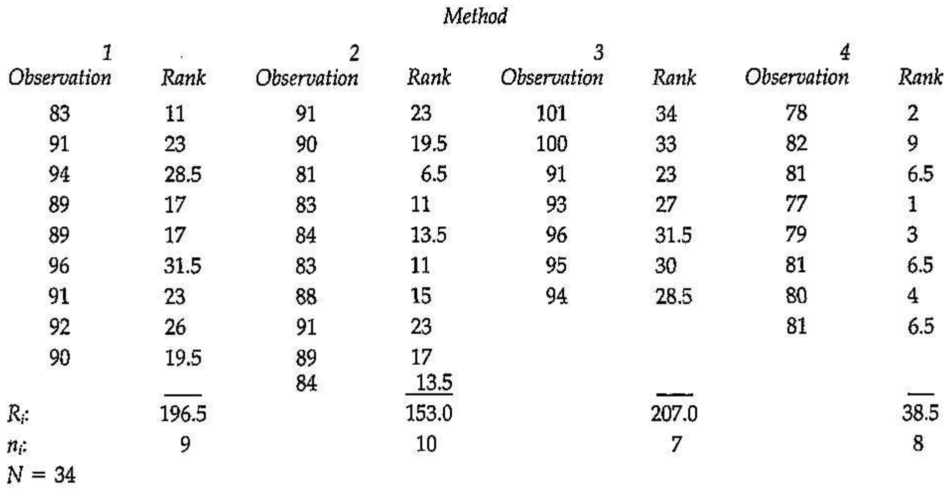
$$\left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right| > t_{N-k,1-(\alpha/2)} \left(S^2 \frac{N-1-T}{N-k}\right)^{1/2} \left(\frac{1}{n_i} + \frac{1}{n_j}\right)^{1/2}$$

Example 5.2.1 *Data from a completely randomized design were given in Example 4.3.1, where four different methods of growing corn resulted in various yields per acre on various plots of ground where the four methods were tried. Ordinarily, only one statistical analysis is used, but here we will use the Kruskal-Wallis test so that a rough comparison may be made with median test, which previously furnished a p-value of slightly less than* 0.001.

H_0 : *The four methods are equivalent*

H_1 : *Some methods of growing corn tend to furnish higher yields than others*

- ▶ The smallest, 77: rank: 1
- ▶ The largest, 101: rank: $N = 34$
- ▶ Tied values receive the average ranks. The ranks of the observations, with the sums R_i, are given next.
- ▶ $n_i = 9, 10, 7, 8$
- ▶ $R_i = 196.5, 153.5, 207.0, 38.5$
- ▶ $T = 25.46$
- ▶ Reject H_0 at size .05 since $T > \chi^2_{3,.95} = 7.815$
- ▶ A rough idea of the power of the Kruskal-Wallis test as compared with the median test may be obtained by comparing the value of the test statistics inboth tests.



	Method							
	1		2		3		4	
	Observation	Rank	Observation	Rank	Observation	Rank	Observation	Rank
	83	11	91	23	101	34	78	2
	91	23	90	19.5	100	33	82	9
	94	28.5	81	6.5	91	23	81	6.5
	89	17	83	11	93	27	77	1
	89	17	84	13.5	96	31.5	79	3
	96	31.5	83	11	95	30	81	6.5
	91	23	88	15	94	28.5	80	4
	92	26	91	23			81	6.5
	90	19.5	89	17				
			84	13.5				
R_i:		196.5		153.0		207.0		38.5
n_i:		9		10		7		8
$N = 34$								

Figure 5.3: *Scores for growing corn data*

- ▷ Both test statistics have identical asymptotic distributions, the chi-squared distribution with 3 degrees of freedom.
- ▷ However, the value 25.46 attained in the Kruskal-Wallis test is somewhat larger than the value 17.6 computed in the median test, indicating more sensitivity to the sample differences.

▶ Multiple comparison: ignore the few ties and use

$$S^2 = N(N+1)/12 = 99.167$$

$$\frac{S^2(N-1-T)}{N-k} = \frac{(99.167)(33-25.464)}{34-4} = 24.911$$

Populations	$\left\lvert \frac{R_i}{n_i} - \frac{R_j}{n_j} \right\rvert$	$2.041(24.911)^{1/2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{1/2}$
1 and 2	6.533	4.681
1 and 3	7.738	5.134
1 and 4	17.021	4.950
2 and 3	14.271	5.020
2 and 4	10.488	4.832
3 and 4	24.759	5.272

▶ In every case the second column exceeds the third column, the multiple comparisons procedure shows every pair of populations to be different. □

▶ The Kruskal-Wallis test is an excellent test to use in a contingency table, where the rows represent ordered categories and the columns represent the different populations:

				Population			Row	
		1	2	3	$\cdots$	k	Totals	$\overline{R}_i$ =Average Rank
Category	1	O_{11}	O_{12}	O_{13}	$\cdots$	O_{1k}	t_1	$(t_1+1)/2$
	2	O_{21}	O_{22}	O_{23}	$\cdots$	O_{2k}	t_2	$t_1+(t_2+1)/2$
	3	O_{31}	O_{32}	O_{33}	$\cdots$	O_{3k}	t_3	$t_1+t_2+(t_3+1)/2$
	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$		
	c	O_{c1}	O_{c2}	O_{c3}	$\cdots$	O_{ck}	t_c	$\sum_1^{c-1} t_i + (t_c+1)/2$
		n_1	n_2	n_3	$\cdots$	n_k	N	

- O_{ij}: Number of the observations in population j fall into the ith category.
- $\overline{R}_i$: Average rank of row i.
- All of the observations in row i are considered equal to each other but less than the observations in row $i+1$.
- $R_j = \sum_{i=1}^{c} O_{ij}\overline{R}_i$
- $S^2 = \frac{1}{N-1}\left[\sum_{i=1}^{c} t_i\overline{R}_i^2 - N(N+1)^2/4\right]$

$$T = \frac{1}{S^2}\left(\sum_{i=1}^{k}\frac{R_i^2}{n_i} - \frac{N(N+1)^2}{4}\right)$$

- If the null hypothesis is rejected the multiple comparisons procedure may be used, to pinpoint differences where they exist.

Example 5.2.2 *Three instructors compared the grades they assigned over the past semester to see if some of them tended to give lower grades than others.*

H_0 : *The three instructors grade evenly with each other.*
H_1 : *Some instructions tend to grade lower than others.*

- Grades examined:

	Instructors			Row	Average
Grades	1	2	3	Totals	Ranks
A	4	10	6	20	10.5
B	14	6	7	27	34
C	17	9	8	34	64.5
D	6	7	6	19	91
F	2	6	1	9	105
Total	43	38	28	109	

- Column rank sums by Eq. 9:

$$R_1 = 2370.5, \quad R_2 = 2156.5, \quad R_3 = 1468$$

- $\sum R_j = N(N+1)/2 = 5995.$
- $S^2 = 941.71$ from Eq. 10.
- $T = 0.3209$ from Eq. 3.

- Accept H_0 at size .05 since $T < \chi^2_{2,.95} = 5.991$. □

Theory

- Each arrangement of the ranks 1 to N into groups of sizes $n_1, n_2, \ldots, n_k$, which is equally likely, and occurs with probability $n_1!n_2!\cdots n_k!/N!$.
- The value of T is computed for each arrangement.
- Example: $n_1 = 2, n_2 = 1$, and $n_3 = 1$

	Sample			
Arrangement	1	2	3	T
1	1, 2	3	4	2.7
2	1, 2	4	3	2.7
3	1, 3	2	4	1.8
4	1, 3	4	2	1.8
5	1, 4	2	3	0.3
6	1, 4	3	2	0.3
7	2, 3	1	4	2.7
8	2, 3	4	1	2.7
9	2, 4	1	3	1.8
10	2, 4	3	1	1.8
11	3, 4	1	2	2.7
12	3, 4	2	1	2.7

- Distribution function:

x	$f(x) = P(T = x)$	$F(x) = P(T \leq x)$
0.3	$2/12 = 1/6$	1/6
1.8	$4/12 = 1/3$	1/2
2.7	$6/12 = 1/2$	1.0

- Large sample approximation for T
- $\frac{R_i - E(R_i)}{\sqrt{\text{Var}(R_i)}} \approx N(0,1)$ where $E(R_i) = \frac{n_i(N+1)}{2}$ and $\text{Var}(R_i) = \frac{n_i(N+1)(N-n_i)}{12}$ (Theorem 1.4.5).
- $\left[\frac{R_i - E(R_i)}{\sqrt{\text{Var}(R_i)}}\right]^2 = \frac{(R_i - [n_i(N+1)/2])^2}{n_i(N+1)(N-n_i)/12} \approx \chi^2_1$
- $T' = \sum_{i=1}^k \frac{(R_i - [n_i(N+1)/2])^2}{n_i(N+1)(N-n_i)/12} \approx \chi^2_k$
- R_i are dependent since $\sum R_i = N(N+1)/2$.
- Kruskal (1952) showed that if the ith term in T' is multiplied by $(N - n_i)/N$, then

$$T = \sum_{i=1}^{k} \frac{(R_i - [n_i(N+1)/2])^2}{n_i(N+1)N/12} \approx \chi^2_{k-1}$$

which is a rearrangement of the terms in Eq. 5.

- For two samples the Kruskal-Wallis test is equivalent to the Mann-Whitney test.

5.3 A test for equal variances

- Usual standard of comparison for several populations is based on the means or other measures of location of the populations.
- In some situations the variances of the populations may be quantity of interest.
- It has been claimed that the effect of seeding clouds with silver iodide is to increase the variance of the resulting rainfall.
- The test for variances is analogous to the test just presented for means.
- That is to test $H_0: E(X) = E(Y)$, the two independent samples were combined, ranked, and the sum of the ranks of the Xs was used as a test statistic.
- The variance is defined as the expected value of $(X - \mu)^2$ where μ is the mean of X.
- Thus to test $H_0: E[(X_i - \mu_x)^2] = E[(Y_j - \mu_y)^2]$ it seems reasonable to record the values of $(X_i - \mu_x)^2$ and $(Y_j - \mu_y)^2$, assign ranks to them, and use the sum of the ranks of the $(X_i - \mu_x)^2$s as the test statistic.
- Although this technique could be used, more power is obtained when the ranks are squared first and then summed.

The squared ranks test for variances

Data

- Population 1: $X_1, X_2, \ldots, X_n$
- Population 2: $Y_1, Y_2, \ldots, Y_m$
- $U_i = |X_i - \mu_1|$ and $V_j = |Y_j - \mu_2|$
- Assign the ranks 1 to $n+m$ to the combined sample of Us and Vs.
- If μ_1 and μ_2 are unknown, use $\overline{X}$ for μ_1 and $\overline{Y}$ for μ_2.
- $R(U_i)$ and $R(V_j)$ are the assigned ranks.

Assumptions

1. Both samples are random samples from their respective populations.
2. In addition to independence within each sample there is mutual independence between the two samples.
3. The measurement scale is at least interval.

Test statistic

- If there are no values of U tied with values of V,

$$T = \sum_{i=1}^{n} R(U_i)^2$$

- If there are ties,

$$T_1 = \frac{T - n\overline{R^2}}{\left[\frac{nm}{N(N-1)}\sum_1^N R_i^4 - \frac{nm}{N-1}\overline{R^2}^2\right]^{1/2}}$$

where $\overline{R^2} = \frac{1}{N}\left\{\sum R(U_i)^2 + \sum R(V_j)^2\right\}$ and $\sum_{i=1}^N R_i^4 = \sum_{j=1}^n [R(U_j)]^4 + \sum_{j=1}^m [R(V_j)]^4$.

Null distribution

- Quantiles of the exact null distribution of T are given in Table A9 for the case of no ties and $n, m \leq 10$.
- For n or $m > 10$, use the standard normal approximation:

$$w_p = \frac{n(N+1)(2N+1)}{6} + z_p \sqrt{\frac{mn(N+1)(2N+1)(8N+11)}{180}}$$

Hypotheses

A. *Two-tailed test*

H_0: X and Y are identically distributed, except for possibly different means

H_1: $\mathrm{Var}(X) \neq \mathrm{Var}(Y)$

- Reject H_0 at the level α if $T(T_1)$ is greater that its $1-\alpha/2$ or less than its $\alpha/2$ quantile, found from Table A9 or Table A1.
- p-value $= 2\cdot$ (smaller of the one-tailed p-values)
- Lower-tailed p-value $= P\left(Z \leq \frac{T-n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}}\right)$
- Upper-tailed p-value $= P\left(Z \geq \frac{T-n(N+1)(2N+1)/6}{\sqrt{mn(N+1)(2N+1)(8N+11)/180}}\right)$

B. *Lower-tailed test*

H_0: X and Y are identically distributed, except for possibly different means

H_1: $\mathrm{Var}(X) < \mathrm{Var}(Y)$

- Reject H_0 at the level α if $T(T_1)$ is less than its $\alpha/2$ quantile, found from Table A9 or Table A1.

C. *Upper-tailed test*

H_0: X and Y are identically distributed, except for possibly different means

H_1: $\mathrm{Var}(X) > \mathrm{Var}(Y)$

- Reject H_0 at the level α if $T(T_1)$ is greater that its $1-\alpha/2$ quantile, found from Table A9 or Table A1.

A test for more than two samples

- If there are three or more samples, this test is modified easily to test the equality of several variances.
- From each observation subtract its population mean (sample mean) and convert the sign of the resulting difference to $+$.
- Rank the combined absolute differences from the smallest to largest.
- Compute the sum of the squares of the ranks of each sample, letting $S_1, S_2, \ldots, S_k$ denote the sums for each of the k samples.

H_0: All k populations are identical, except for possibly different means
H_1: Some of the population variances are not equal to each other.

- Test statistic:

$$T_2 = \frac{1}{D^2}\left[\sum \frac{S_j^2}{n_j} - N(\overline{S})^2\right]$$

where $D^2 = \frac{1}{N-1}\left[\sum R_i^4 - N(\overline{S})^2\right]$.

- If there are no ties $D^2 = N(N+1)(2N+1)(8N+11)/180$ and $\overline{S} = (N+1)(2N+1)/6$.

  - $\sum_{k=1}^{N} k^4 = \frac{N\,(1+N)\,(1+2\,N)\,\left(-1+3\,N+3\,N^2\right)}{30}$

- $T_2 \approx \chi^2_{k-1}$

- p-value $\approx 1 - \chi^2_{k-1}(T_2)$

- If H_0 is rejected, multiple comparisons may be made.

- The variances of populations i and j are said to differ if

$$\left|\frac{S_i}{n_i} - \frac{S_j}{n_j}\right| > t_{N-k,1-\alpha/2}\left(D^2\frac{N-1-T_2}{N-k}\right)^{1/2}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)^{1/2}$$

Example 5.3.1 *A food packaging company would like to be reasonably sure that the boxes of cereal it produces do in fact contain at least the number of ounces of cereal stamped on the outside of the box. In order to do this it must set the average amount per box a little above the advertised amount, because the unavoidable variation caused by the packing machine will sometimes put a little less or a little more cereal in the box. A machine with smaller variation would save the company money because the average amount per box could be adjusted to be closer to the advertised amount.*

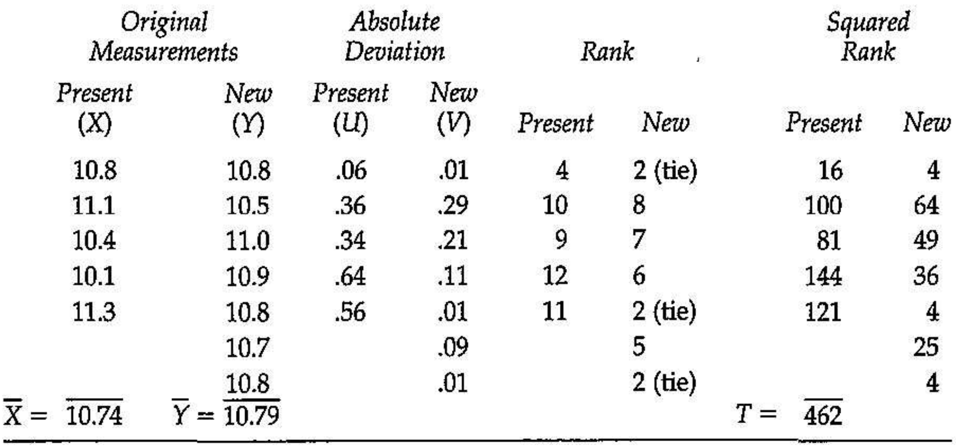
- A new machine is being tested to see if it is less variable than the present machine, in which case it will be purchased to replace the old machine.

- Several boxes are filled with cereal using the present machine and the amount in each box is measured.

- The same is done for the new machine to test:

H_0 : Both machines have equal variability
H_1 : The new machine has a smaller variance.

- The measurements and calculations are as follows.



Original Measurements		*Absolute Deviation*		*Rank*		*Squared Rank*	
Present (X)	*New* (Y)	*Present* (U)	*New* (V)	*Present*	*New*	*Present*	*New*
10.8	10.8	.06	.01	4	2 (tie)	16	4
11.1	10.5	.36	.29	10	8	100	64
10.4	11.0	.34	.21	9	7	81	49
10.1	10.9	.64	.11	12	6	144	36
11.3	10.8	.56	.01	11	2 (tie)	121	4
	10.7		.09		5		25
	10.8		.01		2 (tie)		4
$\overline{X}$ = 10.74	$\overline{Y}$ = 10.79					T = 462	

$$T = \text{sum of the squared ranks (present)} = 462$$

$$\overline{R^2} = \frac{1}{12}(16 + 100 + \cdots + 25 + 4) = 54$$

$$\sum_{i=1}^{N} R_i^4 = (16^2 + 100^2 + \cdots + 25^2 + 4^2) = 60,660$$

$$T_1 = \frac{462 - 5(54)}{\left[\frac{(5)(7)}{(12)(11)}(60660) - \frac{(5)(7)}{11}(54)^2\right]^{1/2}} = 2.3273$$

- Reject H_0 since $T_1 > z_{0.95} = 1.645$
- p-value: $P(Z \geq 2.3273) = 0.01$
- Considerable simplification of the computations results whenever none of the values of U are tied with values of V, as in this example.
- Then ranks rather than average ranks can be used and the exact table consulted.

Theory

- Whenever two random variables X and Y are identically distributed except for having different means μ_1 and μ_2, $X - \mu_1$ and $Y - \mu_2$ not only have zero means, but they are identically distributed also.
- This means $U = |X - \mu_1|$ has the same distribution as $V = |Y - \mu_2|$. Both have the mean zero.
- Every assignment of ranks of the Us is equally likely.
- The ranks of Us and Vs are the same the ranks of U^2s and V^2s.
- Use the squared (score) ranks and not the ranks themselves.

$$a(R) = R^2$$

- $T = \sum a(R_i)$ where R_i denote the ranks of U_i in the combined sample.
- To use the large sample normal approximation for T it is necessary to find the mean and variance of T when H_0 is true.

- $E(T) = \sum_{i=1}^{n} E(a(R_i)) = n \sum_{j=1}^{N} \frac{1}{N} a(j) = n\bar{a}$
- $\text{Var}(T) = \sum_{i=1}^{n} \text{Var}[a(R_i)] + \sum_{i \neq j} \text{Cov}[a(R_i), a(R_j)]$
- $\text{Var}[a(R_i)] = \frac{1}{N} \sum_{k=1}^{N} [a(k) - \bar{a}]^2 = A$
- $\text{Cov}[a(R_i), a(R_j)] = \sum_{k \neq l} \frac{[a(k)-\bar{a}][a(l)-\bar{a}]}{N(N-1)}$
- $\text{Cov}[a(R_i), a(R_j)] = \sum_{k=1}^{N} [a(k) - \bar{a}] \sum_{l=1}^{N} [a(l) - \bar{a}] \frac{1}{N(N-1)} - \sum_{k=1}^{N} [a(k) - \bar{a}]^2 \frac{1}{N(N-1)}$. The first summation equals zero.
- $\text{Cov}[a(R_i), a(R_j)] = -\frac{A}{N-1}$
- $$\begin{aligned} \text{Var}(T) &= \sum_{i=1}^{n} A - \sum_{i \neq j}^{n} \frac{A}{N-1} \\ &= nA - n(n-1)\frac{A}{N-1} = \frac{n(N-n)}{N-1} A \\ &= \frac{nm}{(N-1)N} \sum_{i=1}^{N} [a(i) - \bar{a}]^2 \end{aligned}$$
- Interest in the case $a(R) = R^2$
- The denominator of Eq. 4 is what the square root of Eq. 24 by using
$$\sum_{i=1}^{N} [a(i) - \bar{a}]^2 = \sum_{i=1}^{N} [a(i)]^2 - N(\bar{a})^2$$
- The extension of the two-sample case to the k-sample case is completely analogous to the extension of the two-sample Mann-Whitney test to the k-sample Kruskal-Wallis test.
- $S_1, \ldots, S_k$: Sums of scores for each k samples.
- $E(S_i) = n_i \bar{a}$ and $\text{Var}(S_i) = \frac{n_i(N-n_i)}{(N-1)N} \sum_{i=1}^{N} [a(i) - \bar{a}]^2$
- $T_2 = \sum_{i=1}^{k} \frac{[S_i - E(S_i)]^2}{\text{Var}(S_i)} = \sum_{i=1}^{k} \frac{(S_i - n_i \bar{a})^2}{n_i D^2}$ where $D^2 = \frac{1}{N-1} \left\{ \sum_{i=1}^{N} [a(i)]^2 - N(\bar{a})^2 \right\}$
- $T_2 = \frac{1}{D^2} \left[\sum_{j=1}^{k} \frac{S_j^2}{n_j} - N(\bar{a})^2 \right]$
- If the populations of X and Y have the normal distributions the appropriate statistic to use is the ratio of the two sample variances:
$$F = \frac{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}{\frac{1}{m-1} \sum_{j=1}^{m} (Y_j - \bar{Y})^2}$$
- The F test is very sensitive to the assumption of normality.
- $P\{F_{m,n} \leq x\} = P\{F_{n,m} \geq 1/x\}$

- The F test is not very safe test to use unless one is sure that the populations are normal.

- If the squared rank test is used instead of the F test when the populations are normal the A.R.E. is only $15/(2\pi^2) = 0.76$, 1.08 for the double exponential distribution, 1.00 for the uniform distribution.

5.4 Measures of rank correlation

- A measure of correlation is a random variable that is used in situations where the data consist of pairs of numbers, such as in bivariate data.

- $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \sim (X, Y)$

- Examples of bivariate random variables include one where X_i, represents the height of the ith man and Y_i, represent his father's height, or where X_i, represents a test score of the ith individual and Y_i represents her amount of training.

- By tradition, a measure of correlation between X and Y should satisfy the following requirements in order to be acceptable.

  1. Values between -1 and 1.
  2. If the larger values of X tend to be paired with the larger values of Y, and hence the smaller values of X and Y tend to be paired together, then the measure of correlation should be positive, and close to $+1.0$ if the tendency is strong. Then we would speak of a positive correlation between X and Y.
  3. If the larger values of X tend to be paired with smaller values of Y, and vice versa, then the measure of correlation should be negative and close to -1.0 if the tendency is strong. Then we say that X and Y are negatively correlated.
  4. If the values of X seem to be randomly paired with values of Y, the measure of correlation should be fairly close to zero. This should be the case when X and Y are independent, and possibly some cases where X and Y are not independent. We then say that X and Y are uncorrelated, or have no correlation, or have correlation zero.

  Pearson's product moment correlation coefficient (most commonly used measure):

$$\begin{aligned} r &= \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2\right]^{1/2}} \\ &= \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\left(\sum X_i^2 - n\bar{X}^2\right)^{1/2} \left(\sum Y_i^2 - n\bar{Y}^2\right)^{1/2}} \\ &= \frac{\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^{1/2} \left[\frac{1}{n}\sum(Y_i - \bar{Y})^2\right]^{1/2}} \end{aligned}$$

- *Pearson's r* is a measure of the strength of the linear association between X and Y.

- If a plot of Y versus X shows the points (X, Y) all lie on, or close to, a straight line, then r will equal, or be close to 1.0 if the line is sloping upward, and -1.0 if the line is sloping downward.

- This measure of correlation may be used with any data of a numeric nature.
- The distribution function of r depends on the bivariate distribution function of (X, Y). Therefore r has no value as a test statistic in nonparametric tests or forming confidence intervals unless the distribution of (X, Y) is known.
- Some measures of correlation possess distribution functions that do not depend on the bivariate distribution function of (X, Y).

Spearman's Rho

Data

- $(X_i, Y_i), i = 1, 2, \ldots, n$
- $R(X_i)$: The rank of X_i as compared with the other X values.
- $R(Y_i)$: The rank of Y_i as compared with the other Y values.
- In case of ties, assign to each tied value the average of the ranks that would have been assigned if there had been no ties.

Measure of correlation (Spearman, 1904)

$$\rho = \frac{\sum R(X_i)R(Y_i) - n((n+1)/2)^2}{\left(\sum R(X_i)^2 - n((n+1)/2)^2\right)^{1/2}\left(\sum R(Y_i)^2 - n((n+1)/2)^2\right)^{1/2}} \tag{4}$$

If there are no ties:

$$\rho = 1 - \frac{6\sum[R(X_i) - R(Y_i)]^2}{n(n^2-1)} \tag{5}$$

- Spearman's ρ is what one obtains by replacing the observations by their ranks and then computing Pearson's r on the ranks.
- $\overline{R(X)} = \overline{R(Y)} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{n+1}{2}$
- Then Eq. 2 becomes Eq. 4.

Example 5.4.1 *Twelve MBA graduates are studied to measure the strength of the relationship between their score on the GMAT, which they took prior to entering graduate school, and their grade average while they were in the MBA program.*

- Their GMAT scores and their GPAs are given below:

Student	GMAT(X)	GPA(Y)	$R(X)$	$R(Y)$	$[R(X) - R(Y)]^2$
1	710	4.0	12	11.5	0.25
2	610	4.0	9.5	11.5	4
3	640	3.9	11	10	1
4	580	3.8	8	9	1
5	545	3.7	3	8	25
6	560	3.6	5	7	4
7	610	3.5	9.5	5	20.25
8	530	3.5	1	5	16
9	560	3.5	5	5	0
10	540	3.3	2	3	1
11	570	3.2	7	1.5	30.25
12	560	3.2	5	1.5	12.25

- There are ties involving about half of the observations.

$$\sum_{i=1}^{12}[R(X_i)]^2 = 647.5, \quad \sum_{i=1}^{12}[R(Y_i)]^2 = 647$$

- $\sum_{i=1}^{12} R(X_i)R(Y_i) = 589.75$
- From Eq. 4: $\rho = \frac{589.75-12(\frac{13}{2})^2}{(647.5-12(\frac{13}{2})^2)^{1/2}(647-12(\frac{13}{2})^2)^{1/2}} = 0.59$
- For comparison, Eq. 5: $\rho = 1 - \frac{6(115)}{12(12^2-1)} = 0.5979$ which is slightly larger than the more accurate $\rho = 0.5900$.
- Pearson's $r = 0.6630$ computed on the original data. The linear relationship between X and Y appears stronger than the linear relationship between the rank of X and the rank of Y. □

Hypothesis test The Spearman rank correlation coefficient is often used as a test statistic to test for independence between two random variables. The test statistic is given by Eq. 4.

Null distribution

- Exact quantiles of ρ when X and Y are independent are given in Table A10 for $n \leq 30$ and no ties.
- For larger n, or many ties: (percentile)

$$w_p \approx \frac{z_p}{\sqrt{n-1}}$$

Hypotheses Spearman's ρ is insensitive to some types of dependence, so it is better to be specific as to what type of dependency.

A. *Two-tailed test*

H_0: The X_i and Y_i are mutually independent

H_1: Either (a) there is a tendency for the larger values of X to be paired with the larger values of Y, or (b) there is a tendency for the smaller values of X to be paired with the larger values of Y

Reject H_0 at the level α if $|\rho|$ is greater than its $1-\alpha/2$ quantile obtained from Table A10 or Eq. 11.

B. *Lower-tailed test for negative correlation*

H_0: The X_i and Y_i are mutually independent

H_1: There is a tendency for the smaller values of X to be paired with the larger values of Y, and vice versa

Reject H_0 at the level α if $\rho < -w_{1-\alpha}$ where $w_{1-\alpha}$ is found either in Table A10 or from Eq. 11.

C. Upper-tailed test for positive correlation

H_0: The X_i and Y_i are mutually independent
H_1: There is a tendency for the larger values of X and Y to be paired together

Reject H_0 at the level α if $\rho > w_{1-\alpha}$ where $w_{1-\alpha}$ is found either in Table A10 or from Eq. 11.

Example 5.4.2 *Let us continue with Example 1. Suppose the twelve MBA graduates are a random sample from all recent MBA graduates, and we want to know if there is a tendency for high GPAs to be associated with high GMAT scores.*

H_0 : *GPAs are independent of GMAT scores*
H_1 : *High GPAs tend to be associated with high GMAT scores*

- For $n = 12$, $w_{0.95} = 0.4965$ from Table A10.
- Normal approximation: $w_{0.95} = \frac{1.6449}{\sqrt{11}} = 0.4960$
- The observed value of $\rho = 0.59$.
- Reject H_0 at $\alpha = 5\%$.

$$\begin{aligned} p\text{-value} &= P(Z \geq 0.5900\sqrt{11}) \\ &= P(Z \geq 1.9568) = 0.025 \end{aligned}$$ □

- The next measure of correlation resembles Spearman's ρ in that it is based on the ranks of the observations rather than the numbers themselves, and the distribution of the measure does not depend on the distribution of X and Y if X and Y are independent and continuous. Advantages:
  - The distribution of Kendall's tau approaches the normal distribution quite rapidly so that the normal approximation is better for the Kendall's τ than it is for Spearman's ρ.
  - Another advantage of Kendall's τ is its direct and simple interpretation in terms of probabilities of observing concordant and discordant pairs.

Kendall's tau

Data

- (X_i, Y_i), $i = 1, 2, \ldots, n$
- Two observations, for example $(1.3, 2.2)$ and $(1.6, 2.7)$, are called *concordant* if both members of one observation are larger than their respective members of the other observation.
- N_c: Number of concordant pairs of observations.
- A pair of observations, for example $(1.3, 2.2)$ and $(1.6, 1.1)$, are called *discordant* if the two numbers in one observation differ in opposite directions from the respective members in the other observation.
- N_d: Number of discordant pairs of observations.

- ▷ $N_c + N_d = n(n-1)/2$

Measure of correlation (Kendall, 1938)

- ▷ In case of no ties: $\tau = \frac{N_c - N_d}{n(n-1)/2}$
- ▷ $\tau = 1$ if all pairs are concordant.
- ▷ $\tau = -1$ if all pairs are discordant.

Ties

$$\tau = \frac{N_c - N_d}{N_c + N_d}$$

- ▷ If $\frac{Y_j - Y_i}{X_j - X_i} > 0$, add 1 to N_c (concordant).
- ▷ If $\frac{Y_j - Y_i}{X_j - X_i} < 0$, add 1 to N_d (discordant).
- ▷ If $\frac{Y_j - Y_i}{X_j - X_i} = 0$, add 1/2 to N_c and N_d.
- ▷ $X_i = X_j$, no comparison is made.

- ▶ The computation of τ is simplified if the observations (X_i, Y_i) are arranged in a column according to increasing values of X.

Example 5.4.3 *Again we will use the data in Example 1 for purpose of illustration.*

- ▶ Arrangement of the data (X_i, Y_i) according to increasing values of X gives the following:

	Concordant pairs	Discordant pairs
X_i, Y_i	below (X_i, Y_i)	below (X_i, Y_i)
(530, 3.5)	7	4
(540, 3.3)	8	2
(545, 3.7)	4	5
(560, 3.2)	5.5	0.5
(560, 3.5)	4.5	1.5
(560, 3.6)	4	2
(570, 3.2)	5	0
(580, 3.8)	3	1
(610, 3.5)	2	0
(610, 4.0)	0.5	1.5
(640, 3.9)	1	0
(740, 4.0)		
	$N_c = 44.5$	$N_d = 17.5$

- ▶ *Kendall's τ*:

$$\tau = \frac{N_c - N_d}{N_c + N_d} = \frac{44.5 - 17.5}{44.5 + 17.5} = 0.4355$$

- ▶ There is a positive rank correlation between the GMAT scores and the GPAs as measured by Kendall's τ. □

Hypothesis test

- *Kendall's τ* may also be used as a test statistic to test the null hypothesis of independence between X and Y.

$$\begin{aligned} T &= N_c - N_d, \quad \text{in case of no ties or few ties,} \\ \tau &= \frac{N_c - N_d}{N_c + N_d}, \quad \text{in case of many ties.} \end{aligned}$$

- Exact upper quantiles for τ and T when X and Y are independent are given in Table 11 for $n \leq 60$ in case of no ties.
- Lower quantiles are the negative of the upper quantiles given in the table.
- For larger n or many ties the pth quantile of τ is given approximately by

$$w_p = z_p \, \frac{\sqrt{2(2n+5)}}{3\sqrt{n(n-1)}}$$

The pth quantile of T is given approximately by

$$w_p = z_p \, \sqrt{n(n-1)(2n+5)/18}$$

Hypotheses

A. (Two-tailed test)

H_0: X and Y are independent

H_1: Pairs of observations either tend to be concordant, or tend to be discordant.

- Reject H_0 at the level α if T (or τ) is less than its $\alpha/2$ quantile or greater than its $1-\alpha/2$ quantile in the null distribution.
- p-value: Twice the smaller of the one-tailed p-value, given approximately by

$$\begin{aligned} p(\text{lower-tailed}) &= P\left(Z \leq \frac{(T+1)\sqrt{18}}{\sqrt{n(n-1)(2n+5)}}\right) \\ p(\text{upper-tailed}) &= P\left(Z \geq \frac{(T-1)\sqrt{18}}{\sqrt{n(n-1)(2n+5)}}\right) \end{aligned}$$

B. (Lower-tailed test)

H_0: X and Y are independent

H_1: Pairs of observations trend to be discordant.

Reject H_0 at the level α if T (or τ) is less than its α quantile in the null distribution.

C. (Upper-tailed test)

H_0: X and Y are independent

H_1: Pairs of observations trend to be concordant.

Reject H_0 at the level α if T (or τ) is less than its $1-\alpha$ quantile in the null distribution.

Example 5.4.4 *In Example 3 Kendall's τ was computed by first finding the value of*

$$T = N_c - N_d = 44.5 - 17.5 = 27.$$

- ▶ If we are interested in using T to test the null hypothesis of independence between the student's GMAT score and his or her GPA, to see if higher GPAs tend to be associated with higher GMAT scores, then the null hypothesis is rejected at $\alpha = 0.05$ if $T > w_{.95} = 24$ by Table A11.

$$\begin{aligned} p\text{-value} &= P(T \geq 27) \\ &= P\left(Z \geq \frac{(27-1)\sqrt{18}}{\sqrt{12 \cdot 11 \cdot 29}}\right) \\ &= P(Z \geq 1.7829) \\ &= 0.037 \end{aligned}$$

- ▶ If we use τ as the test statistic, the results are similar. □

Compare Spearman's ρ and Kendall's τ:

- ▶ $\rho = 0.59$ is a larger number than Kendall's $\tau = 0.4355$.
- ▶ Two tests using the two statistic produced nearly identical results.
- ▶ Both of the preceding statements hold true in most, but not all, situations.
- ▶ Spearman's ρ tends to be larger than Kendall's τ, in absolute value.

The Daniel's test for trend

- ▶ Daniels (1950) proposed the use of the Spearman's ρ to test for trend by pairing measurements called X_i, with the time (or order) at which the measurements are taken.
- ▶ X_is are mutually independent, and the null hypothesis is that they are identically distributed.
- ▶ The alternative hypothesis is that the distribution of X_i is related to time so that as times goes on, the X measurements tend to become larger (smaller).
- ▶ Tests of trend based on Spearman's ρ or Kendall's τ are generally considered to be more powerful than the Cox and Stuart test (Sec. 3.5).
- ▶ The A.R.E. of the Cox and Stuart test for trend, when applied to random variables known to be normally distributed, is about 0.78 with respect to the test based on the regression coefficient, while the A.R.E. of these tests using Spearman's ρ or Kendall's τ is about 0.98 under the same conditions.
- ▶ These tests are not as widely applicable as the Cox and Stuart test. For instance, these tests would be inappropriate in Eg. 3.5.3. These tests are appropriate in Eg. 3.5.2.

Example 5.4.5 *In Example 3.5.2, nineteen years of annual precipitation records are given. The two-tailed test for trend involves rejection of the null hypothesis of no trend if Spearman's ρ is too large or too small.*

- The test statistic is given by Eq. 5 because the number of ties is small.

$$\begin{aligned} T &= \sum_{i=1}^{19}[R(X_i) - R(Y_i)]^2 = 1241.5 \\ \rho &= 1 - \frac{6T}{19(19^2-1)} = -0.2469 \\ w_{0.975} &= 0.4579 \\ w_{0.025} &= -0.4579 \end{aligned}$$

- H_0 is accepted.

$$\begin{aligned} p\text{-value} &\cong 2 \cdot P(Z \geq 0.2469\sqrt{19-1}) \\ &= 2(0.147) \\ &= 0.294 \end{aligned}$$

The Jonckheere-Terpstra test

- Either Spearman's ρ or Kendall's τ can be used in the case of several independent samples to test the null hypothesis that all of the samples came from the same distribution.

$$H_0: \ F_1(x) = F_2(x) = \cdots = F_k(x)$$

  against the ordered alternative that the distributions differ in a specified direction

$$H_1: \ F_1(x) \geq F_2(x) \geq \cdots \geq F_k(x)$$

  with at least one inequality.

- The alternative is sometimes written as

$$H_1(x): \ E(Y_1) \leq E(Y_2) \leq \cdots \leq E(Y_k).$$

- The same data setup and the same null hypothesis as in the Kruskal-Wallis test of Section 5.2.

- The Kruskal-Wallis test is sensitive against any differences in means, while this usage of Spearman's ρ or Kendall's τ is sensitive against only the ordering specified in the H_1 given above.

- When Kendall's τ is used, this test is equivalent to the Jonckheere-Terpstra test, which is found in the computer programs *SAS* and *StatXact*.

Example 5.4.6 *As the human eye ages, it loses its ability to focus on objects close to the eye. This is well-recognized characteristic of people over* 40 *years old. In order to see if people in the* 15*- to* 30*-year-old range also exhibit this loss of ability to focus on nearby objects as they get older, eight people were selected from each of four age groups; about* 15 *years old, about* 20*, about* 25*, and about* 30 *years old. It was assumed that these people would behave as a random sample from their age group populations would, with regard to the characteristic being measured. Each person held a printed paper in front of his or her right eye, with left eye covered. The paper was moved closer to the eye until the person declared that the print began to look fuzzy.*

- The closest distance at which the print was still sharp was measured once for each person.

$$\begin{aligned} H_0 &: F_1(x) = F_2(x) = F_3(x) = F_4(x), && \text{for all } x \\ H_1 &: F_i(x) > F_j(x), && \text{for some } x \text{ and some } i < j \end{aligned}$$

- Assume the ability to focus on close objects does not improve with age. The distances are measured in inches.

15 years old		20 years old		25 years old		30 yrs	
X	Y	X	Y	X	Y	X	Y
1	4.6	2	4.7	3	5.6	4	6.0
1	4.9	2	5.0	3	5.9	4	6.8
1	5.0	2	5.1	3	6.6	4	8.1
1	5.7	2	5.8	3	6.7	4	8.4
1	6.3	2	6.4	3	6.8	4	8.6
1	6.8	2	6.6	3	7.4	4	8.9
1	7.4	2	7.1	3	8.3	4	9.8
1	7.9	2	8.3	3	9.6	4	11.5

- If the minimum focusing distance Y tends to get larger with age, then Y and X should show a positive correlation, using either Spearman's ρ or Kendall's τ.
- We have an upper-tailed test with lots of ties among the Xs.
- Instead of $X = 1, 2, 3, 4$ we could have used any increasing sequence of numbers for different samples, such as $X = 15, 20, 25, 30$, and both ρ and τ will not be affected by the change in X values.
- Reject H_0 since $\rho = 0.5680 >$ the approximate 5% upper-tailed test is $1.6449/\sqrt{31} = 0.2954$.
- Kendall's τ for these data, based on $N_c = 290.5$ and $N_d = 93.5$, is $\tau = 0.5130$. Reject H_0 since $0.5130 > w_{0.95} = 0.2056$.
- The upper-tailed p-value is less than 0.001.
- If (X_i, Y_i) are independent and identically distributed bivariate normal random variables, both ρ and τ have an asymptotic relative efficiency of $9/\pi^2 = 0.912$, relative to the parametric test for independence that uses Pearson's r as a test statistic.

Kendall's partial correlation coefficient

- Multivariate random variable $(X_1, X_2, \dots, X_k)$.
- Correlation between X_1 and X_2, between X_2 and X_5.
- Those measures estimate the total influence of one variable on the other.
- Sometimes it is desirable to measure the correlation between two random variables, under the condition that the indirect influence due to the other random variables is somehow eliminated.
- An estimate of this "partial" correlation between X_1 and X_2, say, while the indirect correlation due to $X_3, X_4, \dots, X_n$ is denoted by $r_{12.34\dots n}$ when using the extension of Pearson's r, or by $\tau_{12.34\dots n}$ when using Kendall's τ.
- $n = 3$, Pearson's partial correlation coefficient

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

- $n = 3$, Kendall's τ partial correlation coefficient

$$\tau_{12.3} = \frac{\tau_{12} - \tau_{13}\tau_{23}}{\sqrt{(1 - \tau_{13}^2)(1 - \tau_{23}^2)}}$$

5.5 Nonparametric linear regression methods

- This section is related closely to the previous section on rank correlation.
- Examine a random sample $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ on the bivariate random variable (X, Y).
- Correlation methods emphasize estimating the degree of dependence between X and Y.
- Regression methods are used to inspect the relationship between X and Y more closely.
- One important objective of regression methods is to predict a value of Y in the pair (X, Y) where only the value for X is known, on the basis of information that we can obtain from previous observations (X_1, Y_1) through (X_n, Y_n).
- If X represents the scores on a college entrance examination and Y represents the grade point average of that student four years later, observations on past students may help us to predict how well an incoming student will perform in the four years of college.
- Of course, Y is still a random variable, so we cannot expect to determine Y solely from knowing the associated value of X, but knowing X should help us make a better estimate concerning Y.
- Regression methods also apply to controlled experiments where X may not be random at all, but may be set by the experimenter at various values to determine its effect on Y.

- X may represent a measured amount of medication, such as medication intended to lower blood pressure in a patient.

- Several different levels of X may be selected in an experiment to determine the effect of the medication on Y, which is the patient's response such as the patient's reduction in blood pressure.

Definition 5.5.1 *The regression of Y on X is $E(Y|X = x)$. The regression equation is $y = E(Y|X = x)$.*

- If the regression equation is known, we can represent the regression on a graph by plotting y as the ordinate and x as the abscissa.

- The regression equation is seldom known.

- It is estimated on the basis of past data.

- If we would like to predict Y when $X = 6$, we could use $E(Y|X = 6)$ if we knew it; otherwise, we could use the sample mean or the sample median of several observed values of Y for which X is equal to 6 or close to 6.

- In this way point estimates and confidence intervals may be formed for $E(Y|X = 6)$ using the methods described in Sections 3.2 and 5.7.

- In order to have enough observations so that the regression of Y on X can be estimated for each value of X, many observations are needed.

- A more difficult situation arises when we have only a few observations and wish to estimate the regression of Y on X.

- This is what we will examine in this section.

- It is helpful to know something about the relationship between $E(Y|X = x)$ and x and to be able to use this information when there are only a few observations.

- First we will examine the case where $E(Y|X = x)$ is a linear function of X; in the next section we will consider a more general situation where $E(Y|X = x)$ is a monotonic function of X.

Definition 5.5.2 *The regression of Y on X is linear regression if the regression equation is of the form*

$$E(Y|X = x) = \alpha + \beta x$$

for some constant α, called the y-intercept, and β, called the slope.

- Usually the constants α and β are unknown and must be estimated from the data.

- A commonly accepted method for estimating α and β is called the *least squares method*.

Definition 5.5.3 *The least squares method for choosing estimates a and b of α and β in the regression equation $y = \alpha + \beta x$ is the method that minimizes the sum of squared deviations*

$$SS = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

for the observations $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$.

- The idea behind the least squares method is that an estimate of the regression line should be close to the observed values of X and Y because the true regression line is probably close the observation.
- $SS = \sum_{i=1}^{n} D_i^2$ where $D_i = Y_i - (a + bX_i)$.

Nonparametric methods for linear regression
Data The data consist of a random sample $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ from some bivariate distribution.
Assumptions

1. The sample is a random sample. The methods of this section are valid if the values of X are nonrandom quantities as long as the Ys are independent with identical conditional distributions.
2. The regression of Y on X is linear. This implies an interval scale of measurement on both X and Y.

Least squares estimates

$$\begin{aligned} y &= a + bx \\ b &= \frac{\text{Cov}(X, Y)}{S_x^2} = \rho \frac{S_y}{S_x} = \frac{n \sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \\ a &= \bar{Y} - b\bar{X} \end{aligned}$$

Testing the slope To test the hypothesis concerning the slope, add the following assumption to Assumptions 1 and 2.

3. The residual $Y - E(Y|X)$ is independent of X.

Spearman's ρ may be adapted to test the following hypotheses concerning the slope. Let β_0 represent some specified number. For each pair (X_i, Y_i) compute $Y_i - \beta_0 X_i = U_i$. Then find the Spearman rank correlation coefficient ρ on the pairs (X_i, U_i). Table A10 gives the quantiles of ρ when H_0 is true and there are no ties.

A. Two-tailed test

$$H_0 : \beta = \beta_0 \quad v.s. \quad H_1 : \beta \neq \beta_0$$

Reject H_0 if ρ exceeds the $1 - \alpha/2$ quantile, or less than the $\alpha/2$ quantile as described in the two-tailed test for Spearman's ρ in Section 5.4.

B. Lower-tailed test

$$H_0 : \beta = \beta_0 \quad v.s. \quad H_1 : \beta < \beta_0$$

Reject H_0 if ρ is less than the α quantile as described in the lower-tailed test for Spearman's ρ in Section 5.4.

C. Upper-tailed test

$$H_0 : \beta = \beta_0 \quad v.s. \quad H_1 : \beta > \beta_0$$

Reject H_0 if ρ exceeds the $1-\alpha$ quantile as described in the two-tailed test for Spearman's ρ in Section 5.4.

A confidence interval for the slope For each pair of points (X_i, Y_i) and (X_j, Y_j), such that $i < j$ and $X_i \neq X_j$, compute the two-point slope

$$S_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$$

- N: The number of slopes computed.
- Order the slopes obtained and let

$$S^{(1)} \leq S^{(2)} \leq \cdots \leq S^{(N)}$$

- Find $w_{1-\alpha/2}$ from Table A11.
- $r = (N - w_{1-\alpha/2})/2$ and $s = (N + w_{1-\alpha/2})/2 + 1 = N + 1 - r$.
- $1-\alpha$ CI for β:

$$(S^{(r)}, S^{(s)})$$

Example 5.5.1 *Let us again use the data from the previous section. The GMAT score of each MBA graduates is denoted by X_i and that graduate's GPA is denoted by Y_i. The twelve observations (X, Y) are The twelve observations (X, Y) are* $(710, 4.0), (610, 4.0), (640, 3.9), (580, 3.8), (545, 3.7), (560, 3.6), (610, 3.5), (530, 3.5), (560, 3.5), (540, 3.3),$ *and* $(560, 3.2)$. *These are plotted in Figure 1.*

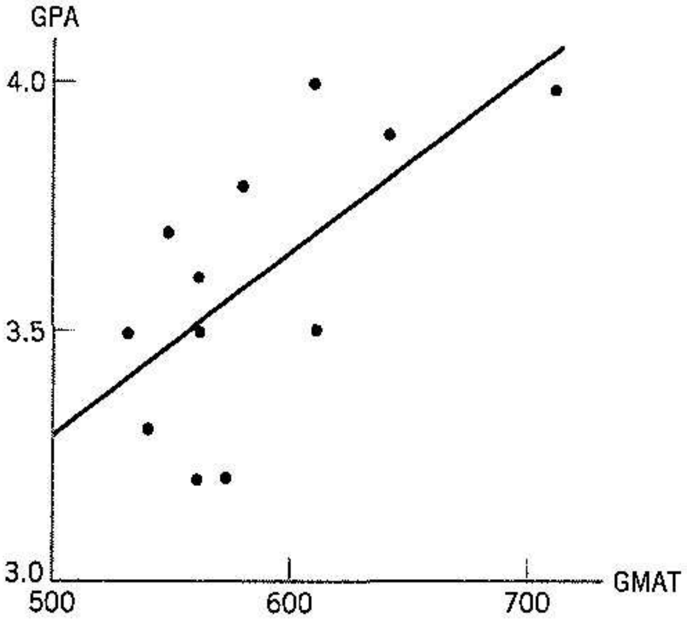


Figure 5.4: *A scatterplot of GMAT scores versus GPAs for 12 MBA graduates, and the least squares regression line.*

- Least squares regression line:

$$y = 1.4287 + .003714x$$

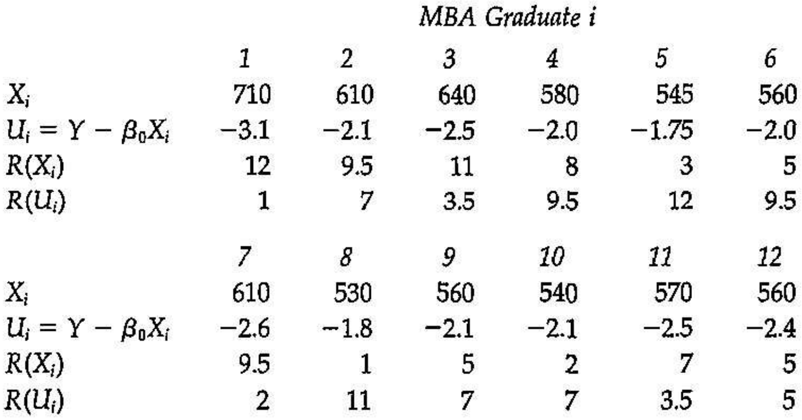
$$\sum_{i=1}^{12} X_i = 7015, \qquad \bar{X} = 584.58, \qquad \sum_{i=1}^{12} X_i^2 = 4129525$$

$$\sum_{i=1}^{12} Y_i = 43.2, \qquad \bar{Y} = 3.6, \qquad \sum_{i=1}^{12} X_iY_i = 25360.5$$

- Suppose that a national study reports that ”a 40 point increase in GMAT scores results in at least 0.4 increase in GPAs.”
- Because slope is a change in Y divided by a change in X, this claim is equivalent to saying the slope in the regression of GPA onto GMAT score is at least $0.4/40 = 0.01$.
- To see if our sample of 12 graduates is consistent with the national study we test

$$H_0 : \beta \geq 0.01 \quad v.s. \quad H_1 : \beta < 0.01$$

- Spearman's ρ calculated between the GMAT scores X and the sample residuals $U = Y - (0.01)X$



	MBA Graduate i											
	1	2	3	4	5	6	7	8	9	10	11	12
X_i	710	610	640	580	545	560	610	530	560	540	570	560
$U_i = Y - \beta_0 X_i$	−3.1	−2.1	−2.5	−2.0	−1.75	−2.0	−2.6	−1.8	−2.1	−2.1	−2.5	−2.4
$R(X_i)$	12	9.5	11	8	3	5	9.5	1	5	2	7	5
$R(U_i)$	1	7	3.5	9.5	12	9.5	2	11	7	7	3.5	5

- Equation 5.4.4 is used to compute $\rho = -0.7273$ which is less than the 0.05 quantile of the null distribution from Table A10.
- H_0 is rejected at $\alpha = 0.05$.
- p-value:

$$P(Z \leq -0.7273\sqrt{11}) = P(Z \leq -2.4121) = 0.008$$

- 95% CI for β:
  - all of the two-point slopes

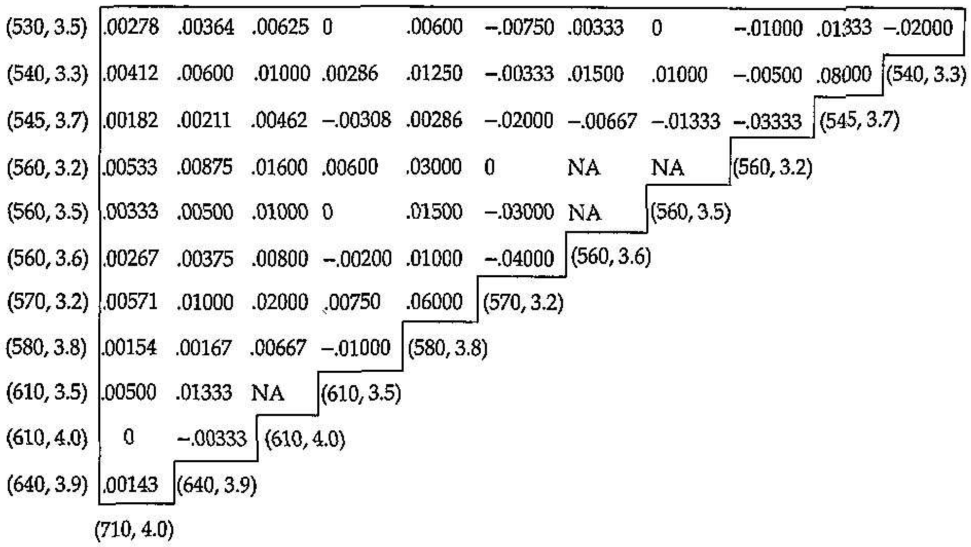
    $$S_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$$

    are computed for pairs of points where $X_i \neq X_j$.
  - See Figure 2 for a convenient spreadsheet-like layout for computing the S_{ij}s.
  - There are $N = 62$ pairs (X_i, Y_i) and (X_i, Y_i) with $X_i \neq X_j$, as seen in Figure 2.

- ▷ Compute $S_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$
- ▷ From Table A11 for $n = 12$, the 0.975 quantile of T is 28.

$$r = \frac{1}{2}(N - w_{1-\alpha/2}) = 17$$
$$s = \frac{1}{2}(N + w_{1-\alpha/2}) = N + 1 - r = 46$$

- ▷ 95% CI for β is $\left(S^{(17)}, S^{(46)}\right) = (0.00000, 0.00800)$



(530, 3.5)	.00278	.00364	.00625	0	.00600	−.00750	.00333	0	−.01000	.01333	−.02000
(540, 3.3)	.00412	.00600	.01000	.00286	.01250	−.00333	.01500	.01000	−.00500	.08000	(540, 3.3)
(545, 3.7)	.00182	.00211	.00462	−.00308	.00286	−.02000	−.00667	−.01333	−.03333	(545, 3.7)	
(560, 3.2)	.00533	.00875	.01600	.00600	.03000	0	NA	NA	(560, 3.2)		
(560, 3.5)	.00333	.00500	.01000	0	.01500	−.03000	NA	(560, 3.5)			
(560, 3.6)	.00267	.00375	.00800	−.00200	.01000	−.04000	(560, 3.6)				
(570, 3.2)	.00571	.01000	.02000	.00750	.06000	(570, 3.2)					
(580, 3.8)	.00154	.00167	.00667	−.01000	(580, 3.8)						
(610, 3.5)	.00500	.01333	NA	(610, 3.5)							
(610, 4.0)	0	−.00333	(610, 4.0)								
(640, 3.9)	.00143	(640, 3.9)									
	(710, 4.0)										

Figure 5.5: *A spreadsheet arrangement of points (X, Y), arranged by increasing Xs, to find the value of S_{ij}.*

Theory

- ▶ To derive a and b such that SS in Equation 2 is minimized, add and subtract the quantity $(\bar{Y} - b\bar{X})$ inside the brackets to get

$$SS = \sum_{i=1}^{n} [(Y_i - \bar{Y}) - b(X_i - \bar{X}) + (\bar{Y} - b\bar{X} - a)]^2$$

- ▶ Because of the algebraic identity

$$(c - d + e)^2 = c^2 + d^2 + e^2 - 2cd + 2ce - 2de$$

we can expand Equation 14 using $c = Y_i - \bar{Y}$ and so on, to get

$$\begin{aligned} SS = &\sum_{i=1}^{n}(Y_i - \bar{Y})^2 + b^2 \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{Y} - b\bar{X} - a)^2 \\ &- 2b \sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) + 2(\bar{Y} - b\bar{X} - a) \sum_{i=1}^{n}(Y_i - \bar{Y}) \\ &- 2b(\bar{Y} - b\bar{X} - a) \sum_{i=1}^{n}(X_i - \bar{X}) \end{aligned}$$

- Because $\sum(Y_i - \bar{Y}) = 0$ and $\sum(X_i - \bar{X}) = 0$ by the definition of $\bar{X}$ and $\bar{Y}$ the last two summations equal zero in Equation 16. The third summation is smallest (zero) when

$$a = \bar{Y} - b\bar{X}$$

which gives the least squares solution for a.

- We are left with the problem of finding the value of b that minimizes the sum of the second and fourth summations, that is, that minimizes

$$b^2 S_x - 2bS_{xy}$$

where

$$S_x = \sum_{i=1}^{n}(X_i - \bar{X})^2$$

and

$$S_{xy} = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

- By adding and subtracting S_{xy}^2/S_x to Equation 18, the sum of the second and fourth summations becomes

$$S_x\left[b^2 - 2b\frac{S_{xy}}{S_x} + \left(\frac{S_{xy}}{S_x}\right)^2\right] - \frac{S_{xy}^2}{S_x} = S_x\left(b - \frac{S_{xy}}{S_x}\right)^2 - \frac{S_{xy}^2}{S_x}$$

which is obviously a minimum when

$$b = \frac{S_{xy}}{S_x}$$

in agreement with Equation 6.

- Note that this reduces the second and fourth summation to $-S_{xy}^2/S_x$, so that the minimum sum of squares is

$$\begin{aligned} SS_{\min} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \frac{S_{xy}^2}{S_x} \\ &= (1 - r^2)\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \end{aligned}$$

where r is the Pearson product moment correlation coefficient given by Equation 5.4.1.

- Also note that no assumptions regarding the distribution of (X, Y) were made, so the least squares method is distribution-free.

- In fact, the only purpose of assumptions 1 and 2 is to assure us that there is a regression line somewhere to be estimated.

- Under assumption 3, the residuals

$$Y_i - E[Y_i|X_i] = Y_i - (\alpha + \beta X_i)$$

are independent of X_i, so the assumptions of Section 5.4 regarding Spearman's ρ are met.

- ▶ Note that the ranks of $(Y_i - \alpha - \beta X_i),\ i = 1$ to n, are the same as the ranks of $U_i = (Y_i - \beta X_i),\ i = 1$ to n, so we can test $H_0 : \beta = \beta_0$ without knowing α.

- ▶ Just as Spearman's ρ is merely Pearson's r computed on ranks, this test is the rank analogue of computing r on the pairs (X_i, U_i) which is the usual parametric procedure for testing the same null hypothesis, valid with the additional assumption that (X, Y) has the bivariate normal distribution.

- ▶ Under that condition and the condition that the observations on X are equally spaced, the A.R.E. of this procedure is $(3/\pi)^{1/3} = 0.98$ according to Stuart (1954, 1956); for other distributions the A.R.E. is always greater than or equal to 0.95 (Lehmann, 1975).

- ▶ To see the relationship between the slopes S_{ij} and Kendall's τ, note that for any hypothesized slope β_0 we have

$$\begin{aligned} S_{ij} &= \frac{Y_i - Y_j}{X_i - X_j} = \frac{U_i + \beta_0 X_i - U_j - \beta_0 X_j}{X_i - X_j} \\ &= \beta_0 + \frac{U_i - U_j}{X_i - X_j} \end{aligned}$$

where $U_i = Y_i - \beta_0 X_i - \alpha$ is the residual of Y_i from the hypothesized regression line $y = \alpha + \beta_0 x$.

- ▶ The slope S_{ij} is greater than β_0 or less than β_0 according to whether the pair (X_i, U_i) and (X_j, U_j) is concordant or discordant in the sense described in Section 5.4 in the discussion of Kendall's τ.

- ▶ If we use the number of S_{ij} less than β_0 as our test statistic for determining whether to accept $H_0 : \beta = \beta_0$, we accept β_0 as long as the number of discordant pairs N_d is not too small or too large.

- ▶ Because N_d is related to the number of concordant pairs N_c by

$$N_c + N_d = N$$

where N is the total number of pairs, and because the quantiles of $N_c - N_d$ are given in Table A11 if we have the true slope and Assumption 3 of independence, we can say N_d is too small if $N_c - N_d$ is greater than $\omega_{1-\alpha/2}$ from Table A11. This is equivalent to saying N_d is less than $r = (N - \omega_{1-\alpha/2})/2$.

- ▶ In other words, β_0 is acceptable if β_0 is greater than at least r of the S_{ij}, or $\beta_0 > S^{(r)}$.

- ▶ The same argument gives an upper bound for β_0, and the confidence interval is obtained.

- ▶ This method, due to Theil (1950), was modified to handle ties by Sen (1968a). □

5.6 Methods for monotonic regression

- In Section 5.5 nonparametric methods for linear regression were presented. These may be used in situations such as in Example 5.5.1, where the assumption of linear regression seems reasonable.
- In other situations it may be unreasonable to assume that the regression function is a straight line, but it may be reasonable to assume that $E(Y|X)$ increases (at least, it does not decrease) as X increases.
- In such a case we say the regression is *monotonically increasing*.
- If $E(Y|X)$ becomes smaller as X increases the regression is *monotonically decreasing*.

Nonparametric methods for monotonic regression

Data The data consist of a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from some bivariate distribution.

Assumptions

1. The sample is a random sample.
2. The regression of Y on X is monotonic.

An estimate of $E(Y|X)$ at a Point To estimate the regression of Y on X at a particular value of $X = x_0$:

1. Obtain the ranks $R(X_i)$ of the Xs and $R(Y_i)$ of the Ys. Use average ranks in case of ties.
2. Find the least squares regression line on the ranks.

$$y = a_2 + b_2 x \tag{1}$$

where

$$b_2 = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n(n+1)^2/4}{\sum_{i=1}^{n}[R(X_i)]^2 - n(n+1)^2/4} \tag{2}$$

and

$$a_2 = (1 - b_2)(n+1)/2. \tag{3}$$

3. Obtain a rank $R(x_0)$ for x_0 as follows:

   (a) If x_0 equals one of the observed X_is, let $R(x_0)$ equal the rank of that X_i.

   (b) If x_0 lies between two adjacent values X_i and X_j where $X_i < x_0 < X_j$, interpolate between their respective ranks to get $R(x_0)$:

$$R(x_0) = R(X_i) + \frac{x_0 - X_i}{X_j - X_i}[R(X_j) - R(X_i)]. \tag{4}$$

   This "rank" will not necessarily be an integer.

   (c) If x_0 is less than the smallest observed X or greater than the largest observed X, do not attempt to extrapolate. Information on the regression of Y on X is available only within the observed range of X.

4. Substitute $R(x_0)$ for x in Equation 1 to get an estimated rank $R(y_0)$ for the corresponding value of $E(Y|X = x_0)$.

$$R(y_0) = a_2 + b_2 R(x_0) \tag{5}$$

5. Convert $R(y_0)$ into $\hat{E}(Y|X = x_0)$, an estimate of $E(Y|X = x_0)$, by referring to the observed Y_is as follows.
   (a) If $R(y_0)$ equals the rank of one of the observations Y_i, let the estimate $\hat{E}(Y|X = x_0)$ equal that observation Y_i.
   (b) If $R(y_0)$ lies between the ranks of two adjacent values of Y, say Y_i and Y_j where $Y_i < Y_j$, so that $R(Y_i) < R(y_0) < R(Y_j)$, interpolate between Y_i and Y_j.
   $$\hat{E}(Y|X = x_0) = Y_i + \frac{R(y_0) - R(Y_i)}{R(Y_j) - R(Y_i)}(Y_j - Y_i) \tag{6}$$
   (c) If $R(y_0)$ is greater than the largest observed rank of Y, let $\hat{E}(Y|X = x_0)$ equal the largest observed Y. If $R(y_0)$ is less than the smallest observed rank of Y, let $\hat{E}(Y|X = x_0)$ equal the smallest observed Y.

An Estimate of the Regression of Y on X To obtain the entire regression curve consisting of all points that can be obtained in the manner just described, the following procedure may be used.

1. For each X_i from $X^{(1)}$ to $X^{(n)}$ use the previously described procedure to estimate $E(Y|X)$.

2. For each rank of Y, $R(Y_i)$, find the estimated rank of X_i, $\hat{R}(X_i)$ from Equation 1.
$$\hat{R}(X_i) = [R(Y_i) - a_2]/b_2, \quad i = 1, 2, \ldots, n \tag{7}$$

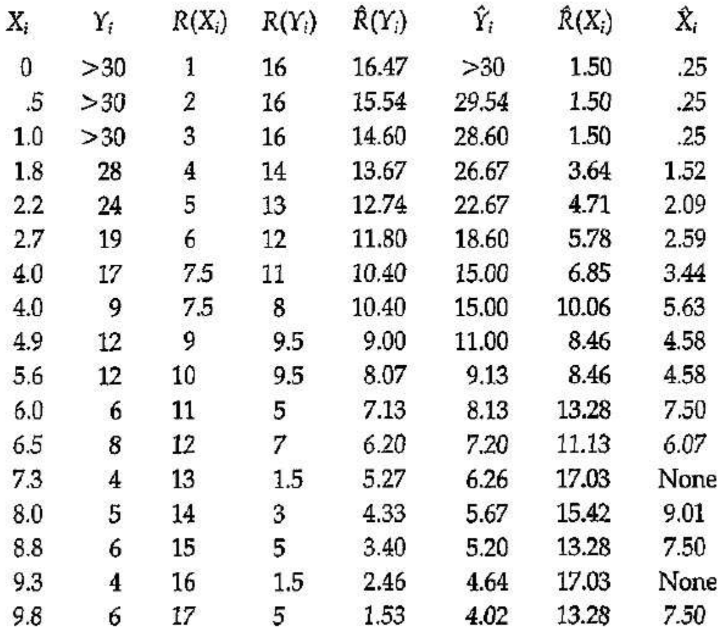
3. Convert each $\hat{R}(X_i)$ to an estimate $\hat{X}_i$ in the manner of the preceding step 5. More specifically:
   (a) If $\hat{R}(X_i)$ equals the rank of some observation X_j, let $\hat{X}_i$ equal that observed value.
   (b) If $\hat{R}(X_i)$ falls between the ranks of two adjacent observations X_j and X_k, where $X_j < X_k$, then use interpolation,
   $$\hat{X}_i = X_j + \frac{\hat{R}(X_i) - R(X_j)}{R(X_k) - R(X_j)}(X_k - X_j) \tag{8}$$
   to get $\hat{X}_i$.
   (c) If $\hat{R}(X_i)$ is less than the smallest observed rank of X or greater than the largest observed rank, no estimate $\hat{X}_i$ is found.

4. Plot each of the points found in steps 1 and 3 on graph paper. That is, plot each $(X_i, \hat{Y}_i)$ and each $(\hat{X}_i, Y_i)$. All of these points should be monotonic, increasing if $b_2 > 0$ and decreasing if $b_2 < 0$.

5. Connect the adjacent points in step 4 with straight lines. This series of connected line segments is the estimate of the regression of Y on X.

Example 5.6.1 *Seventeen jars of fresh grape juice were obtained to study how long it took for the grape juice to turn into wine as a function of how much sugar was added to the juice. Various amounts of sugar, ranging from none to about* 10 *pounds, were added to the jars, and each day the jars were checked to see if the transition to wine was complete. At the end of* 30 *days the experiment was terminated, with three jars still unfermented. An estimate of the regression curve of Y (number of days till fermentation) v.s. X (pounds of sugar) is desired.*

- $(X_i, Y_i), R(X_i), R(Y_i)$, and the values $\hat{R}(Y_i), \hat{Y}_i = \hat{E}(Y|X_i), \hat{R}(X_i)$, and $\hat{X}_i$ computed from the preceding steps 1, 2, and 3 are given in Fig. 5.6.



X_i	Y_i	$R(X_i)$	$R(Y_i)$	$\hat{R}(Y_i)$	$\hat{Y}_i$	$\hat{R}(X_i)$	$\hat{X}_i$
0	>30	1	16	16.47	>30	1.50	.25
.5	>30	2	16	15.54	29.54	1.50	.25
1.0	>30	3	16	14.60	28.60	1.50	.25
1.8	28	4	14	13.67	26.67	3.64	1.52
2.2	24	5	13	12.74	22.67	4.71	2.09
2.7	19	6	12	11.80	18.60	5.78	2.59
4.0	17	7.5	11	10.40	15.00	6.85	3.44
4.0	9	7.5	8	10.40	15.00	10.06	5.63
4.9	12	9	9.5	9.00	11.00	8.46	4.58
5.6	12	10	9.5	8.07	9.13	8.46	4.58
6.0	6	11	5	7.13	8.13	13.28	7.50
6.5	8	12	7	6.20	7.20	11.13	6.07
7.3	4	13	1.5	5.27	6.26	17.03	None
8.0	5	14	3	4.33	5.67	15.42	9.01
8.8	6	15	5	3.40	5.20	13.28	7.50
9.3	4	16	1.5	2.46	4.64	17.03	None
9.8	6	17	5	1.53	4.02	13.28	7.50

Figure 5.6: *Calculations for finding the monotonic regression curve estimate.*

- Least squares line on ranks are computed from Equations 2 and 3:

$$y = 17.4037 - 0.9337x$$

- The observations are plotted in Fig. 5.7. The regression curve, consisting of line segments joining successive values of $(X, \hat{Y}_i)$ and $(\hat{X}_i, Y_i)$, is also plotted in Fig. 5.7.

- An estimate $\hat{E}(Y|X = x_0)$ is obtained easily from Fig. 5.7 by finding the ordinate that corresponds to the abscissa x_0.

- It is interesting to note how a set of observations, with a regression curve that is obviously nonlinear, is converted to ranks that have a regression curve that seems to be linear. The ranks are plotted in Fig. 5.8 along with Equation 9.

Theory

- The procedures for monotonic regression are based on the fact that if two variables have a monotonic relationship, their ranks will have a linear relationship.

- A scattering of the observations around the monotonic regression line should correspond to a scattering of the ranks around their linear regression line.

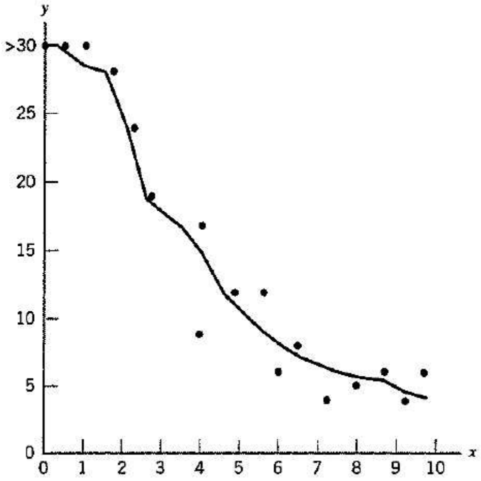


Figure 5.7: *Number of days till fermentation (*y*) versus pounds of sugar (*x*), and the estimated monotonic regression curve.*

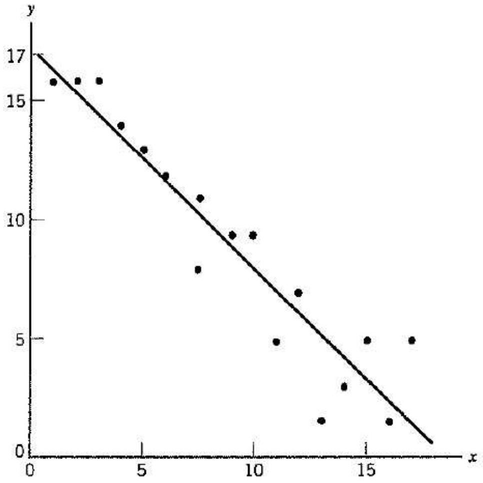


Figure 5.8: $R(Y_i)$ *versus* $R(X_i)$ *and the least squares regression line.*

- The ranks serve as transformed variables, where the transformation seeks to convert the monotonic regression function to a linear regression function.

- Interval estimates of $E(Y|X)$ can be found using the bootstrap method described in Section 2.2. □

5.7 The one-sample or matched-pairs case

- The rank test of this section deals with the single random sample and the random sample of matched pairs that is reduced to a single sample by considering differences.

- A matched pair (X_i, Y_i) is actually a single observation on a bivariate random variable.

- The sign test of Section 3.4 analyzed matched airs of data by reducing each pair to a plus, a minus, or a tie and applying the binomial test to the resultant single sample.

- The test of this section also reduces the matched pair (X_i, Y_i) to a single observation by considering the difference

$$D_i = Y_i - X_i, \quad \text{for } i = 1, 2, \ldots, n \tag{1}$$

- The analysis is then performed on the D_is as a sample of single observations.

- Whereas the sign test merely noted whether D_i was positive, negative, or zero, the test of this section notes the sizes of the positive D_is relative to the negative D_is.

- The model of this section resembles the model used in the sign test. Also, the hypotheses resemble the hypotheses of the sign test.

- The important difference between the sign test and this test is an additional assumption of symmetry of the distribution of differences.

- Before we introduce the test, we should clarify the meaning of the adjective symmetric as it applies to a distribution and discuss the influence of symmetry on the scale of measurement

- Symmetry is easy to define if the distribution is discrete. A discrete distribution is symmetric if the left half of the graph of the probability function is the mirror image of the right half.

- For example, the binomial distribution is symmetric if $p = 1/2$ (see Fig. 5.9) and the discrete uniform distribution is always symmetric (see Fig. 5.10).

- The dotted lines in the figures represent the lines about which the distributions are symmetric.

- For other than discrete distributions we are not able to draw a graph of the probability function. Therefore a more abstract definition of symmetry is required, such as the following.

Definition 5.7.1 *The distribution of a random variable X is symmetric about a line $x = c$, for some constant c, if the probability of $X \leq c - x$ equals the probability of $X \geq c + x$ for each possible value of x.*

- In Fig. 5.9, $c = 2$ and the definition is easily verified for all real numbers x. In Fig. 5.10, $c = 3.5$.
- Even though we may not know the exact distribution of a random variable, we are often able to say, “It is reasonable to assume that the distribution is symmetric.”
- Such an assumption is not as strong as the assumption of a normal distribution; while all normal distributions are symmetric, not all symmetric distributions are normal.

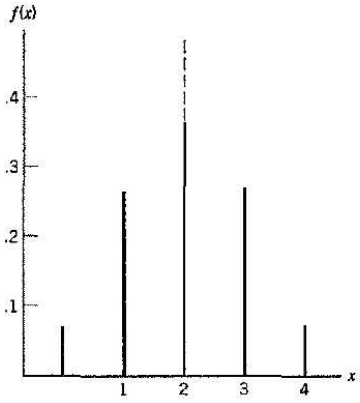


Figure 5.9: *Symmetry in a binomial distribution.*

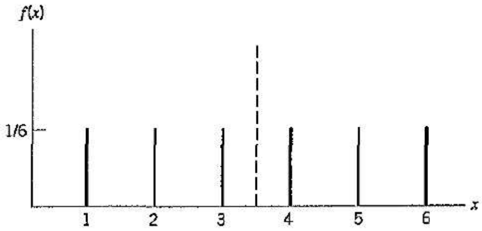


Figure 5.10: *Symmetry in a discrete uniform distribution.*

- If a distribution is symmetric, the mean (if it exists) coincides with the median because both are located exactly in the middle of the distribution, at the line of symmetry.
- One consequence of adding the assumption of symmetry to the model is that any inferences concerning the median are also valid statements for the mean.
- A second consequence of adding the assumption of symmetry to the model is that the required scale of measurement is changed from ordinal to interval.
- With an ordinal scale of measurement, two observations of the random variable need only to be distinguished on the basis of which is larger and which is smaller.

- It is not necessary to know which one is farthest from the median, such as when the two observations are on opposite sides of the median.
- If the assumption of symmetry is a meaningful one, the distance from the median is a meaningful measurement and, therefore, the distance between two observations is a meaningful measurement.
- As a result, the scale of measurement is more than just ordinal, it is interval.
- A test presented by Wilcoxon (1945) is designed to test whether a particular sample came from a population with a specified mean or median.
- It may also be used in situations where observations are paired, such as "before" and "after" observations on each of several subjects, to see if the second random variable in the pair has the same mean as the first.
- Note that in a symmetric distribution the mean equals the median, so the two terms can be used interchangeably.

The Wilcoxon signed ranks test

Data The data consist of n' observations $(x_1, y_1), (x_2, y_2), \ldots, (x_{n'}, y_{n'})$ on the respective bivariate random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n'}, Y_{n'})$. Find the n' differences $D_i = Y_i - X_i$. (In the one-sample problem, the Ds are the observations in the sample, as illustrated in Example 2.) The absolute differences (without regard to sign)

$$|D_i| = |Y_i - X_i|, \quad i = 1, 2, n' \tag{2}$$

are then computed for each of the n' pairs $(X_i - Y_i)$.

Omit from further consideration all pairs with a difference of zero (i.e., where $X_i = Y_i$, or $D_i = 0$). Let the number of pairs remaining be denoted by n, $n \leq n'$. Ranks from 1 to n are assigned to these n pairs according to the relative size of the absolute difference, as follows. The rank 1 is given to the pair (X_i, Y_i) with the smallest absolute difference $|D_i|$; the rank 2 is given to the pair with the second smallest absolute difference; and so on, with the rank n being assigned to the pair with the largest absolute difference.

If several pairs have absolute differences that are equal to each other, assign to each of these several pairs the average of the ranks that would have otherwise been assigned [i.e., if the ranks $3, 4, 5$, and 6 belong to four pairs, but we do not know which rank to assign to which pair because all four absolute differences are exactly equal to each other, assign the average rank $\frac{1}{4}(3 + 4 + 5 + 6) = 4.5$ to each of the four pairs.]

Assumptions

1. The distribution of each D_i is symmetric.
2. The D_is are mutually independent.
3. The D_is all have the same mean.
4. The measurement scale of the D_is is at least interval.

Test statistic

- ▷ Let R_i, called the signed rank, be defined for each pair (X_i, Y_i) as follows.

$$\begin{aligned} R_i = &\ \text{the rank assigned to } (X_i, Y_i) \text{ if } D_i = Y_i - X_i \text{ is positive} \\ &\ (\text{i.e., } Y_i > X_i) \\ R_i = &\ \text{the negative of the rank assigned to } (X_i, Y_i) \text{ if } D_i \text{ is negative} \\ &\ (\text{i.e., } Y_i < X_i) \end{aligned}$$

- ▷ The test statistic is the sum of the positive signed ranks

$$T^+ = \sum (R_i \text{ where } D_i \text{ is positive}) \tag{3}$$

Null distribution

- ▷ Lower quantiles of the exact distribution of T^+ when there are no ties and $n \le 50$ are given in Table A12, under the null hypothesis that the D_is have mean 0.
- ▷ Upper quantiles are found from the relation

$$w_p = \frac{n(n+1)}{2} - w_{1-p} \tag{4}$$

- ▷ If there are many ties, or if $n > 50$, the normal approximation should be used.
- ▷ The normal approximation uses the sum of all of the signed ranks, with their + or − signs, and the statistic

$$T = \frac{\sum_{i=1}^n R_i}{\sqrt{\sum_{i=1}^n R_i^2}} \tag{5}$$

- ▷ In case there are no ties, it simplifies to

$$T = \frac{\sum_{i=1}^n R_i}{\sqrt{n(n+1)(2n+1)/6}} \tag{6}$$

  with the aid of Lemma 1.4.2.
- ▷ The null distribution of T is approximately standard normal, as in Table Al.

Hypotheses

A. *Two-tailed test*:

$$\begin{aligned} &H_0 : E(D) = 0 \qquad (\text{i.e., } E(Y_i) = E(X_i)) \\ &H_1 : E(D) \neq 0 \end{aligned}$$

$$\text{lower-tailed } p\text{-value} = P\left(Z \le \frac{\sum_{i=1}^n R_i + 1}{\sqrt{\sum_{i=1}^n R_i^2}} \right) \tag{7}$$

$$\text{upper-tailed } p\text{-value} = P\left(Z \ge \frac{\sum_{i=1}^n R_i - 1}{\sqrt{\sum_{i=1}^n R_i^2}} \right) \tag{8}$$

B. Lower-tailed test:

$$\begin{aligned} &H_0 : E(D) \geq 0 \qquad (\text{i.e., } E(Y_i) \geq E(X_i)) \\ &H_1 : E(D) < 0 \end{aligned}$$

Reject H_0 at the level of or if T^+ (or T) is less than its a quantile from Table A12 (for T^+) or from Table A1 (for T).
The lower-tailed p-value is given approximately by

$$p\text{-value} = P\left(Z \leq \frac{\sum_{i=1}^n R_i + 1}{\sqrt{\sum_{i=1}^n R_i^2}}\right).$$

C. Upper-tailed test:

$$\begin{aligned} &H_0 : E(D) \leq 0 \qquad (\text{i.e., } E(Y_i) \leq E(X_i)) \\ &H_1 : E(D) > 0 \end{aligned}$$

Reject H_0 at the level of or if T^+ (or T) is greater than its a quantile from Table A12 (for T^+) or from Table A1 (for T).
The upper-tailed p-value is given approximately by

$$p\text{-value} = P\left(Z \geq \frac{\sum_{i=1}^n R_i - 1}{\sqrt{\sum_{i=1}^n R_i^2}}\right).$$

Example 5.7.1 *Twelve sets of identical twins were given psychological tests to measure in some sense the amount of aggressiveness in each person's personality. We are interested in comparing the twins with each other to see if the firstborn twin tends to be more aggressive than the other.*

▶ The results are as follows, where the higher score indicates more aggressiveness.

	Twin set											
	1	2	3	4	5	6	7	8	9	10	11	12
Firstborn X_i	86	71	77	68	91	72	77	91	70	71	88	87
Second twin Y_i	88	77	76	64	96	72	65	90	65	80	81	72
$D_i = Y_i - X_i$	+2	+6	−1	−4	+5	0	−12	−1	−5	+9	−7	−15
Rank of D_i	3	7	1.5	4	5.5	_	10	1.5	5.5	9	8	11
R_i	3	7	−1.5	−4	5.5	_	−10	−1.5	−5.5	9	−8	−11

▶ Lower-tailed test:

H_0: The firstborn twin does not tend to be more aggressive than the other $(E(X_i) \leq E(Y_i))$

H_1: The firstborn twin tends to be more aggressive than the other $(E(X_i) > E(Y_i))$

▶

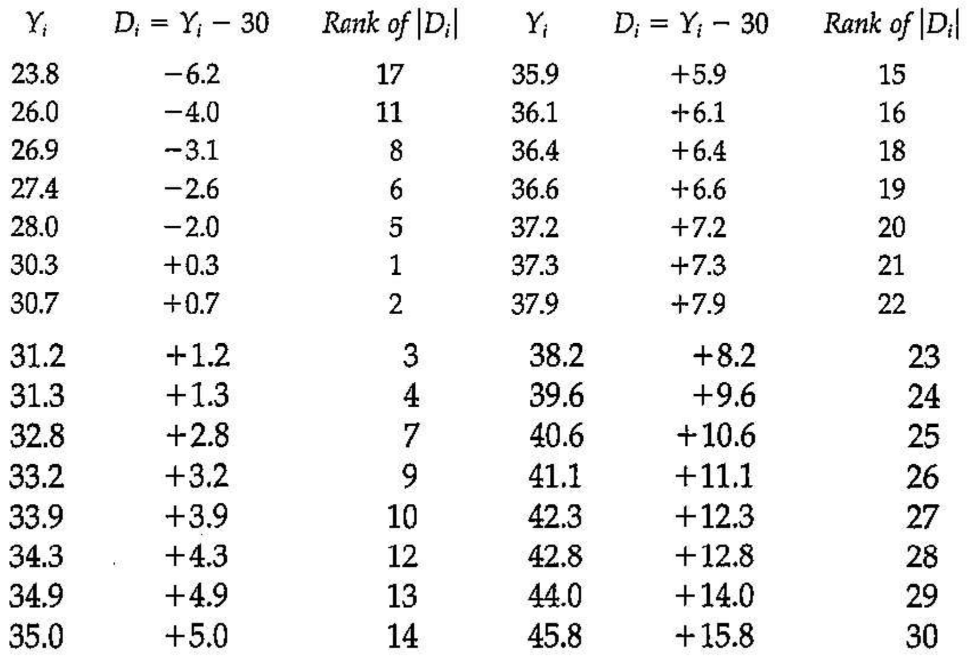
$$T = \frac{\sum R_i}{\sqrt{\sum R_i^2}} = \frac{-17}{\sqrt{505}} = -0.7565$$

- The critical region of size $\alpha = 0.05$ corresponds to values of T less than -1.6449 (from Table A1). Therefore H_0 is readily accepted.
- The p-value, from Equation 7, is 0.238.
- If we had used T^+ and Table A12 we would have obtained $T^+ = 24.5$ and a critical region corresponding to values of T^+ less than 14.
- So the same conclusion would have been reached and a similar p-value would have been obtained by interpolation between $W_{0.20}$ and $W_{0.30}$ in Table A12.

Example 5.7.2 *Thirty observations on a random variable Y are obtained in order to test the hypothesis that $E(Y)$ is no longer than 30 (hypotheses set C).*

$$H_0 : E(Y) \leq 30 \quad v.s. \quad H_1 : E(Y) > 30$$

- The observations, the difference $Y_i - m$, and the ranks of the pairs are as follows.



Y_i	$D_i = Y_i - 30$	Rank of $\lvert D_i \rvert$	Y_i	$D_i = Y_i - 30$	Rank of $\lvert D_i \rvert$
23.8	−6.2	17	35.9	+5.9	15
26.0	−4.0	11	36.1	+6.1	16
26.9	−3.1	8	36.4	+6.4	18
27.4	−2.6	6	36.6	+6.6	19
28.0	−2.0	5	37.2	+7.2	20
30.3	+0.3	1	37.3	+7.3	21
30.7	+0.7	2	37.9	+7.9	22
31.2	+1.2	3	38.2	+8.2	23
31.3	+1.3	4	39.6	+9.6	24
32.8	+2.8	7	40.6	+10.6	25
33.2	+3.2	9	41.1	+11.1	26
33.9	+3.9	10	42.3	+12.3	27
34.3	+4.3	12	42.8	+12.8	28
34.9	+4.9	13	44.0	+14.0	29
35.0	+5.0	14	45.8	+15.8	30

- The 0.05 quantile from Table A12 is 152 so the 0.95 quantile is $465 - 152 = 313$. Therefore the critical region of size 50.05 corresponds to values of the test statistic greater than 313.
- The test statistic is defined by Equation 3. In this case T^+ equals the sum of the ranks associated with the positive D_i.

$$T^+ = 418$$

The large value of T^+ results in rejection of H_0. We conclude that the mean of Y is greater than 30.

- The approximate p-value is given by Equation 8.

$$P\left(Z \geq \frac{\sum R_i - 1}{\sqrt{n(n+1)(2n+1)/6}}\right) = P\left(Z \geq \frac{371 - 1}{\sqrt{9455}}\right)$$
$$= P(Z \geq 3.8051)$$

- Table A1 shows that the p-value is smaller than 0.0001.

Theory

- The model states that all of the differences D_i share a common median, say $d_{0.50}$, which equals zero when H_0 is true.
- By the definition of symmetry, the probability of each D_i being negative is the same as its probability of being positive, which equals 0.5 for continuous distributions, or for discrete distributions where values of D_i equal to zero are discarded.
- The purpose of these considerations is to find the distribution of the test statistic T^+ when H_0 is true.
- First, we will consider the null hypothesis of the two-tailed test. The resulting distribution applies equally well in the one-tailed tests.
- Consider n chips numbered from 1 to n, corresponding to the n ranks if there are no ties.
- Suppose each chip has its number written on one side and the negative of its number on the other side (like 6 and -6).
- Each chip is tossed into the air so that it is equally likely to land with either side showing, corresponding to the ranks of (X_i, Y_i), which are equally likely to correspond to a positive D_i, in which case the signed rank R_i equals the rank, or a negative D_i, in which case R_i is a negative rank.
- Let T^+ be the sum of the positive numbers showing after all n chips are tossed, corresponding to the definition of T^+ in Equation 3.
- The probability distribution of T^+ is the same in the game with the chips as it is when H_0 is true, but the game with the chips is easier to imagine.
- The sample space in the game with the chips consists of points such as $(1, 2, 3, -4, -5, 6, 7, \ldots, n)$, simply a reordering of the R_i associated with a set of data like in Example 1.
- The tosses are independent of each other, so each of the 2^n points has probability $(1/2)^n$.
- The test statistic T^+ equals the sum of the positive numbers in the sample point.
- Therefore the probability that T^+ equals any number x is found by counting the points whose positive numbers add to x, then multiplying that count by the probability $(1/2)^n$.
- For example, if $n = 8$, then T^+ can equal 0 one way (all the positive numbers landed face down), and so $P(T = 0) = (1/2)^8$.
- $T^+ = 1$ only one way, $T^+ = 2$ only one way, but $T^+ = 3$ two ways, points $(-1, -2, 3, -4, -5, -6, -7, -8)$ and $(1, 2, -3, -4, -5, -6, -7, -8)$. Also, $T^+ = 4$ two ways. That is,

$$\begin{array}{ll} P(T^+ = 0) = (1/2)^8 = 1/256 & P(T^+ \leq 0) = 0.0039 \\ P(T^+ = 1) = 1/256 & P(T^+ \leq 1) = 0.0078 \\ P(T^+ = 2) = 1/256 & P(T^+ \leq 2) = 0.0117 \end{array}$$

$P(T^+ = 3) = 2/256$	$P(T^+ \leq 3) = 0.0195$
$P(T^+ = 4) = 2/256$	$P(T^+ \leq 4) = 0.0273$
etc.	etc.

- The distribution function of T^+ is tabulated in Owen (1962) for $n \leq 20$ and in Harter and Owen (1970) for $n \leq 50$.

- A table of selected quantiles for $n \leq 100$ is given by McCornack (1965). That table is more extensive than we need here, so the more useful quantiles were selected and are given in Table A12.

- The use of Table A12 will generally result in a slightly conservative test, because the probability of being less than the p quantile may be less than p.

- For example, if $n = 8$, as in the preceding paragraph, the 0.025 quantile of T^+ is given in Table A12 as 4, while the actual size of the critical region corresponding to values of T^+ less than 4 is 0.0195.

- Further results on the exact distribution of T^+ are given by Claypool (1970) and Chow and Hodges (1975).

- For the one-tailed tests, the probability of getting a point in the critical region is a maximum when the median difference is 0, so this is the situation to be considered. Thus the preceding distribution of T^+ is equally valid when H_0 is true in the one-tailed tests.

- To find the conditional distribution of T^+ when there are ties, only the initial step in the discussion is changed. That is, the numbers on the chips must agree with the ranks and average ranks assigned to the pairs (X_i, Y_i) in the particular set of data under consideration. Call these ranks and average ranks $a_1, a_2, \ldots, a_n$.

- In Example 1 we have $a_1 = 1.5$, $a_2 = 1.5$, $a_3 = 3$, and so on. For this set of numbers we can find the distribution T^+. Because there are 11 numbers in Example 1, there are $2^{11} = 2048$ points in the sample space. The smallest 5% of these, about 102 points, constitute the critical region. This is a large number of points to tabulate by hand, so the normal approximation is used.

- To use the normal approximation, let S equal the sum of all the R_i. Then, to apply the central limit theorem from Section 1.5, we need the mean and variance of S when H_0 is true.

- Note that under H_0,

  $$P(R_i = a_i) = 1/2 \quad \text{and} \quad P(R_i = -a_i) = 1/2$$

  so that

  $$E(R_i) = a_i(\frac{1}{2}) + (-a_i)(\frac{1}{2}) = a_i^2.$$

- Since the R_is are independent of each other (the tosses of the chips are independent), we can apply Theorems 1.4.1 and 1.4.3 to get

  $$E(S) = \sum_{i=1}^{n} E(R_i) = 0$$

and

$$Var(S) = \sum_{i=1}^{n} Var(R_i) = \sum_{i=1}^{n} a_i^2.$$

- ▶ But since a_i^2 always equals R_i^2 (the sign always becomes +), we can say

$$Var(S) = \sum_{i=1}^{n} R_i^2$$

and apply the central limit theorem to

$$T = \frac{\sum_{i=1}^{n} R_i}{\sqrt{\sum_{i=1}^{n} R_i^2}}$$

and use the normal distribution with a continuity correction as an approximation whenever exact tables are not available.

- ▶ Justification for the treatment of ties is given by Vorlickova (1970) and Conover (1973a). □

Confidence interval for the median difference

Data The data consist of n observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ on the bivariate random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, respectively. Compute the difference

$$D_i = y_i - X_i$$

for each pair and arrange them in order from the smallest (the most negative) to the largest (the most positive), denoted as follows.

$$D^{(1)} \leq D^{(2)} \leq \cdots \leq D^{(n-1)} \leq D^{(n)}$$

Assumptions

1. The distribution of each D_i is symmetric.
2. The D_is are mutually independent.
3. The D_is all have the same median.
4. The measurement scale of the D_is is at least interval.

Method

- ▷ To obtain a $1 - \alpha$ confidence interval, obtain the $\alpha/2$ quantile $w_{\alpha/2}$ from Table A12.
- ▷ Then consider the $n(n+1)/2$ possible averages $(D_i + D_j)/2$ for all i and j, including $i = j$, which is the average of D_i with itself, giving just D_i.
- ▷ The $w_{\alpha/2}$th largest of these averages and the $w_{\alpha/2}$th smallest of these averages constitute the upper and lower bounds for the $1 - \alpha$ confidence interval.
- ▷ It is not necessary to compute all $n(n+1)/2$ averages; only the averages near the largest and the smallest need to be computed to obtain a confidence interval.

Example 5.7.3 (Example 1 (continued)) *The 12 values of D_i, arranged in order, are*

$$-15, -12, -7, -5, -4, -1, -1, 0, 2, 5, 6, 9$$

Find a 95% CI for the median difference

$$P(-6.5 \leq d_{0.50} \leq 2.5) \geq 0.95$$

$$H_0 : d_{0.50} = m$$

$$P(L \leq d_{0.50} \leq U) \geq 1 - \alpha$$

5.8 Several related samples

- In Section 5.2 we presented the Kruskal-Wallis rank test for several independent samples, which is an extension of the Mann-Whitney test for two independent samples introduced in Section 5.1.

- In this section we consider the problem of analyzing several related samples, whichis an extension of the problem of matched pairs, or two related samples, examined in the previous section.

- First we will present the Friedman test, which is an extension of the sign test of Sections 3.4 and 3.5. Then we will present the Quade test, whichis an extensionof the Wilcoxon signed ranks test of the previous section.

- The Friedman test is the better-known test of the two and requires fewer assumptions, but it suffers from a lack of power when there are only three treatments, just as the sign test has less power than the Wilcoxon signed ranks test when there are only two treatments.

- When there are four or five treatments the Friedman test has about the same power as the Quade test, but when the number of treatments is six or more the Friedman test tends to have more power.

- See Iman et al. (1984) and Hora and Iman (1988) for power and A.R.E. comparisons.

- The problem of several related samples arises in an experiment that is designed to detect differences in k possibly different treatments, $k \geq 2$.

- The observations are arranged in blocks, which are groups of k experimental units similar to each other in some important respects, such as k puppies that are littermates and therefore may tend to respond to a particular stimulus more similarly than would randomly selected puppies from various litters.

- The k experimental units within a block are matched randomly with the k treatments being scrutinized, so that each treatment is administered once and only once within each block.

- In this way the treatments may be compared with each other without an excess of unwanted effects confusing the results of the experiment.

- The total number of blocks used is denoted by b, $b > 1$.

- The experimental arrangement described here is usually called a randomized complete block design.

- This design may be compared with the incomplete block design described in the next section, in which the blocks do not contain enough experimental units to enable all the treatments to be applied in all the blocks, and so each treatment appears in some blocks but not in others.

- Examples of randomized complete block designs are as follows.

  1. *Psychology.* Five litters of mice, with four mice per litter, are used to examine the relationship between environment and aggression. Each litter is considered to be a block. Four different environments are desi-m ed. One mouse from each litter is placed in each environment, so that the four mice from each litter are in four different environments. After a suitable length of time, the mice are regrouped with their littermates and are ranked a cordiigto degree of aggressiveness.
  2. *Home economics.* Six different types of bread dough are compared to see which bakes the fastest by forming three loaves with each type of dough. Three different ovens are used, and each oven bakes the six different types of bread at the same time. The ovens are the blocks and the doughs are the treatments.
  3. *Environmental engineering.* One experimental unit may form a block if the different treatments may be applied to the same unit without leaving residual effects. Seven different men are used in a study of the effect of color schemes on work efficiency. Each man is considered to be a block and spends some time in each of three rooms, each with its own type of color scheme. While in the room, each man performs a work task and is measured for work efficiency. The three rooms are the treatments.

- By now the reader should have some idea of the nature of a randomized complete block design.

- The usual parametric method of testing the null hypothesis of no treatment differences is called the two-way analysis of variance.

- The following nonparametric method depends only on the ranks of the observations within each block.

- Therefore it may be considered a two-way analysis of variance on ranks.

- This test is named after its inventor, the noted economist Milton Friedman.

The Friedman Test

Data The data consist of b mutually independent k-variate random variables $(X_{i1}, X_{i2}, \ldots, X_{ik})$, called b blocks, $i = 1, 2, \ldots, b$. The random variable X_{ij} is in block i and is associated with treatment j. The b blocks are arranged as follows.

	Treatment			
Block	1	2	...	k
1	X_{11}	X_{12}	...	X_{1k}
2	X_{21}	X_{22}	...	X_{2k}
3	X_{31}	X_{32}	...	X_{3k}
...	...	...	...	...
b	X_{b1}	X_{b2}	...	X_{bk}

Let $R(X_{ij})$ be the rank, from 1 to k, assigned to X_{ij} within block (row) i. That is, for block i the random variables $X_{i1}, X_{i2}, \ldots, X_{ik}$ are compared with each other and the rank 1 is assigned to the smallest observed value, the rank 2 to the second smallest, and so on to the rank k, which is assigned to the largest observation in block i. Ranks are assigned in all of the b blocks. Use average ranks in case of ties.

Then sum the ranks for each treatment to obtain R_j where:

$$R_j = \sum_{i=1}^{b} R(X_{ij}) \tag{5.8.1}$$

for $j = 1, 2, \ldots, k$.

Assumptions

1. The b k-variate random variables are mutually independent. (The results within one block do not influence the results within the other blocks.)
2. Within each block the observations may be ranked according to some criterion of interest.

Test statistic

▷ Friedman suggested using the statistic

$$T_1 = \frac{12}{bk(k+1) \sum_{j=1}^{k} \left(R_j - \frac{b(k+1)}{2} \right)^2} \tag{5.8.2}$$

▷ If there are ties present an adjustment needs to be made. Let A_1 be the sum of the squares of the ranks and average ranks.

$$A_1 = \sum_{i=1}^{b} \sum_{j=1}^{k} [R(X_{ij})]^2 \tag{5.8.3}$$

▷ Also compute the "correction factor" C_1 given by

$$C_1 = bk(k+1)^2/4 \tag{5.8.4}$$

▷ Then the statistic T_1, adjusted for the presence of ties, becomes

$$T_1 = \frac{(k-1)\left[\sum_{j=1}^{k} R_j^2 - bC_1\right]}{A_1 - C_1} = \frac{(k-1)\sum_{j=1}^{k}\left(R_j - \frac{b(k+1)}{2}\right)^2}{A_1 - C_1} \tag{5.8.5}$$

▷ Current research indicates the preferred statistic, because of its more accurate approximate distribution, is the two-way analysis of variance statistic computed on the ranks $R(X_{ij})$, which simplifies to the following function of T_1 given above.

$$T_2 = \frac{(b-1)T_1}{b(k-1) - T_1} \tag{5.8.6}$$

▷ See Iman and Davenport (1980) for more details on the closeness of these approximations.

Null distribution

- The exact distribution of T_1 (or T_2) is difficult to find and so an approximation is usually used.
- The approximate distribution of T_1 is the chi-squared distribution with $k-1$ degrees of freedom.
- However, this approximation is sometimes rather poor, so it is recommended to use T_2 instead of T_1, which has the approximate quantiles given by the F distribution (Table A22) with $k_1 = k-1$ and $k_2 = (b-1)(k-1)$, when the null hypothesis is true.

Hypotheses

H_0 : Each ranking of the random variables within a block is equally likely (i.e., the treatments have identical effects)

H_1 : At least one of the treatments tends to yield larger observed values than at least one other treatment

- Reject H_0 at the approximate level a if T_2 exceeds the $1-\alpha$ quantile of the F distribution given by Table A22 for $k_1 = k-1$ and $k_2 = (b-l)(k-1)$.
- This approximation is fairly good and improves as b gets larger. The approximate p-value may be estimated from Table A22.

Multiple Comparisons

- The following method for comparing individual treatments may be used only if the Friedman test results in rejection of the null hypothesis.
- Treatments i and j are considered different if the following inequality is satisfied.

$$|R_j - R_i| > t_{1-\alpha/2}\left[\frac{2(bA_1 - \sum_{j=1}^{k} R_j^2)}{(b-1)(k-1)}\right]^{\frac{1}{2}} \tag{5.8.7}$$

where R_i, R_j and A_1 are given previously and where $t_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the f distribution given by Table A21 with $(b-1)(k-1)$ degrees of freedom.

- The value for a is the same one used in the Friedman test.
- Alternatively, Equation 7 can be expressed as a function of T_1

$$|R_j - R_i| > t_{1-\alpha/2}\left[\frac{(A_1 - C_1)2b}{(b-1)(k-1)}\left(1 - \frac{T_1}{b(k-1)}\right)\right]^{\frac{1}{2}} \tag{5.8.8}$$

- If there are no ties A, in Equation 7 simplifies to

$$A_1 = bk(k+1)(2k+1)/6$$

and $(A_1 - C_1)$ in Equation 8 simplifies to

$$A_1 - C_1 = bk(k+1)(k-1)/12$$

Computer Assistance The Friedman test appears in *Minitab*, *S-Plus*, *SAS*, and *StatXact*.

Example 5.8.1 *Twelve homeowners are randomly selected to participate in an experiment with a plant nursery. Each homeowner was asked to select four fairly identical areas in his yard and to plant four different types of grasses, one in each area. At the end of a specified length of time each homeowner was asked to rank the grass types in order of preference, weighing important criteria such expense, maintenance and upkeep required, beauty, hardiness, wife's preference, and so on.*

- The rank 1 was assigned to the last preferred grass and the rank 4 to the favorite.
- The null hypothesis was that there is no difference in preferences of the grass types, and the alternative was that some grass types tend to be preferred over others.
- Each of the 12 blocks consists of four fairly identical plots of land, each receiving care of approximately the same degree of skill because the four plots are presumably cared for by the same homeowner.
- The results of the experiment are as follows.

	Grass			
Homeowner	1	2	3	4
1	4	3	2	1
2	4	2	3	1
3	3	1.5	1.5	4
4	3	1	2	4
5	4	2	1	3
6	2	2	2	4
7	1	3	2	4
8	2	4	1	3
9	3.5	1	2	3.5
10	4	1	3	2
11	4	2	3	1
12	3.5	1	2	3.5
R_j	38	23.5	24.5	34

- $A_1 = 356.5$
- $C_1 = \frac{12(4)(25)}{4} = 300$ by Eq. 4.
- By Eq. 5:

$$\begin{aligned} T_1 &= \frac{3[(38)^2 + (23.5)^2 + (24.5)^2 + (34)^2 - 12(300)]}{356.5 - 300} \\ &= 8.097 \end{aligned}$$

- $T_2 = \frac{11(8.097)}{12(3)-8.097} = 3.19$ by Eq. 6.

The Quade Test

Data

- ▷ Find the ranks within blocks $R(X_{ij})$ as described in the previous test. The next step again uses the original observations X_{ij}.
- ▷ Ranks are assigned to the blocks themselves according to the size of the sample range in each block.
- ▷ The sample range within block i is the difference between the largest and the smallest observations within that block.

$$\text{Range in block}i = \text{maximum}\{X_{ij}\} - \text{minimum}\{X_{ij}\} \tag{5.8.9}$$

- ▷ There are b sample ranges, one for each block. Assign rank 1 to the block with the smallest range, rank 2 to the second smallest, and so on to the block with the largest range, which gets rank b.
- ▷ Use average ranks in case of ties. Let $Q_1, Q_2, \ldots, Q_b$ be the ranks assigned to blocks $1, 2, \ldots, b$, respectively.
- ▷ Finally, the block rank Q_i is multiplied by the difference between the rank within block i, $R(X_{ij})$, and the average rank within blocks, $(k+1)/2$ to get the product S_{ij}, where

$$S_{ij} = Q_i \left[R(X_{ij} - \frac{k+1}{2}) \right] \tag{5.8.10}$$

  is a statistic that represents the relative size of each observation within the block, adjusted to reflect the relative significance of the block in which it appears.
- ▷ Let S_j denote the sum for each treatment:

$$S_j = \sum_{i=1}^{b} S_{ij} \tag{5.8.11}$$

  for $j = 1, 2, \ldots, k$.

Assumptions The first two assumptions are the same as the two assumptions of the previous test. A third assumption is needed because comparisons are made between blocks.

3. The sample range may be determined within each block so that the blocks may be ranked.

Test statistic

- ▷ First calculate the term

$$A_2 = \sum_{i=1}^{b} \sum_{j=1}^{k} S_{ij}^2$$

  where S_{ij} is given by (5.8.10). This is called the "total sum of squares."
- ▷ If there are no ties, A_2 simplifies to

$$A_2 = b(b+1)(2b+1)k(k+1)(k-1)/72$$

  where S_{ij} is given by (5.8.11). This is called the "treatment sum of squares."
- ▷ The test statistic is

$$T_3 = \frac{(b-1)B}{A_2 - B}$$

If $A_2 = B$, consider the point to be in the critical region and calculate the p-value as $(l/k!)^{b-l}$.

- Note that T_3 is the two-way analysis of variance test statistic computed on the scores S_{ij} given by (5.8.10).

Null distribution

- The exact distribution of T_3 is difficult to find, so the F distribution, whose quantiles are given in Table A22, is used as an approximation, with $k_1 = k - 1$ and $k_2 = (b-1)(k-1)$ as before in the Friedman test.

Hypotheses The hypotheses are the same as in the Friedman test.

- Reject the null hypothesis at the level α if T_3 exceeds the $1 - \alpha$ quantile of the F distribution as given in Table A22 with $k_1 = k - 1$ and $k_2 = (b-1)(k-1)$.
- Actually, the F distribution only approximates the exact distribution of T_3, but exact tables are not available at this time.
- As b becomes large, the F approximation comes closer to being exact.

Multiple comparisons

- Only if the preceding procedure results in rejection of the null hypothesis are multiple comparisons made.
- Treatments i and j are considered different if the inequality

$$|S_i - S_j| > t_{1-\alpha/2}\left[\frac{2b(A_2 - B)}{(b-1)(k-1)}\right]^{1/2}$$

  is satisfied, where S_i, S_j, A_2, and B are given previously, and where $t_{1-\alpha/2}$ is obtained from Table A21 with $(b-1)(k-1)$ degrees of freedom.
- This comparison is made for all pairs of treatments, using the same a used in the Quade test.

Example 5.8.2 *Seven stores are selected for a marketing survey. In each store five different brands of a new type of hand lotion are placed side by side. At the end of the week, the number of bottles of lotion sold for each brand is tabulated, with the following results.*

Numbers of customers

Store	*A*	*B*	*C*	*D*	*E*
1	*5(2)*	*4(1)*	*7(3)*	*10(4)*	*12(5)*
2	*1(2.5)*	*3(5)*	*1(2.5)*	*0(1)*	*2(4)*
3	*16(2)*	*12(1)*	*22(3.5)*	*22(3.5)*	*35(5)*
4	*5(4.5)*	*4(2.5)*	*3(1)*	*5(4.5)*	*4(2.5)*
5	*10(3.5)*	*9(2)*	*7(1)*	*13(5)*	*10(3.5)*
6	*19(2)*	*18(1)*	*28(3)*	*37(4)*	*58(5)*
7	*10(5)*	*7(2.5)*	*6(1)*	*8(4)*	*7(2.5)*

- The observations are ranked from 1 to 5 within each store, with average ranks assigned when there are ties. These ranks $R(X_{ij})$ appear in parentheses.

- Next, the sample range within each store is computed by subtracting the smallest observation from the largest. In store 1 the sample range is 12-4 = 8.

- These sample ranges are listed next, along with the ranks Q_i of the sample ranges, and the products

$$S_{ij} = Q_i[R(X_{ij}) - (k+1)/2]$$

			$S_{ij} = Q_i[R(X_{ij}) - 3]$				
Store	Sample	Rank	Brand				
Number	Range	Q_i	A	B	C	D	E
1	8	5	-5	-10	0	$+5$	$+10$
2	3	2	-1	$+4$	-1	-4	$+2$
3	23	6	-6	-12	$+3$	$+3$	$+12$
4	2	1	$+1.5$	-0.5	-2	$+1.5$	-0.5
5	6	4	$+2$	-4	-8	$+8$	$+2$
6	40	7	-7	-14	0	$+7$	$+14$
7	4	3	$+6$	-1.5	-6	$+3$	-1.5
		$S_j =$	-9.5	-38	-14	$+23.5$	$+38$

- From Equation 12,

$$A_2 = \sum_{i=1}^{7}\sum_{j=1}^{5} S_{ij}^2 = (-5)^2 + (-10)^2 + \cdots = 1366.5$$

  which is slightly less than the more easily obtained value 1400.

- Equation 14 yields

$$B = \frac{1}{7}\sum_{j=1}^{5} S_j^2 = \frac{1}{7}[(-9.5)^2 + (-38)^2 + \cdots] = 532.4$$

  which gives, when substituted into Equation 15, the test statistic

$$T_3 = \frac{6(532.4)}{1366.5 - 532.4} = 3.83$$

- This value of T_3 is greater than 2.78, the 0.95 quantile of the F distribution with $k_1 = 4$ and $k_2 = 24$, obtained from Table A22; therefore the null hypothesis is rejected at $\alpha = 0.05$.

- In fact, perusal of Table A22 shows the p-value to be slightly less than 0.025.

- Some brands seem to be preferred over others by the store customers.

- Because the null hypothesis is rejected, multiple comparisons are made.

- From Equation 16 two treatments are considered different if the difference between their sums $|S_i - S_j|$ exceeds

$$t_{1-\alpha/2}\left[\frac{2b(A_2 - B)}{(b-1)(k-1)}\right]^{1/2} = 2.064\left[\frac{14(834.1)}{24}\right]^{1/2} = 45.53$$

where $t_{1-\alpha/2} = t_{0.975}$ is obtained from Table A21 for $(b-l)(k-1) = 24$ degrees of freedom.

- Thus the brands that may be considered different from each other are brands A and E, brands B and D, brands B and E, and brands C and E.

- Note that a summary of the multiple comparisons procedure may be presented by listing the treatments in order of increasing average scores, and underlining the groups of treatments that are not significantly different with a single underline, as follows.

$$\begin{array}{ccccc} \text{B} & \text{C} & \text{A} & \text{D} & \text{E} \\ \cline{1-3} \cline{4-5} & & & & \\ \cline{2-4} \end{array}$$

Theory

- The exact distribution of T_1, T_2 and T_3 is found under the assumption that each ranking within a block is equally likely, which is the null hypothesis.

- There are $k!$ possible arrangements of ranks $R(X_{ij})$ within a block and, therefore, $(k!)^b$ possible arrangements of ranks in the entire array of b blocks.

- The preceding statements imply that each of these $(k!)^b$ arrangements is equally likely under the null hypothesis.

- Therefore the probability distributions of T_1, T_2 and T_3 may be found for a given number of samples k and blocks b, merely by listing all possible arrangements of ranks and by computing T_1, T_2, or T_3 for each arrangement.

- For example, if $k = 2$ and $b = 3$, there are $(2!)^3 = 8$ equally likely arrangements of the ranks, which are listed next along with their associated values of T_1, and T_2. We will consider T_3 later.

	Arrangements							
Blocks	1	2	3	4	5	6	7	8
1	1, 2	1, 2	1, 2	2, 1	2, 1	2, 1	1, 2	2, 1
2	1, 2	1, 2	2, 1	1, 2	2, 1	1, 2	2, 1	2, 1
3	1, 2	2, 1	1, 2	1, 2	1, 2	2, 1	2, 1	2, 1
Probability	1/8	1/8	1/8	1/8	1/8	1/8	1/8	1/8
Value of T_2	∞	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
Value of T_1	3	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$	3

- Therefore the probability distribution of T_1 is given by $P(T_1 = \frac{1}{3}) = 3/4$ and $P(T_1 = 3) = 1/4$ under H_0. The probability distribution of T_2 is given by $P(T_2 = \frac{1}{4}) = 3/4$ and $P(T_2 = \infty) = 1/4$.

- To examine the behavior of T_3 under the null hypothesis we again start out with the eight equally likely arrangements of ranks $R(X_{ij})$, as just given.

- The average rank 1.5 is subtracted from each rank and, for the moment, we consider the case where the block ranks are given by $Q_l = 1$, $Q_2 = 2$, $Q_3 = 3$. The resulting arrays of S_{ij} are given here.

	Arrangements			
Blocks	1	2	3	4

1	$-0.5, +0.5$	$-0.5, +0.5$	$-0.5, +0.5$	$+0.5, -0.5$
2	$-1, +1$	$-1, +1$	$+1, -1$	$-1, +1$
3	$-1.5, +1.5$	$+1.5, -1.5$	$-1.5, +1.5$	$-1.5, +1.5$
Conditional Probability	1/8	1/8	1/8	1/8
Value of T_3	12	0	$\frac{4}{19}$	$1\frac{3}{13}$

	Arrangements			
Blocks	5	6	7	8
1	$+0.5, -0.5$	$+0.5, -0.5$	$-0.5, +0.5$	$+0.5, -0.5$
2	$+1, -1$	$-1, +1$	$+1, -1$	$+1, -1$
3	$-1.5, +1.5$	$+1.5, -1.5$	$+1.5, -1.5$	$+1.5, -1.5$
Conditional Probability	1/8	1/8	1/8	1/8
Value of T_3	0	$\frac{4}{19}$	$1\frac{3}{13}$	12

- The probability for each value of T_3 is:

$$\frac{1}{8} \cdot P(Q_1 = 1, Q_2 = 2, Q_3 = 3)$$

  because 1/8 represents the conditional probability for that value of T_3, given the assignment of ranks Q_1, Q_2, and Q_3.

- Suppose a different assignment of ranks Q_1, Q_2, Q_3 is considered, say $Q_1 = 2$, $Q_2 = 1$, $Q_3 = 3$. Then the reader may easily verify, by listing the eight arrangements of values of S_{ij} as we just did, that again we observe the same eight values of T_3, and each of these eight values has probability

$$\frac{1}{8} \cdot P(Q_1 = 2, Q_2 = 1, Q_3 = 3)$$

- By considering all six (3!) permutations of ranks for Q_1, Q_2, and Q_3, we arrive at the total probability for each value of T_3: 1/8. Thus, for purposes of calculating the null distribution of T_3, only the one case given here, $Q_i = i$, for $i = 1, 2, 3$, must be considered.

- The probability distribution of T_3 is obtained by collecting identical values of T_3 to get

$$P(T_3 = 0) = 1/4 \quad P(T_3 = \frac{4}{19}) = 1/4$$

$$P(T_3 = 1\frac{3}{13}) = 1/4 \quad P(T_3 = 12) = 1/4$$

- The approximation of the distributions of T_1, T_2 and T_3 that use the F or chi-squared distributions are justified using the central limit theorem.

- Some of the details are beyond the scope of this book, so the entire development of the asymptotic distributions is omitted.

- The reader is referred to Quade (1972,1979) or Lawler (1978) for T_3, Iman and Davenport (1979) for T_2, and Friedman (1937) for T_1. □

The Page Test for Ordered Alternatives

- In Section 5.4 we presented the Jonckheere-Terpstra test of k independent samples when the alternative of interest specifies an ordering of the treatment effects.

- It is equivalent to computing Kendall's τ between the observations and the ordering of the treatments specified in the alternative hypothesis.

- We mentioned that Spearman's ρ could have been used just as well.

- In the randomized complete block design, Spearman's ρ is used to test for k related samples against the alternative hypothesis of a specified ordering of the treatment effects.

- The correlation between the Friedman within-block rankings and the ordering of the treatments as specified by H_1 is used in a test introduced by Page (1963).

- Because of the many ties inherent in the data, Page uses a simpler statistic, which is a monotonic function of Spearman's ρ if there are no ties within blocks, namely

  $$T_4 = \sum_{j=1}^{k} jR_j = R_1 + 2R_2 + \cdots + kR_K$$

  where R_j is the treatment rank sum in the Friedman test, arranged in increasing order of the treatment effects as specified by H_1.

- Although exact tables are given by Page (1963), we will consider only the large sample approximation, and reject H_0 when

  $$T_5 = \frac{T_4 - bk(k+1)^2/4}{[b(k^3-k)^2/144(k-1)]^{1/2}}$$

  exceeds the $1-\alpha$ quantile from a standard normal distribution, as given in Table Al, for an upper-tailed test of size α.

- *StatXact* finds exact p-values for Page's test.

Example 5.8.3 *Health researchers suspect that regular exercise has a tendency to lower the pulse rate of a resting individual. To test this theory eight healthy volunteers, who did not exercise on a regular basis, were enrolled in a supervised exercise program. their resting pulse rate was measured at the beginning of the program, and again after each month for four months.*

$$\begin{aligned} H_0 &: \quad \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 \\ H_1 &: \quad \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4 \leq \mu_5 \end{aligned}$$

where μ_1 is the mean at the end of the fourth month, μ_5 is the initial mean, and there is at least one strict inequality in H_1.

- The observed pulse rates are as follows, along with their Friedman within- blocks ranks.

Person	Initial	Month 1	Month 2	Month 3	Month 4
1	82(4)	84(5)	77(2)	76(1)	79(3)
2	80(4.5)	80(4.5)	76(1.5)	76(1.5)	78(3)
3	75(3)	78(5)	77(4)	74(2)	72(1)
4	65(1.5)	72(5)	68(4)	65(1.5)	66(3)
5	77(5)	74(2)	72(1)	75(3.5)	75(3.5)
6	68(4)	69(5)	65(2)	66(3)	64(1)
7	70(3.5)	74(5)	68(1.5)	70(3.5)	68(1.5)
8	77(4)	76(3)	78(5)	72(2)	70(1)
	$R_5 = 29.5$	$R_4 = 34.5$	$R_3 = 21$	$R_2 = 18$	$R_1 = 17$

- Notice that R_1 is the rank sum predicted by H_1 to be the smallest, R_2 is predicted to be the second smallest, and so on.

$$T_4 = 17 + 2(18) + 3(21) + 4(34.5) + 5(29.5) = 401.5$$

$$T_5 = \frac{401.5 - 8(5)(36)/4}{[8(5^3 - 5)^2/144(4)}]^{1/2} = \frac{41.5}{\sqrt{200}} = 2.9345$$

- A comparison of T_5 with Table A1 shows $p = 0.002$ and H_0 is easily rejected at $\alpha = 0.05$. Page's tables, exact only if there are no ties, show the same p-value.

5.9 The balanced incomplete block design

The Durbin Test

Data We will use the following notation.

t =the number of treatments to be examined.
k =the number of experimental units per block ($k < t$).
b =the total number of blocks.
r =the number of times each treatment appears ($r < b$).
λ =the number of blocks in which the ith treatment and the jth treatment appear together.
(λ is the same for all pairs of treatments.)

- The data are arrayed in a balanced incomplete block design, just defined. Let X_{ij} represent the result of treatment j in the ith block, if treatment j appears in the ith block.
- Rank the X_{ij} within each block by assigning the rank 1 to the smallest observation in block i, the rank 2 to the second smallest, and so on to the rank k, which is assigned to the largest observation in block i, there being only k observations within each block.
- Let $R(X_{ij})$ denote the rank of X_{ij} where X_{ij} exists.
- Compute the sum of the ranks assigned to the r observed values under the jth treatment and denote this sum by R_j.
- Then R_j may be written as

$$R_j = \sum_{i-1}^{b} R(X_{ij})$$

where only r values of $R(X_{ij})$ exist under treatment j; therefore only r ranks are added to obtain R_j.

- ▷ If the observations are nonnumeric but such that they are amenable to ordering and ranking within blocks according to some criterion of interest, the ranking of each observation is noted and the values R_j for $j = 1, 2, \ldots, t$ are computed as described.
- ▷ If the ranks may be assigned in several different ways because of several observations being equal to each other, we recommend assigning the average of the disputed ranks to each of the tied observations.
- ▷ This procedure changes the null distribution of the test statistic, but the effect is negligible if the number of ties is not excessive.

Assumptions

1. The blocks are mutually independent of each other.
2. Within each block the observations have an ordinal scale of measurement. Ties cause no problem.

Test statistic Durbin (1951) suggested using the test statistic

$$T_1 = \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^{t} \left(R_j - \frac{r(k+1)}{2} \right)^2$$

- ▷ If there are ties within blocks, average ranks are used, and an adjustment needs to be made.
- ▷ Let A be the sum of the squares of the ranks and average ranks used.

$$A = \sum_{i=1}^{b} \sum_{j=1}^{t} [R(X_{ij})]^2$$

- ▷ Also compute the “correction factor” C given by

$$C = \frac{bk(k+1)^2}{4}$$

- ▷ Then the statistic T_1, corrected for ties, becomes

$$T_1 = \frac{(t-1)\sum_{j=1}^{t} \left(R_j - \frac{r(k+1)}{2} \right)^2}{A - C} = \frac{(t-1)\left[\sum_{j=1}^{t} R_j^2 - rC \right]}{A - C}$$

- ▷ An alternative procedure, equivalent to this one, is to use the ordinary analysis of variance procedure on the ranks and average ranks.
- ▷ This results in the following statistic T_2, which is merely a function of T_1. Current research indicates the approximate quantiles for T_2 are slightly more accurate than the approximate quantiles for T_1, making T_2 the preferred statistic.

$$T_2 = \frac{T_1/(t-1)}{(b(k-1) - T_1)/(bk - b - t + 1)}$$

Null Distribution

- ▷ The exact distribution of T_1 (or T_2) is difficult to find and so an approximation is usually used.
- ▷ The approximate distribution of T_1 is the chi-squared distribution with $t-1$ degrees of freedom. This approximation tends to be very conservative.
- ▷ The approximate distribution of T_2 is the F distribution (Table A22) with $k_1 = t-1$ and $k_2 = bk - b - t + 1$. This approximation tends to give inflated values of α, but closer than the values obtained from using T_1.

Hypotheses

H_0 : Each ranking of the random variables within each block is
equally likely (i.e., the treatments have identical effects)

H_1 : At least one treatment tends to yield larger observed
values than at least one other treatment

- ▷ Reject H_0 at the approximate level α if T_2 exceeds the $(1-\alpha)$ quantile of the F distribution given by Table A22 with $k_1 = t-1$ and $k_2 = bk - b - f + 1$.

Multiple Comparisons

- ▷ If the null hypothesis is rejected, then multiple comparisons between pairs of treatments may be made as follows.
- ▷ Consider treatments i and j to have different means if their rank sums R_i and R_j satisfy the inequality

$$|R_i - R_j| > t_{1-\alpha/2}\left[\frac{(A-C)2r}{bk-b-t+1}\left(1-\frac{T_1}{b(k-1)}\right)\right]^{1/2}$$

- ▷ If there are no ties, then

$$|R_i - R_j| > t_{1-\alpha/2}\left[\frac{rk(k+1)}{6(bk-b-t+1)}(b(k-1)-T_1)\right]^{1/2}$$

Example 5.9.1 *Suppose an ice cream manufacturer wants to test the taste preferences of several people for her seven varieties of ice cream. She asks each person to taste three varieties and to rank them 1, 2, and 3, with rank 1 being assigned to the favorite variety. in order to design the experiment so that each variety is compared with every other variety an equal number of times, a Youden square layout given by Federer (1963) is used. Seven people are each given three varieties to taste, and the resulting ranks as follows.*

	Variety						
Person	1	2	3	4	5	6	7
1	2	3		1			
2		3	1		2		
3			2	1		3	
4				1	2		3
5	3				1	2	
6		3				1	2
7	3		1			2	
$R_j =$	8	9	4	3	5	6	7

- $t = 7$ = total number of varieties.

- $k = 3$ = number of varieties compared at one time.

- $b = 7$ = number of people (blocks)

- $r = 3$ = number of times each variety is tested.

- $\lambda = 1$ =number of times each variety is compared with each other variety.

- The critical region of approximate size $\alpha = 0.05$ corresponds to all values of T_2 greater than 3.58, which is the 0.95 quantile of the F distribution with $k_1 = t - 1 = 6$ and $k_2 = bk - b - t + 1 = 8$, found in Table A22.

- First T_1 is found using Equation 2 because there are no ties.

$$\begin{aligned} T_1 &= \frac{12(t-1)}{rt(k-1)(k+1)} \sum_{j=1}^{t} [R_j - \frac{r(k+1)}{2}]^2 \\ &= \frac{(12)(6)}{(3)(7)(2)(4)} [(8-6)^2 + (9-6)^2 + \cdots + (7-6)^2] \\ &= 12 \end{aligned}$$

- Then T_2 is found using Equation 6.

$$\begin{aligned} T_2 &= \frac{T_1/(t-1)}{(b(k-1) - T_1)/(bk - b - t + 1)} \\ &= \frac{12/6}{(14-12)/8} = 8 \end{aligned}$$

  The test statistic T_2 is in the critical region, so the null hypothesis is rejected.

- The probability of getting perfect agreement, such as we have in this example, is the p-value, which is exactly

$$P(T_2 \geq 8) = P(T_2 = 8) = \frac{7!}{6^7} = 0.018$$

Theory

- The theoretical development of the Durbin test is very similar to that of the Friedman test. That is, the exact distribution of the Durbin test statistic is found under the assumption that each arrangement of the k ranks within a block is equally likely because of no differences between treatments.

- There are $k!$ equally likely ways of arranging the ranks within each block, and there are b blocks. Therefore each arrangement of ranks over the entire array of b blocks is equally likely and has probability $l/(k!)^b$ associated with it, because there are $(k!)^b$ different arrays possible.

- The Durbin test statistic is calculated for each array and then the distribution function is determined, just as it was for the Friedman test statistic in the previous section.

- The exact distribution is not practical to find in most cases, so the distribution of the Durbin test statistic T_1 is approximated by the chi-squared distribution with $t-1$ degrees of freedom, if the number of repetitions r of each treatment is large. The justification for this approximation is as follows.

- If the number r of repetitions of each treatment is large, the sum of the ranks, R_j, under the jth treatment is approximately normal, according to the central limit theorem. Therefore the random variable

  $$\frac{R_j - E(R_j)}{\sqrt{\mathrm{Var}(R_j)}}$$

  has approximately a standard normal distribution.

- As in the previous section, if the R_j were independent, the statistic

  $$T' = \sum_{j=1}^{t} \frac{[R_j - E(R_j)]^2}{\mathrm{Var}(R_j)}$$

  could be considered as the sum of t independent, approximately chi-squared, random variables and the distribution of T' then could be approximated with a chi-squared distribution with t degrees of freedom. But the R_j are not independent.

- Their sum is fixed as

  $$\sum_{j=1}^{l} R_j = \frac{bk(k+1)}{2}$$

  so that the knowledge of $t-1$ of the R_j enables us to state the value of the remaining R_i.

- Durbin (1951) shows that multiplication of T' by $(t-1)/t$ results in a statistic that is approximately chi-squared with $t-1$ degrees of freedom, with the form

  $$T_1 = \frac{t-1}{t} T' = \frac{t-1}{t} \sum_{j=1}^{t} \frac{[R_j - E(R_j)]^2}{\mathrm{Var}(R_j)}$$

- The sum of ranks R_j is the sum of independent random variables $R(X_{ij})$.

  $$R_j = \sum_{i=1}^{b} R(X_{ij})$$

  Each $R(X_{ij})$, where it exists, is a randomly selected integer from 1 to k.

- Therefore the mean and variance of $R(X_{ij})$ are given by Theorem 1.4.5 as

  $$E[R(X_{ij})] = \frac{k+1}{2}$$

  and

  $$\mathrm{Var}[R(X_{ij})] = \frac{(k+1)(k-1)}{12}$$

- Then the mean and variance of the R_i are easily found to be

$$E(R_j) = \sum_{i=1}^{b} e[R(X_{ij})] = \frac{r(k+1)}{2}$$

and

$$\text{Var}(R_j) = \sum_{i=1}^{b} \text{Var}[R(X_{ij})] = \frac{r(k+1)(k-1)}{12}$$

5.10 Tests with A.R.E. of 1 or more

Example 5.10.1 *The same example that was used to illustrate the Kruskal-Wallis test in Sec. 5.2 and the median test in Sec. 4.3 will also be used here for ease in comparing these methods. Four methods of growing corn resulted in the following observations and their ranks.*

Example 5.10.2 *Use the same data given in Example 5.7.1 to test*

H_0 : *The firstborn twin does not tend to be more aggressive than the other*
H_1 : *The firstborn twin tends to be more aggressive than the second twin*

The data are as follows.

Set	X_i	Y_i	D_i	$\lvert D_i \rvert$	R_i	A_i
1	86	88	+2	3	3	0.3186
2	71	77	+6	7	7	0.8134
3	77	76	−1	1.5	−1.5	−0.1560
4	68	64	−4	4	−4	−0.4316
5	91	96	+5	5.5	5.5	0.6098
6	72	72	0			
7	77	65	−12	10	−10	−1.3852
8	91	90	−1	1.5	−1.5	−0.1560
9	70	65	−5	5.5	−5.5	−0.6098
10	71	80	+9	9	9	1.1503
11	88	81	−7	8	−8	−0.9661
12	87	72	−15	11	−11	−1.7279

$$T_2 = \frac{\sum_{i=1}^{n} A_i}{\sqrt{\sum_{i=1}^{n} A_i^2}} = \frac{-2.5405}{\sqrt{8.9027}} = -0.8514$$

Example 5.10.3 *Refer to Example 5.3.1 for details of this example and a comparison with the squared ranks test.*

H_0 : *Both machines have the same variability*
H_1 : *The new machine has a smaller variance*

$$T_3 = \frac{6.2280 - 3.2669}{1.2629} = 2.3447$$

5.11 Fisher's method of randomization

Example 5.11.1 *Suppose that a random sample yielded X_is of $0, 1, 1, 0,$ and -2 and an independent random sample of Y_is gave $6, 7, 7, 4, -3, 9,$ and 14. The null hypothesis*

$$\begin{aligned} H_0 &: \quad E(X) = E(Y) \\ H_1 &: \quad E(X) \neq E(Y) \end{aligned}$$

with the randomization test for two independent samples.

- $n = 5$ and $m = 7$, so there are $\binom{12}{5} = 792$ ways of forming a subset containing 5 of the 12 numbers.
- $792 \cdot 0.025 = 19.8 \approx 20$ groups.
- $$T_1 = \sum_{i=1}^{5} X_i = 0 + 1 + 1 + 0 - 2 = 0$$ $$p\text{-value} = \frac{2(11)}{792} = 0.028$$

Example 5.11.2 *Suppose that eight matched pairs resulted in the following differences: $-7, -3, 0, +5, +1, -10$. The zero is discarded, and we have $D_1 = -16, D_2 = -4, D_3 = -7, D_4 = -3, D_5 = +5, D_6 = +1, D_7 = -10$.*

$$H_0 : d_{0.50} = 0 \quad v.s. \quad H_1 : d_{0.50} \neq 0$$

$$w_{0.025} = 4$$

$$w_{0.975} = \sum_{i=1}^{7} |D_i| - w_{0.025} = 42$$

$$T_2 = \sum \text{positive} D_i = 5 + 1 = 6$$

$$p\text{-value} = \frac{2(8)}{2^7} = \frac{16}{128} = 0.125$$

5.12 Summary

1. *Stem-and-leaf method (Tukey, 1977)*: A convenient method of arranging observations in increasing order. 126

2. - # of $(X_i, Y_j) = mn$
   - If $k = \#$ of $X_i - Y_j > 0$, then $T = k + n(n+1)/2$.
   - $T = n(n+1)/2$ if no Ys are smaller than any of Xs.
   - The borderline value of T, where H_0 is barely accepted, is given in Table A7 as $w_{\alpha/2}$.
   - By subtracting $n(n+1)/2$ from $w_{\alpha/2}$, we find the borderline value of k.
   - Want to find the value of d that we can add to the Ys to achieve barely this borderline value of k.

- If we add the maximum of all the differences $X_i - Y_j$ to each of the Ys, then none of the Xs will be greater than the adjusted Ys.
- Add the kth largest difference $X_i - Y_j$ to each of Ys, we achieve the borderline case: fewer than k pairs satisfy $X_i > Y_j + d$, and at least k pairs satisfy $X_i > Y_j + d$. In this way we obtain the largest value of d that results in acceptance of $H_0 :\ E(X) = E(Y) + d$.
- By reversing the procedure and working from the lower end, we obtain the smallest value of d that results in acceptance of the same hypothesis.

. 134

3.
- Each arrangement of the ranks 1 to N into groups of sizes $n_1, n_2, \ldots, n_k$, which is equally likely, and occurs with probability $n_1!n_2! \cdots n_k!/N!$.
- The value of T is computed for each arrangement.
- Example: $n_1 = 2, n_2 = 1$, and $n_3 = 1$

	Sample			
Arrangement	1	2	3	T
1	1, 2	3	4	2.7
2	1, 2	4	3	2.7
3	1, 3	2	4	1.8
4	1, 3	4	2	1.8
5	1, 4	2	3	0.3
6	1, 4	3	2	0.3
7	2, 3	1	4	2.7
8	2, 3	4	1	2.7
9	2, 4	1	3	1.8
10	2, 4	3	1	1.8
11	3, 4	1	2	2.7
12	3, 4	2	1	2.7

- Distribution function:

x	$f(x) = P(T = x)$	$F(x) = P(T \leq x)$
0.3	$2/12 = 1/6$	$1/6$
1.8	$4/12 = 1/3$	$1/2$
2.7	$6/12 = 1/2$	1.0

- Large sample approximation for T
- $\frac{R_i - E(R_i)}{\sqrt{\mathrm{Var}(R_i)}} \approx N(0, 1)$ where
  $E(R_i) = \frac{n_i(N+1)}{2}$ and
  $\mathrm{Var}(R_i) = \frac{n_i(N+1)(N-n_i)}{12}$
- $T' = \sum_{i=1}^{k} \frac{(R_i - [n_i(N+1)/2])^2}{n_i(N+1)(N-n_i)/12} \approx \chi^2_k$
- R_i are dependent since $\sum R_i = N(N+1)/2$.
- Kruskal (1952) showed that if the ith term in T' is multiplied by $(N - n_i)/N$, then

$$T = \sum_{i=1}^{k} \frac{(R_i - [n_i(N+1)/2])^2}{n_i(N+1)N/12} \approx \chi^2_{k-1}$$

which is a rearrangement of the terms in Eq. 5.

- For two samples the Kruskal-Wallis test is equivalent to the Mann-Whitney test.

. . . . 139

4.
- Whenever two random variables X and Y are identically distributed except for having different means μ_1 and μ_2, $X-\mu_1$ and $Y-\mu_2$ not only have zero means, but they are identically distributed also.
- This means $U=|X-\mu_1|$ has the same distribution as $V=|Y-\mu_2|$. Both have the mean zero.
- Every assignment of ranks of the Us is equally likely.
- The ranks of Us and Vs are the same the ranks of U^2s and V^2s.
- Use the squared (score) ranks and not the ranks themselves.

$$a(R)=R^2$$

- $T=\sum a(R_i)$ where R_i denote the ranks of U_i in the combined sample.
- To use the large sample normal approximation for T it is necessary to find the mean and variance of T when H_0 is true.
- $E(T)=\sum_{i=1}^{n} E(a(R_i))=n\sum_{j=1}^{N}\frac{1}{N}a(j)=n\bar{a}$
- $\mathrm{Var}(T)=\sum_{i=1}^{n}\mathrm{Var}[a(R_i)]+\sum_{i\neq j}\mathrm{Cov}[a(R_i),a(R_j)]$
- $\mathrm{Var}[a(R_i)]=\frac{1}{N}\sum_{k=1}^{N}[a(k)-\bar{a}]^2=A$
- $\mathrm{Cov}[a(R_i),a(R_j)]=\sum_{k\neq l}\frac{[a(k)-\bar{a}][a(l)-\bar{a}]}{N(N-1)}$
- $\mathrm{Cov}[a(R_i),a(R_j)]=\sum_{k=1}^{N}[a(k)-\bar{a}]\sum_{l=1}^{N}[a(l)-\bar{a}]\frac{1}{N(N-1)}-\sum_{k=1}^{N}[a(k)-\bar{a}]^2\frac{1}{N(N-1)}$
- $\mathrm{Cov}[a(R_i),a(R_j)]=-\frac{A}{N-1}$
- $\mathrm{Var}(T)=nA-n(n-1)\frac{A}{N-1}=\frac{n(N-n)}{N-1}A$
- $\mathrm{Var}(T)=\frac{nm}{(N-1)N}\sum_{i=1}^{N}[a(i)-\bar{a}]^2$
- Interest in the case $a(R)=R^2$
- The denominator of Eq. 4 is what the square root of Eq. 24 by using

$$\sum_{i=1}^{N}[a(i)-\bar{a}]^2=\sum_{i=1}^{N}[a(i)]^2-N(\bar{a})^2$$

- The extension of the two-sample case to the k-sample case is completely analogous to the extension of the two-sample Mann-Whitney test to the k-sample Kruskal-Wallis test.
- $S_1,\ldots,S_k$: Sums of scores for each k samples.
- $E(S_i)=n_i\bar{a}$ and $\mathrm{Var}(S_i)=\frac{n_i(N-n_i)}{(N-1)N}\sum_{i=1}^{N}[a(i)-\bar{a}]^2$
- $T_2=\sum_{i=1}^{k}\frac{[S_i-E(S_i)]^2}{\mathrm{Var}(S_i)}=\sum_{i=1}^{k}\frac{(S_i-n_i\bar{a})^2}{n_iD^2}$ where $D^2=\frac{1}{N-1}\left\{\sum_{i=1}^{N}[a(i)]^2-N(\bar{a})^2\right\}$
- $T_2=\frac{1}{D^2}\left[\sum_{j=1}^{k}\frac{S_j^2}{n_j}-N(\bar{a})^2\right]$
- If the populations of X and Y have the normal distributions the appropriate statistic to use is the ratio of the two sample variances:

$$F=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}{\frac{1}{m-1}\sum_{j=1}^{m}(Y_j-\bar{Y})^2}$$

- The F test is very sensitive to the assumption of normality.
- $P\{F_{m,n} \leq x\} = P\{F_{n,m} \geq 1/x\}$
- The F test is not very safe test to use unless one is sure that the populations are normal.
- If the squared rank test is used instead of the F test when the populations are normal the A.R.E. is only $15/(2\pi^2) = 0.76$, 1.08 for the double exponential distribution, 1.00 for the uniform distribution.

. . . . 145

5. *Pearson's product moment correlation coefficient*:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2\right]^{1/2}}$$

. . . . 145

6. *Pearson's r:* Pearson's r is a measure of the strength of the linear association between X and Y. 145

7. *Spearman's Rho*:

$$\rho = \frac{\sum R(X_i)R(Y_i) - n((n+1)/2)^2}{\left(\sum R(X_i)^2 - n((n+1)/2)^2\right)^{1/2} \left(\sum R(Y_i)^2 - n((n+1)/2)^2\right)^{1/2}}$$

If there are no ties:

$$\rho = 1 - \frac{6\sum[R(X_i) - R(Y_i)]^2}{n(n^2-1)}$$

. . . . 146

8. *Spearman rank correlation coefficient* The Spearman rank correlation coefficient is often used as a test statistic to test for independence between two random variables. The test statistic is

$$\rho = \frac{\sum R(X_i)R(Y_i) - n((n+1)/2)^2}{\left(\sum R(X_i)^2 - n((n+1)/2)^2\right)^{1/2} \left(\sum R(Y_i)^2 - n((n+1)/2)^2\right)^{1/2}}$$

If there are no ties:

$$\rho = 1 - \frac{6\sum[R(X_i) - R(Y_i)]^2}{n(n^2-1)}$$

. . . . 147

9. *Null distribution*:

- Exact quantiles of ρ when X and Y are independent are given in Table A10 for $n \leq 30$ and no ties.
- For larger n, or many ties: (percentile)

$$w_p \approx \frac{z_p}{\sqrt{n-1}}$$

. . . . 147

10. *Spearman's ρ* Spearman's ρ is insensitive to some types of dependence, so it is better to be specific as to what type of dependency 147

11. *Two-tailed test*

    H_0: The X_i and Y_i are mutually independent
    H_1: Either (a) there is a tendency for the larger values of X to be paired with the larger values of Y, or (b) there is a tendency for the smaller values of X to be paired with the larger values of Y

    Reject H_0 at the level α if $|\rho|$ is greater than its $1-\alpha/2$ quantile obtained from Table A10 or Eq. 11. 147

12. *Lower-tailed test for negative correlation*

    H_0: The X_i and Y_i are mutually independent
    H_1: There is a tendency for the smaller values of X to be paired with the larger values of Y, and vice versa

    Reject H_0 at the level α if $\rho < -w_{1-\alpha}$ where $w_{1-\alpha}$ is found either in Table A10 or from Eq. 11. 147

13. *Upper-tailed test for positive correlation*

    H_0: The X_i and Y_i are mutually independent
    H_1: There is a tendency for the larger values of X and Y to be paired together

    Reject H_0 at the level α if $\rho > w_{1-\alpha}$ where $w_{1-\alpha}$ is found either in Table A10 or from Eq. 11. 148

14. *Kendall's τ*: No ties

    $$\tau = \frac{N_c - N_d}{n(n-1)/2}$$

    where N_c and N_d are the number of concordant and discordant pairs of observations, respectively.
    Ties

    $$\tau = \frac{N_c - N_d}{N_c + N_d}$$

    - If $\frac{Y_j - Y_i}{X_j - X_i} > 0$, add 1 to N_c (concordant).
    - If $\frac{Y_j - Y_i}{X_j - X_i} < 0$, add 1 to N_d (discordant).
    - If $\frac{Y_j - Y_i}{X_j - X_i} = 0$, add 1/2 to N_c and N_d.
    - $X_i = X_j$, no comparison is made.

   149

15. *Measure of correlation (Kendall, 1938)* In case of no ties: $\tau = \frac{N_c - N_d}{n(n-1)/2}$. $\tau = 1$ if all pairs are concordant. $\tau = -1$ if all pairs are discordant.
    Ties

    $$\tau = \frac{N_c - N_d}{N_c + N_d}$$

    If $\frac{Y_j - Y_i}{X_j - X_i} > 0$, add 1 to N_c (concordant). If $\frac{Y_j - Y_i}{X_j - X_i} < 0$, add 1 to N_d (discordant). If $\frac{Y_j - Y_i}{X_j - X_i} = 0$, add 1/2 to N_c and N_d. $X_i = X_j$, no comparison is made. 149

16. *Kendall's τ test*: Kendall's τ may also be used as a test statistic to test the null hypothesis of independence between X and Y.

$$T = \begin{cases} N_c - N_d & \text{in case of no ties or few ties,} \\ \frac{N_c - N_d}{N_c + N_d} & \text{in case of many ties.} \end{cases}$$

....150

17. *Two-tailed test*

H_0: X and Y are independent
H_1: Pairs of observations either tend to be concordant, or tend to be discordant.

Reject H_0 at the level α if T (or τ) is less than its $\alpha/2$ quantile or greater than its $1 - \alpha/2$ quantile in the null distribution. 150

18. *Lower-tailed test*

H_0: X and Y are independent
H_1: Pairs of observations trend to be discordant.

Reject H_0 at the level α if T (or τ) is less than its α quantile in the null distribution. 150

19. *Upper-tailed test*

H_0: X and Y are independent
H_1: Pairs of observations trend to be concordant.

Reject H_0 at the level α if T (or τ) is less than its $1-\alpha$ quantile in the null distribution. 151

20. *Daniel's test for trend*: Tests of trend based on Spearman's ρ or Kendall's τ are generally considered to be more powerful than the Cox and Stuart test (Sec. 3.5). 152

21. *Jonckheere-Terpstra test*: Either Spearman's ρ or Kendall's τ can be used in the case of several independent samples to test the null hypothesis that all of the samples came from the same distribution.

$$H_0 : \; F_1(x) = F_2(x) = \cdots = F_k(x)$$

against the ordered alternative that the distributions differ in a specified direction

$$H_1 : \; F_1(x) \geq F_2(x) \geq \cdots \geq F_k(x)$$

with at least one inequality. The alternative is sometimes written as

$$H_1(x) : \; E(Y_1) \leq E(Y_2) \leq \cdots \leq E(Y_k).$$

....152

22. *Kendall's partial correlation coefficient*: $n = 3$, Pearson's partial correlation coefficient

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$n = 3$, Kendall's τ partial correlation coefficient

$$\tau_{12.3} = \frac{\tau_{12} - \tau_{13}\tau_{23}}{\sqrt{(1-\tau_{13}^2)(1-\tau_{23}^2)}}$$

. . . . 154

23. The *regression of Y on X* is $E(Y|X = x)$. The regression equation is $y = E(Y|X = x)$. 155

24. The regression of Y on X is linear regression if the regression equation is of the form

$$E(Y|X = x) = \alpha + \beta x$$

for some constant α, called the *y-intercept*, and β, called the *slope*. 155

25. The *least squares* method for choosing estimates a and b of α and β in the regression equation $y = \alpha + \beta x$ is the method that minimizes the sum of squared deviations

$$SS = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

for the observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. 156

26. *Least squares estimates*:

$$\begin{aligned} y &= a + bx \\ b &= \frac{\operatorname{Cov}(X,Y)}{S_x^2} = \rho\frac{S_y}{S_x} = \frac{n\sum_{i=1}^n X_iY_i - \left(\sum_{i=1}^n X_i\right)\left(\sum_{i=1}^n Y_i\right)}{n\sum_{i=1}^n X_i^2 - \left(\sum_{i=1}^n X_i\right)^2} \\ a &= \bar{Y} - b\bar{X} \end{aligned}$$

. . . . 156

27. *Testing the slope*: Let β_0 represent some specified number. For each pair (X_i, Y_i) compute $Y_i - \beta_0 X_i = U_i$. Then find the Spearman rank correlation coefficient ρ on the pairs (X_i, U_i). 156

28. *A confidence interval for the slope*: For each pair of points (X_i, Y_i) and (X_j, Y_j), such that $i < j$ and $X_i \neq X_j$, compute the two-point slope

$$S_{ij} = \frac{Y_j - Y_i}{X_j - X_i}$$

- ▶ N: The number of slopes computed.
- ▶ Order the slopes obtained and let

$$S^{(1)} \leq S^{(2)} \leq \cdots \leq S^{(N)}$$

- ▶ Find $w_{1-\alpha/2}$ from Table A11.

- $r = (N - w_{1-\alpha/2})/2$ and $s = (N + w_{1-\alpha/2})/2 + 1 = N + 1 - r$.
- $1 - \alpha$ CI for β:

$$(S^{(r)}, S^{(s)})$$

. . . . 157

29. *Minimum sum of squares*:

$$\begin{aligned} SS_{\min} &= \sum_{i=1}^{n}(Y_i - \bar{Y})^2 - \frac{S_{xy}^2}{S_x} \\ &= (1 - r^2)\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \end{aligned}$$

where r is the Pearson product moment correlation coefficient. 160

30. *Relationship between the slopes S_{ij} and Kendall's τ*: $S_{ij} = \frac{Y_i - Y_j}{X_i - X_j} = \beta_0 + \frac{U_i - U_j}{X_i - X_j}$. The slope S_{ij} is greater than β_0 or less than β_0 according to whether the pair (X_i, U_i) and (X_j, U_j) is concordant or discordant. 161

31. *Monotonically increasing (decreasing)*: If $E(Y|X)$ becomes smaller as X increases the regression is *monotonically increasing (decreasing)*. 162

32. *Symmetric distribution*: The distribution of a random variable X is *symmetric* about a line $x = c$, for some constant c, if the probability of $X \leq c - x$ equals the probability of $X \geq c + x$ for each possible value of x. 167

33. If a distribution is symmetric, the mean (if it exists) coincides with the median because both are located exactly in the middle of the distribution, at the line of symmetry. 167

34. *Signed rank*:

$$\begin{aligned} R_i = &\text{ the rank assigned to } (X_i, Y_i) \text{ if } D_i = Y_i - X_i \text{ is positive} \\ &\quad (\text{i.e., } Y_i > X_i) \\ R_i = &\text{ the negative of the rank assigned to } (X_i, Y_i) \text{ if } D_i \text{ is negative} \\ &\quad (\text{i.e., } Y_i < X_i) \end{aligned}$$

. . . . 169

35. *Statistics of Wilcoxon signed rank test*: $T^+ = \sum(R_i$ where D_i is positive$)$ 169

36. The distribution of a random variable X is *symmetric* about a line $x = c$, for some constant c, if the probability of $X \leq c - x$ equals the probability of $X \geq c + x$ for each possible value of x. 170

Part II

Appendices

Appendix A

Preface

- The author decided to write a book on nonparametric statistics over thirty years ago.
- He wanted a book that could be used as a textbook in one-semester course on nonparametric statistics.
- But he also wanted a book that could be used as a quick reference to the most useful nonparametric methods.
- The second edition was published twenty years ago.
- This third edition adds many new exercises and problems, some new worked-out examples, and many updated references to related material.
- On the one hand, users of this book would find computer instructions very useful, but on the other hand, computer packages change more rapidly than this book does, and computer tips become outdated quickly.
- Computer software packages with extensive nonparametric programs:
    - Minitab-`http://www.minitab.com`
    - PASS or NCSS-`http://www.ncss.com`
    - Resampling Stats-`http://www.statistics.com`
    - *SAS-`http://www.sas.com`*
    - *SPSS-`http://www.spss.com`*
    - STATA-`http://www.stata.com`
    - *STATISTICA-`http://www.statsoft.com`*
    - StatMost-`http://www.dataxiom.com`
    - StatXact-`http://www.cytel.com`
    - SYSTAT-`http://www.spss.com/software/science/systat`
    - *S-Plus- `http://www.insightful.com/products/product.asp?PID=3`*
- The STATA website contains links to these and other software providers.
- Prerequisite: College algebra and a modicum of mathematical ability.
- Chapters 1 and 2 are to bring such a student up to the level of knowledge required to understand the theory and methods in the rest of the book.

- This book has been used successfully as a textbook both at the graduate and undergraduate levels.

- At the undergraduate level most instructors find the Problems and Theory sections too challenging for their students.

- I have taught this course at the graduate level countless times, and only once did I omit Chapters 1 and 2, with disastrous results I might add!

- Other instructors have told me they omitted Chapters 1 and 2 without a problem, but their students have sometimes told me they had to go back and read those chapters on their own before they could grasp the later material.

Appendix B

Introduction

- *Science*: Truth ascertained by observation, experiment, and induction.
- A vast amount of time, money, and energy is being spent by society today in the pursuit of science.
- One experiment, with one set of observations, may lead two scientists to two different conclusions.
- *Example*: A scientist places a rat into a pen with two doors, both closed. One door is painted red and the other blue. The rat is subjected to 20 minutes of music of the type popular with today's teenagers. After this experience, both doors are opened and the rat run out of the pen. The scientist notes which door the rat chose. This experiment is repeated 10 times, each time using a different rat.
- Later the scientist conducts a second experiment.
  - He injects a certain drug into the bloodstream of each of 10 rats.
  - Five minutes later he examines the rats and finds that 7 are dead, and the other 3 are apparently healthy.
  - However, since only 7 are dead, he recalls the previous experiment and concludes that such a result could easily have occurred by chance and there is no proof that the drug injections are dangerous.
- *Statistics:* Provide the means for measuring the amount of subjectivity that goes into the scientists' conclusions and thus to separate "science" from "opinion".
- This is accomplished by setting up a theoretical "model" for the experiment.
- *Nonparametric methods* have become essential tools in the workshop of the applied scientist who needs to do statistical analysis.
- When the price for making a wrong decision is high, applied scientists are very concerned that the statistical methods they are using are not based on assumptions that appear to be invalid, or are impossible to verify.
- *Nonparametric statistical methods*
  - Use a simpler model.
  - Involve less computation work and easier and quicker to apply than other statistical methods.

- Much of the theory behind the nonparametric methods may be developed rigorously, using no mathematics beyond high school algebra.
- They are often more powerful than the parametric methods if the assumptions behind the parametric model are not true.

Appendix C

無母數統計報告

C.1　報告寫作的注意事項

請自己從各大學圖書館的考古題網站：如

1. 臺閩地區圖書館暨資料單位名錄
   `http://wwwsrch.ncl.edu.tw/libdir/`

   碩士班入學考古題

   國立大學

2. 國立臺灣大學圖書館
   `http://www.lib.ntu.edu.tw/exam/graduate/college.htm`
3. 國立臺灣師範大學圖書館
   `http://www.lib.ntnu.edu.tw/libweb/qlink/exam.php`
4. 國立政治大學圖書館
   `http://www.lib.nccu.edu.tw/exam/index.htm`
5. 國立交通大學浩然圖書館
   `http://www.lib.nctu.edu.tw/n_exam/index.html`
6. 國立清華大學圖書館
   `http://www.lib.nthu.edu.tw/library/department/ref/exam/index.htm`
7. 國立中央大學圖書館
   `http://www.lib.ncu.edu.tw/cexamn.html`
8. 國立中興大學圖書館
   `http://recruit.nchu.edu.tw/`
9. 國立中正大學圖書館
   `http://www.lib.ccu.edu.tw/gradexam/kind.htm`
10. 國立成功大學圖書館
    `http://eserv.lib.ncku.edu.tw/exam/index.php`
11. 國立中山大學圖書館
    `http://www.lib.nsysu.edu.tw/exam`

12. 國立高雄師範大學圖書館
`http://www.nknu.edu.tw/~math/mathweb/frame.asp`

13. 國立東華大學圖書館
`http://www.lib.ndhu.edu.tw/index.phtml?path=,175,164&language=zh_tw`

14. 國立海洋大學圖書館
`http://www.lib.ntou.edu.tw/exam/exam.htm`

15. 國立台北大學圖書館
`http://www.ntpu.edu.tw/library/lib/ntpulib_exam.htm`

16. 國立暨南大學圖書館
`http://www.library.ncnu.edu.tw/download/old_exam.htm`

17. 國立中正理工學院
`http://www.lib.ccit.edu.tw/search/search6.htm`

私立大學

18. 私立文化大學圖書館
`http://www.lib.pccu.edu.tw/exam.html`

19. 私立淡江大學圖書館
`http://www.lib.tku.edu.tw/exam/exam-tku.shtml`

20. 私立中原大學圖書館
`http://www.lib.cycu.edu.tw/exams_new/exams_new.html`

21. 私立輔仁大學圖書館
`http://lib.fju.edu.tw/collection/examine.htm`

22. 私立逢甲大學圖書館
`http://www.admission.fcu.edu.tw/test_question.htm`

23. 私立元智大學圖書館
`http://www.yzu.edu.tw/library/index.php/content/view/152/253/`

24. 私立銘傳大學圖書館
`http://140.131.66.3/`

25. 私立靜宜大學圖書館
`http://www.lib.pu.edu.tw/new/exam/`

26. 私立東吳大學
`http://www.scu.edu.tw/entrance/exam92/index.htm`

27. 私立中山醫學大學圖書館
`http://www.lib.csmu.edu.tw/overlib/2206.php`

28. 私立高雄醫學大學圖書館
`http://www.kmu.edu.tw/%7Elib/kmul/exam.htm`

29. 私立大同大學圖書館
`http://www.library.ttu.edu.tw/eresource/exam.htm`

30. 私立義守大學
`http://www1.isu.edu.tw/exam/exam/`

31. 私立世新大學圖書館
`http://lib.shu.edu.tw/search_taskpaper.asp`

32. 私立南華大學圖書館
`http://libserver2.nhu.edu.tw/20.htm`

33. 私立華梵大學圖書館
`http://huafan.hfu.edu.tw/~lib/srvc/exam2/exam2.htm`

34. 私立玄奘大學
`http://www.hcu.edu.tw/hcu2/old/old.asp`

或研究所升學的統計或機率參考書，找出十題研究所入學考題與無母數統計所教授的觀念相關的題目，每章題目找兩題。寫作的注意事項：

1. 第一行標明出處及關鍵詞，第二行題目所在的網址或參考書的作者、年代、版次及書名。
2. 題目不變，中文就用中文，英文就用英文，不用翻譯成中文。
3. 解題過程詳細講解每一個解題步驟，請參考範例。
4. 預計第三、五章結束需要上台報告自己每階段的作品（五題）。
5. 使用提供的範本作報告，檔名：m962040002蔡仲信.tex。
6. 學期末繳交列印報告，m962040002蔡仲信.tex和m962040002蔡仲信.pdf等檔案。

C.2 報告範例

無母數統計報告

蔡仲信
國立中山大學應用數學系
`tsaijs1@gmail.com`
2008-06-10

1. 【94中山應數統計組，隨機變數分解求期望值】
`http://www.lib.nsysu.edu.tw/exam/master/sci/math/94.pdf`
An urn contains $n+m$ balls, of which n are red and m are black. They are withdrawn from the urn, one at a time and without replacement. Let Y denote the number of red balls chosen after the first but before the second black ball has been chosen. Number the red balls from 1 to n. Find $E[Y]$.

Ans: 在第一個黑球被選中後且第二個黑球被選中前，若紅球i 被選中，則令$Y_i = 1, i = 1, \ldots, n$。所以$Y = \sum_{i=1}^{n} Y_i$。

$$\begin{aligned} E[Y_i] &= P(Y_i = 1) \\ &= P(\text{從}m+1 \text{ 個球中選中紅球}i\ ) \\ &= 1/(m+1) \text{ 因爲}m+1 \text{ 個球被選中的機率皆相等} \end{aligned}$$

因此，

$$E[Y] = n/m + 1$$

2. 【95東華國經，多樣本The Kruskal-Wallis 檢定】
`http://econ.ndhu.edu.tw/attachment/490_3.pdf`
將30個體質相類似的人隨機分成5組，試吃五種不同的減肥藥，三個月後，紀錄每個人減肥前後的體重差，結果如下:

A 餐	2	7	9	5	3	10
B 餐	-1	6	7	0	2	4
C 餐	5	7	13	11	2	10
D 餐	3	7	6	-1	-1	4
E 餐	-3	3	5	-4	-2	7

(a) 在$\alpha = 0.05$下，以無母數統計檢定法檢定五種減肥餐的效果是否相同?

(b) 若上述資料符合變異數分析的各種假設，在在$\alpha = 0.05$下，檢定五種減肥餐的效果是否相同?

Ans:

(a) 先將30筆資料混合然後排序，之後依序取等級: （註：如果遇到相同的資料，取平均等級）

A餐		B餐		C餐		D餐		E餐	
體重差	等級	體重差	等級	體重差	等級	體重差	等級	體重差	等級
2	9	-1	5	5	17	3	12	-3	2
7	23	6	19.5	7	23	7	23	3	12
9	26	7	23	13	30	6	19.5	5	17
5	17	0	7	11	29	-1	5	-4	1
3	12	2	9	2	9	-1	5	-2	3
10	27.5	4	14.5	10	27.5	4	14.5	7	23
$n_1 = 6$	$\overline{X}_{1\cdot} = \frac{229}{12}$	$n_2 = 6$	$\overline{X}_{2\cdot} = 13$	$n_3 = 6$	$\overline{X}_{3\cdot} = \frac{271}{12}$	$n_4 = 6$	$\overline{X}_{4\cdot} = \frac{79}{6}$	$n_5 = 6$	$\overline{X}_{5\cdot} = \frac{29}{3}$

上面的表，其中n_i爲第i組樣本數、$\overline{X}_{i\cdot}$爲第i組樣本等級的平均

$$\overline{X}_{\cdot\cdot} = \frac{\sum_{i=1}^{5} \overline{X}_{i\cdot}}{5} = 15.5$$

$$SSR = \sum_{i=1}^{5} (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 = 747.186$$

$$n = \sum_{i=1}^{5} n_i = 30$$

H_0：五種減肥餐效果相同 v.s H_1：五種減肥餐效果不全相同

$$\begin{aligned} H^* &= \frac{12}{n(n+1)} \times SSR \\ &= \frac{12}{30(30+1)} \times 747.186 \\ &= 9.641 > \chi^2_{0.05}(4) = 9.488 \end{aligned}$$

由上面可知，在$\alpha = 0.05$下，有足夠的證據能夠拒絕H_0。因此，這五種減肥餐效果不全相同。

(b) 如果上述資料符合變異數分析的各種假設，就直接對資料取平均。$\overline{X}_{i\cdot}$爲第i組資料的平均，X_{ij}爲第i組的第j筆資料：$\overline{X}_{1\cdot} = 6, \overline{X}_{2\cdot} = 3, \overline{X}_{3\cdot} = 8, \overline{X}_{4\cdot} = 3, \overline{X}_{5\cdot} = 1,$

$$\begin{aligned}
\overline{X}_{\cdot\cdot} &= \frac{\sum_{i=1}^{5} \overline{X}_{i\cdot}}{5} = 4.2 \\
SSR &= \sum_{i=1}^{5} (\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2 = 30.8 \\
SST &= \sum_{j=1}^{5} \sum_{i=1}^{5} (X_{ij} - \overline{X}_{\cdot\cdot})^2 = 536.8 \\
SSE &= SST - SSR = 506
\end{aligned}$$

則

$$\begin{aligned}
F^* &= \frac{SSR/4}{SSE/25} \\
&= \frac{30.8/4}{506/25} \\
&= 0.380 < F_{0.05}(4, 25) = 2.76
\end{aligned}$$

由上面可知，在$\alpha = 0.05$下，沒有足夠的證據能夠拒絕H_0。因此，這五種減肥餐效果相同。

3. 【95台大財金乙，Mann-Whitney-Wilcoxon雙樣本檢定法】
`http://www.lib.ntu.edu.tw/exam/graduate/95/378.pdf`
研究機構欲了解重要之A管理理論與B管理理論何者較有效? (即績效中位數η_A、η_b何者較大?)，乃進行一項實驗：隨機抽出32員工施以A 理論的環境，另外抽出32員工施以B理論的環境，年終各員工之績效分別爲$Y_{ij}, i = A, B, j = 1, 2, \ldots, 32$。若已求得$Y_{Aj}$對應之等級和$R_A = 1393$，令顯著水準$\alpha = 0.05$，依Mann-Whitney-Wilcoxon檢定法檢定之。

Ans: 考慮假說檢定H_0：B理論有效 v.s H_1：A理論有效

$$\begin{aligned}
T^* &= R_A - \frac{n_A(n_A + 1)}{2} \\
&= 1393 - \frac{32 \times 33}{2} \\
&= 865
\end{aligned}$$

利用標準常態近似

$$\begin{aligned}
Z^* &= \frac{T^* - \frac{n_A n_B}{2}}{\sqrt{\frac{n_A n_B (n_A + n_B + 1)}{12}}} \\
&= \frac{865 - \frac{32 \times 32}{2}}{\sqrt{\frac{32 \times 32 \times (32 + 32 + 1)}{12}}} \\
&= 4.74 > Z_{0.05} = 1.645
\end{aligned}$$

由上面可知，在$\alpha = 0.05$下，有足夠的證據能夠拒絕H_0。因此，A理論有效。

Appendix D

What is Statistics?

Contents

Statistics: The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling. (*American Heritage Dictionary®*)

D.1 Introduction

Statistics, branch of mathematics that deals with the collection, organization, and analysis of numerical data and with such problems as experiment design and decision making.

D.2 History

Simple forms of statistics have been used since the beginning of civilization, when pictorial representations or other symbols were used to record numbers of people, animals, and inanimate objects on skins, slabs, or sticks of wood and the walls of caves. Before 3000 BC the Babylonians used small clay tablets to record tabulations of agricultural yields and of commodities bartered or sold. The Egyptians analyzed the population and material wealth of their country before beginning to build the pyramids in the 31st century BC. The biblical books of Numbers and 1 Chronicles are primarily statistical works, the former containing two separate censuses of the Israelites and the latter describing the material wealth of various Jewish tribes. Similar numerical records existed in China before 2000

BC. The ancient Greeks held censuses to be used as bases for taxation as early as 594 BC. *See* Census.

The Roman Empire was the first government to gather extensive data about the population, area, and wealth of the territories that it controlled. During the Middle Ages in Europe few comprehensive censuses were made. The Carolingian kings Pepin the Short and Charlemagne ordered surveys of ecclesiastical holdings: Pepin in 758 and Charlemagne in 762. Following the Norman Conquest of England in 1066, William I, king of England, ordered a census to be taken; the information gathered in this census, conducted in 1086, was recorded in the Domesday Book. Registration of deaths and births was begun in England in the early 16th century, and in 1662 the first noteworthy statistical study of population, *Observations on the London Bills of Mortality*, was written. A similar study of mortality made in Breslau, Germany, in 1691 was used by the English astronomer Edmond Halley as a basis for the earliest mortality table. In the 19th century, with the application of the scientific method to all phenomena in the natural and social sciences, investigators recognized the need to reduce information to numerical values to avoid the ambiguity of verbal description.

At present, statistics is a reliable means of describing accurately the values of economic, political, social, psychological, biological, and physical data and serves as a tool to correlate and analyze such data. The work of the statistician is no longer confined to gathering and tabulating data, but is chiefly a process of interpreting the information. The development of the theory of probability increased the scope of statistical applications. Much data can be approximated accurately by certain probability distributions, and the results of probability distributions can be used in analyzing statistical data. Probability can be used to test the reliability of statistical inferences and to indicate the kind and amount of data required for a particular problem.

D.3 Statistical methods

The raw materials of statistics are sets of numbers obtained from enumerations or measurements. In collecting statistical data, adequate precautions must be taken to secure complete and accurate information.

The first problem of the statistician is to determine what and how much data to collect. Actually, the problem of the census taker in obtaining an accurate and complete count of the population, like the problem of the physicist who wishes to count the number of molecule collisions per second in a given volume of gas under given conditions, is to decide the precise nature of the items to be counted. The statistician faces a complex problem when, for example, he or she wishes to take a sample poll or straw vote. It is no simple matter to gauge the size and constitution of the sample that will yield reasonably accurate predictions concerning the action of the total population.

In protracted studies to establish a physical, biological, or social law, the statistician may start with one set of data and gradually modify it in light of experience. For example, in early studies of the growth of populations, future change in size of population was predicted by calculating the excess of births over deaths in any given period. Population statisticians soon recognized that rate of increase ultimately depends on the number of births, regardless of the number of deaths, so they began to calculate future population growth on the basis of the number of births each year per 1000 population. When predictions based on this method yielded inaccurate results, statisticians realized that other limiting factors exist in population growth. Because the number of births possible depends on the number of women rather than the total population, and because women bear children during only part of their total lifetime, the basic datum used to calculate future population size is now the number of live births per 1000 females of childbearing age. The predictive value of this basic datum can be further refined by combining it with other data on the percentage of women who remain childless because of choice or circumstance, sterility, contraception, death before the end of the childbearing period, and other limiting factors. The excess of births over deaths, therefore, is meaningful only as an indication of gross population growth over a definite period in the past; the number of births per 1000 population is meaningful only as an expression of the proportion of increase during a similar period; and the number of live births per 1000 women of childbearing age is meaningful for predicting future size of populations.

D.4 Tabulation and presentation of data

INTERVALS	INTERVAL MIDPOINTS	FREQUENCY	RELATIVE FREQUENCY	CUMULATIVE FREUQENCY	CUMULATIVE RELATIVE FREUQENCY
0-10	5	20	0.017	20	0.017
10-20	15	15	0.012	35	0.029
20-30	25	18	0.015	53	0.044
30-40	35	25	0.021	78	0.065
40-50	45	44	0.037	122	0.102
50-60	55	88	0.073	210	0.175
60-70	65	222	0.185	432	0.360
70-80	75	335	0.279	767	0.639
80-90	85	218	0.182	985	0.821
90-100	95	215	0.179	1200	1.000

The collected data must be arranged, tabulated, and presented to permit ready and meaningful analysis and interpretation. To study and interpret the examination-grade distribution in a class of 30 pupils, for instance, the grades are arranged in ascending order: $30, 35, 43, 52, 61, 65, 65, 65, 68, 70, 72, 72, 73, 75, 75, 76, 77, 78, 78, 80, 83, 85, 88, 88, 90, 91, 96, 97, 100, 100$. This progression shows at a glance that the maximum is 100, the minimum 30, and the range, or difference, between the maximum and minimum is 70.

In a cumulative-frequency graph, such as Fig. 1, the grades are marked on the horizontal axis and double marked on the vertical axis with the cumulative number of the grades on the left and the corresponding percentage of the total number on the right. Each dot represents the accumulated number of students who have attained a particular grade or less. For example, the dot A corresponds to the second 72; reading on the vertical axis, it is evident that there are 12, or 40 percent, of the grades equal to or less than 72.

In analyzing the grades received by 10 sections of 30 pupils each on four examinations, a total of 1200 grades, the amount of data is too large to be exhibited conveniently as in Fig. 1. The statistician separates the data into suitably chosen groups, or intervals. For example, ten intervals might be used to tabulate the 1200 grades, as in column (a) of the accompanying frequency-distribution table; the actual number in an interval, called the frequency of the interval, is entered in column (c). The numbers that define the interval range are called the interval boundaries. It is convenient to choose the interval boundaries so that the interval ranges are equal to each other; the interval midpoints, half the sum of the interval boundaries, are simple numbers, because they are used in many calculations. A grade such as 87 will be tallied in the 80-90 interval; a boundary grade such as 90 may be tallied uniformly throughout the groups in either the lower or upper intervals. The relative frequency, column (d), is the ratio of the frequency of an interval to the total count; the relative frequency is multiplied by 100 to obtain the percent relative frequency. The cumulative frequency, column (e), represents the number of students receiving grades equal to or less than the range in each succeeding interval; thus, the number of students with grades of 30 or less is obtained by adding the frequencies in column (c) for the first three intervals, which total 53. The cumulative relative frequency, column (f), is the ratio of the cumulative frequency to the total number of grades.

The data of a frequency-distribution table can be presented graphically in a frequency histogram, as in Fig. 2, or a cumulative-frequency polygon, as in Fig. 3. The histogram is a series of rectangles with bases equal to the interval ranges and areas proportional to the frequencies. The polygon in Fig. 3 is drawn by connecting with straight lines the interval midpoints of a cumulative frequency histogram.

Newspapers and other printed media frequently present statistical data pictorially by using different lengths or sizes of various symbols to indicate different values.

D.5 Measures of central tendency

After data have been collected and tabulated, analysis begins with the calculation of a single number, which will summarize or represent all the data. Because data often exhibit a cluster or central point, this number is called a measure of central tendency.

Let $x_1, x_2, \ldots, x_n$ be the n tabulated (but ungrouped) numbers of some statistic; the most frequently used measure is the simple arithmetic average, or mean, written $\overline{x}$, which is the sum of the numbers divided by n:

$$\overline{x} = \frac{\sum x}{n}$$

If the x's are grouped into k intervals, with midpoints $m_1, m_2, \ldots, m_k$ and frequencies $f_1, f_2, \ldots, f_k$, respectively, the simple arithmetic average is given by

$$\frac{\sum f_1 m_1}{\sum f_1} \qquad \text{with } i = 1, 2, \ldots, k.$$

The median and the mode are two other measures of central tendency. Let the x's be arranged in numerical order; if n is odd, the median is the middle x; if n is even, the median is the average of the two middle x's. The mode is the x that occurs most frequently. If two or more distinct x's occur with equal frequencies, but none with greater frequency, the set of x's may be said not to have a mode or to be bimodal, with modes at the two most frequent x's, or trimodal, with modes at the three most frequent x's.

D.6 Measures of variability

The investigator frequently is concerned with the variability of the distribution, that is, whether the measurements are clustered tightly around the mean or spread over the range. One measure of this variability is the difference between two percentiles, usually the 25th and the 75th percentiles. The pth percentile is a number such that p percent of the measurements are less than or equal to it; in particular, the 25th and the 75th percentiles are called the lower and upper quartiles, respectively. The pth percentile is readily found from the cumulative-frequency graph, (Fig. 1) by running a horizontal line through the p percent mark on the vertical axis on the graph, then a vertical line from this point on the graph to the horizontal axis; the abscissa of the intersection is the value of the pth percentile.

The standard deviation is a measure of variability that is more convenient than percentile differences for further investigation and analysis of statistical data. The standard deviation of a set of measurements $x_1, x_2, \ldots, x_n$, with the mean $\overline{x}$ is defined as the square root of the mean of the squares of the deviations; it is usually designated by the Greek letter sigma (σ). In symbols

$$\sigma = \sqrt{\frac{1}{n}[(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2]} = \sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2}$$

The square, σ^2, of the standard deviation is called the variance. If the standard deviation is small, the measurements are tightly clustered around the mean; if it is large, they are widely scattered.

D.7 Correlation

When two social, physical, or biological phenomena increase or decrease proportionately and simultaneously because of identical external factors, the phenomena are correlated positively; under the same conditions, if one increases in the same proportion that the other decreases, the two phenomena are negatively correlated. Investigators calculate the degree of correlation by applying a coefficient of correlation to data concerning the two phenomena. The most common correlation coefficient is expressed as

$$\frac{\sum\left(\frac{x}{\sigma^x} \cdot \frac{y}{\sigma^y}\right)}{N}$$

in which x is the deviation of one variable from its mean, y is the deviation of the other variable from its mean, and N is the total number of cases in the series. A perfect positive correlation between the two variables results in a coefficient of +1, a perfect negative correlation in a coefficient of -1, and a total absence of correlation in a coefficient of 0. Intermediate values between +1 and 0 or -1 are interpreted by degree of correlation. Thus, .89 indicates high positive correlation, -.76 high negative correlation, and .13 low positive correlation.

D.8 Mathematical models

A mathematical model is a mathematical idealization in the form of a system, proposition, formula, or equation of a physical, biological, or social phenomenon. Thus, a theoretical, perfectly balanced die that can be tossed in a purely random fashion is a mathematical model for an actual physical die. The probability that in n throws of a mathematical die a throw of 6 will occur k times is

$$p(k) = \binom{n}{k}\left(\frac{1}{6}\right)^n (\frac{5}{6})^{n-k}$$

in which $\binom{n}{k}$ is the symbol for the binomial coefficient

$$\frac{n(n-1)\cdots(n-k+1)}{1 \cdot 2 \cdot \cdots \cdot k} \cdot \left(\binom{n}{0} = 1\right)$$

The statistician confronted with a real physical die will devise an experiment, such as tossing the die n times repeatedly, for a total of Nn tosses, and then determine from the observed throws the likelihood that the die is balanced and that it was thrown in a random way.

In a related but more involved example of a mathematical model, many sets of measurements have been found to have the same type of frequency distribution. For example, let $x_1, x_2, \ldots, x_N$ be the number of 6's cast in the N respective runs of n tosses of a die and assume N to be moderately large. Let $y_1, y_2, \ldots, y_N$ be the weights, correct to the nearest 1/100 g, of N lima beans chosen haphazardly from a 100-kg bag of lima beans. Let $z_1, z_2, \ldots, z_N$ be the barometric pressures recorded to the nearest 1/1000 cm by N students in succession, reading the same barometer. It will be observed that the x's, y's, and z's have amazingly similar frequency patterns. The statistician adopts a model that is a mathematical prototype or idealization of all these patterns or distributions. One form of the mathematical model is an equation for the frequency distribution, in which N is assumed to be infinite:

$$y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

in which e (approximately 2.7) is the base for natural logarithms (see Logarithm). The graph of this equation (Fig. 4) is the bell-shaped curve called the normal, or Gaussian, probability curve. If a variate x is normally distributed, the probability that its value lies between a and b is given by

$$\frac{1}{\sqrt{2\pi}} \int_d^b e^{(-(x^2)/2)} dx$$

The mean of the x's is 0, and the standard deviation is 1. In practice, if N is large, the error is exceedingly small.

D.9 Tests of reliability

The statistician is often called upon to decide whether an assumed hypothesis for some phenomenon is valid or not. The assumed hypothesis leads to a mathematical model; the model, in turn, yields certain predicted or expected values, for example, 10, 15, 25. The corresponding actually observed values are 12, 16, 21. To determine whether the hypothesis is to be kept or rejected, these deviations must be judged as normal fluctuations caused by sampling techniques or as significant discrepancies. Statisticians have devised several tests for the significance or reliability of data. One is the chi-square (χ^2) test. The deviations (observed values minus expected values) are squared, divided by the expected values, and summed:

$$x^2 = \frac{(12-10)^2}{10} + \frac{(16-15)^2}{15} + \frac{(21-25)^2}{25} = 1.11$$

The value of χ^2 is then compared with values in a statistical table to determine the significance of the deviations.

D.10 Higher statistics

The statistical methods described above are the simpler, more commonly used methods in the physical, biological, and social sciences. More advanced methods, often involving advanced mathematics, are used in further statistical studies, such as sampling theory, inference and estimation theory, and design of experiments.

Contributed By: James Singer

D.11 Difference between statistics and probability

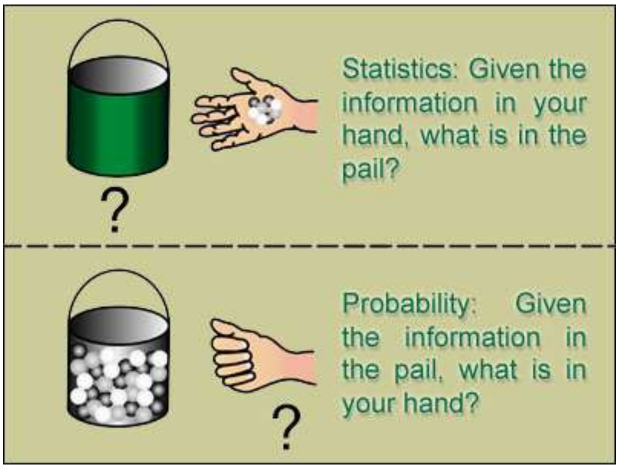

Statistics: Given the information in your hand, what is in the pail?
?
Probability: Given the information in the pail, what is in your hand?
?

Figure D.1: Diagram showing the difference between statistics and probability

(Image by MIT OpenCourseWare. Based on Gilbert, Norma. Statistics. W.B. Saunders Co., 1976.)

Index