Lecture Notes on Statistics I

Shiu-Sheng Chen[†]

First Version: August 2004
This Version: September 2007

 Department of Economics, National Taiwan University. email: sschen@ntu.edu.tw

Acknowledgement

Parts of this lecture note use the course materials from **?**). I thank Professor William H. Sandholm for allowing me to use his materials. Any remaining errors are my own responsibility.

Part I

Introduction, Data and Statistical Measures

Chapter 1

Introduction

What is this course about?

Probability The mathematics of uncertainty. An interesting and very useful area of mathematics. Useful for reasoning whenever you are uncertain about future events-which is basically all the time.

ex: Investment Decisions All sorts of ways to invest many stocks, bonds, mutual funds. Many of these investments are **risky**: their outcomes are uncertain. How should we invest our money to keep **risks low** while maintaining a high probability of a **strong return**? The key to understanding investment decisions is probabilistic reasoning.

What do we know about probability? Consider following example:

ex: The Doctor's Problem Suppose one in a thousand students in NTU is afflicted with lovesickness. A new test has been proposed to detect this disease. Experience with the test has shown that of in **99%** of the case in which the patient has had lovesickness, the test has turned out **positive**. Moreover, in **95%** of cases in which the patient has NOT had lovesickness, the test result has been **negative**. During a routine physical, a doctor finds that his patient has tested positive for lovesickness. Based on the information above, approximately what probability should the doctor assign to the patient having lovesickness?

(A) Above 99%
(B) Between 95% and 99%
(C) Between 50% and 95%
(D) Less than 50%

What is your answer? Most people fail to figure out the correct answer. It indicates that we do not seem to be naturally well equip to reason probabilistically-many common intuitions turn out to be wrong.

So for the first half of the course we will learn the tools of probability theory so that we can reason about uncertainty correctly. To help intuition, we will consider a number of applications, especially in finance.

Statistical Inference The process of drawing conclusions from random samples. Ex: In an exit of supermarket, you ask 100 randomly chosen shoppers which cola they bought, Coca or Pepsi? Suppose 54 say Coca. How strong is this as evidence that more people prefer Coca rather than Pepsi? The key to answer this question is to use the fact that we have drawn a random sample of shoppers. Probability theory has a lot to say about how random samples behave. By using probability theory, we can draw conclusions about the shoppers' preference on the basis of the random sample.

1.1 Data and Statistical Measures

1.1.1 Data

Types of Data:

1. Qualitative Data: nonnumeric variables and can't be measured. For example, gender, state of birth, education.

2. Quantitative Data: numerical variables and can be measured. For example, number of children in your family, consumer price index, gross domestic product (GDP).

Types of Quantitative Data:

1. Discrete Data: only certain values, and there are usually "gaps" between the values, such as the number of children in your family

2. Continuous Data: any value within a specific range, such as the air pressure in a tire.

1.1.2 Descriptive Measures (Descriptive Statistics)

Data Organizing and Presenting

- **Frequency Distribution:** The easiest method of organizing data is a frequency distribution, which converts raw data into a meaningful pattern for statistical analysis. The following are the steps of constructing a frequency distribution:

    1. Specify the number of class intervals. A class is a group (category) of interest. No totally accepted rule tells us how many intervals are to be used. Between 5 and 15 class intervals are generally recommended. Note that the classes must be both mutually exclusive and all-inclusive. Mutually exclusive means that classes must be selected such that an item can not fall into two classes, and all-inclusive classes are classes that together contain all the data.

    2. When all intervals are to be the same width, the following rule may be used to find the required class interval width:

$$\text{class width} = \frac{(\text{the largest data} - \text{the smallest data})}{\text{number of classes}}$$

- **Frequency Plot:** Graphs, curves, and charts are used to present data. ex: Histograms.

Measures of Locations

- **Mean:** The mean is computed by summing all numbers and dividing by the number of observations.

$$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}$$

The mean uses all the observations and each observation affects the mean. Even though the mean is sensitive to extreme values (i.e., extremely large or small data can cause the mean to be pulled toward the extreme data) it is still the most widely used measure of location. This is due to the fact that the mean has valuable mathematical properties that make it convenient for use with inferential statistics analysis. For example, the sum of the deviations of the numbers in a set of data from the mean is zero, and the sum of the squared deviations of the numbers in a set of data from the mean is minimum value

- **Median:** The median is the middle value in an ordered array of observations. If there is an even number of data in the array, the median is the average of the two middle numbers. If there is an odd number of data in the array, the median is the middle number.

$$\text{Median item number} = \frac{N+1}{2}$$

- **Mode:** The mode is the most frequently occurring value in a set of observation. For example, given 2, 3, 4, 5, 4, the mode is 4, because there are more fours than any other number. Data may have two modes. In this case we say the data are bimodal, and observations with more than two modes are referred to as multi-modal. Note that the mode does not have important mathematical properties for future use. Also, the mode is not a helpful measure of location, because there can be more than one mode or even no mode.

Measures of Variability

- **Range:** The range is the difference between the largest observation of a data set and the smallest observation. The major disadvantage of the range is that it does not include all of the observations. Only the two most extreme values are included and these two numbers may be untypical observations. For example, given that the ages for a sample of 8 students at CSC are: 24, 18, 22, 19, 25, 20, 23, and 21, the range for this data set is: 25 - 18 = 7.

- **Variance:** An important measure of variability is variance. Variance is the average of the squared deviations from the arithmetic mean

$$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N}$$

- **Standard Deviation:** The standard deviation is the square root of the variance. Both variance and standard deviation provide the same information; one can always be obtained from the other. In other words, the process of computing a standard deviation always involves computing a variance.

Other Important Measures

- **Skewness:** Skewness is a measure of asymmetry of the distribution of the series around its mean. Skewness is computed as:

$$S = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_x}{\sigma_x}\right)^3$$

  The skewness of a symmetric distribution, such as the normal distribution, is zero. Positive skewness means that the distribution has a long right tail and negative skewness implies that the distribution has a long left tail.

- **Kurtosis:** Kurtosis measures the peakedness or flatness of the distribution of the data. Kurtosis is computed as

$$K = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i - \mu_x}{\sigma_x}\right)^4$$

  The kurtosis of the normal distribution is 3. If the kurtosis exceeds 3, the distribution is peaked (leptokurtic) relative to the normal; if the kurtosis is less than 3, the distribution is flat (platykurtic) relative to the normal.

Remarks: It is now easier to carry out the descriptive measures of the data using computer software. For instance, Microsoft EXCEL provides **Data Analysis** option under **Tools**.

An Example: Consider the following midterm grades from an Econ 101 course:
73.26 76.59 76.59 86.58 93.24 76.59 86.58 83.25 86.58 73.26 53.28 86.58 99.9 86.58 93.24 89.91 86.58 73.26 59.94 93.24 79.92 99.9 59.94 86.58 79.92 73.26 69.93 73.26 56.61 79.92 63.27 89.91 76.59 89.91 93.24 96.57 66.6 69.93 56.61 69.93 63.27 73.26 89.91 83.25 96.57 99.9 83.25 96.57 86.58 86.58 93.24 83.25 66.6 86.58 63.27 83.25 53.28 96.57 89.91 86.58 96.57 49.95 93.24 69.93 83.25 66.6 79.92 96.57 76.59 66.6 73.26 83.25 89.91 93.24 86.58 69.93 59.94 79.92 86.58 83.25 83.25 99.9 56.61 69.93 99.9 76.59 73.26 46.62 93.24 66.6 73.26 79.92 83.25 83.25 93.24 79.92 86.58 76.59 86.58 83.25 69.93 99.9

Table 1.1: Frequency Distribution

Bin	Frequency
(0,50]	2
(50,55]	2
(55,60]	6
(60,65]	3
(65,70]	12
(70,75]	9
(75,80]	14
(80,85]	12
(85,90]	21
(90,95]	9
(95,100]	12

People may be curious about why the Kurtosis reported here is negative (-0.33)? Since most people are interested in the peakedness or flatness **relative to Normal Distribution**, some computer software would report $\tilde{K} = K - 3$.

Figure 1.1: Econ 101 Midterm Grades

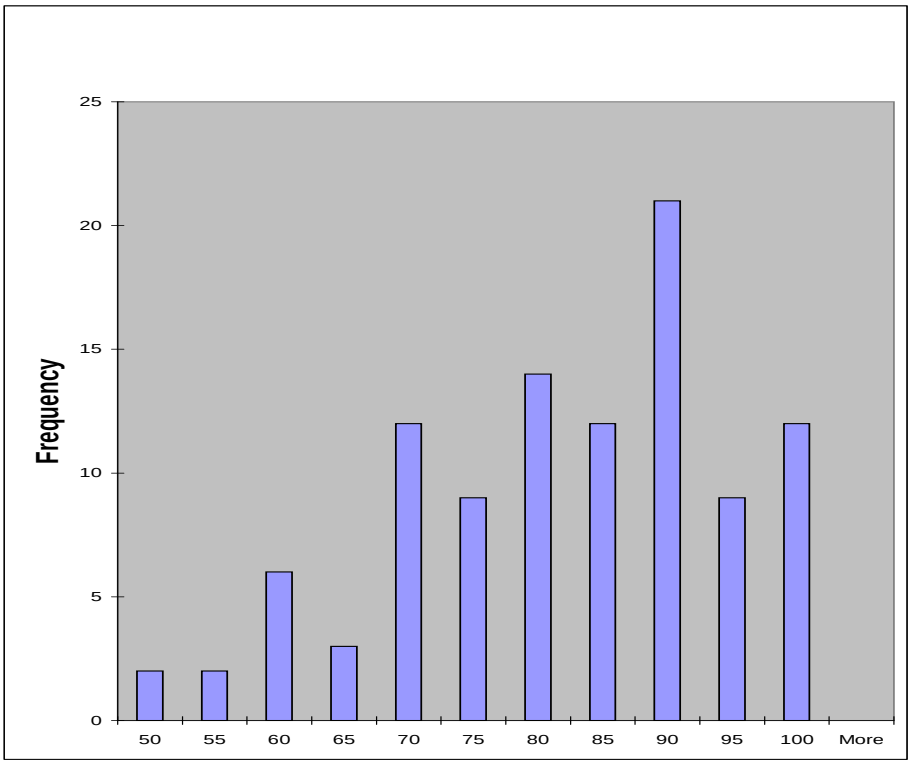


Table 1.2: Descriptive Statistics

Statistics		Statistics	
Mean	80.15	Skewness	-0.51
Median	83.25	Range	53.28
Mode	86.58	Minimum	46.62
Standard Deviation	12.66	Maximum	99.9
Variance	160.35	Sum	8175.15
Kurtosis	-0.33	Count	102

Part II

Probability Theory

Chapter 2

Set Theory and Probability Models

2.1 Set Theory

Set A set is a collection of elements. ex: A={red, blue, green}. Moreover, B is a **subset** of A if every element of B is also in A. ex: B={red, blue}. Notation: $B \subseteq A$ (B is a subset of A) or $A \supseteq B$ (A contains B).

In probability theory, set theory is used to describe possible outcomes. The **state space** S, is a set containing all **states** (possible outcomes). Exactly one state will occur. Each subset of S is called an **event**.

Set Theory	Probability Theory
set	state space
elements	states, possible outcomes
subsects	events

ex: Die Roll $S = \{1, 2, 3, 4, 5, 6\}$, $F = \{1, 2, 3\}$, $E = \{2, 4, 6\}$, $O = \{1, 3, 5\}$.

Four Operations

Name	Notation	Example
Complement	A^c	$F^c = \{4, 5, 6\}$
Union	$A \cup B$	$F \cup E = \{1, 2, 3, 4, 6\}$
Intersection	$A \cap B$	$F \cap E = \{2\}$
Relative Complement	$A - B = A \cap B^c$	$F - E = \{1, 3\}$

Venn Diagram A Venn Diagram is a way to graphically represent sets and set operations. Each diagram begins with a rectangle representing the state space (universal set). Then each set is represented by a circle.

Empty Set The empty set is the set containing no elements. We denote it as $\emptyset$. (Do not misread it as the Greek Alphabet ϕ. It is called "empty set").

Disjoint (Mutually Exclusive) Two sets, A and B, are said to be disjoint (mutually exclusive) if their intersection is empty:

$$A \cap B = \emptyset$$

Partition $A_1, \ldots, A_n \subseteq S$ is said to be a partition of S if

1. $A_i \cap A_j = \emptyset$ for $i \neq j$ (disjoint)
2. $A_1 \cup A_2 \cup \cdots \cup A_n = S$

Example Suppose $S = \{1, 2, 3, 4, 5, 6\}$. Consider the following subsets of S: A_1, A_2 and A_3.

1. $A_1 = \{1, 2\}$, $A_2 = \{3, 6\}$, $A_3 = \{4, 5\}$
2. $A_1 = \{1, 2\}$, $A_2 = \{5\}$, $A_3 = \{4, 6\}$
3. $A_1 = \{1, 2\}$, $A_2 = \{3, 4\}$, $A_3 = \{2, 5, 6\}$

Check if A_1, A_2 and A_3 form a partition of S for each case.

A List of Connections between the Set Operations

1. Complementation:
$$(A^c)^c = A, \quad \emptyset^c = S, \quad S^c = \emptyset$$
2. Commutativity of set union and intersection:
$$A \cup B = B \cup A, \quad A \cap B = B \cap A$$

3. Associativity of union and intersection:

$$(A \cup B) \cup C = A \cup (B \cup C), \quad (A \cap B) \cap C = A \cap (B \cap C)$$

4. *De Morgan's laws*:

$$(A_1 \cup A_2 \cup \cdots \cup A_n)^c = A_1^c \cap A_2^c \cap \cdots \cap A_n^c$$

$$(A_1 \cap A_2 \cap \cdots \cap A_n)^c = A_1^c \cup A_2^c \cup \cdots \cup A_n^c$$

(i) $(A \cap B)^c \subseteq (A^c \cup B^c)$

Proof: (by contradiction) Assume: $(A \cap B)^c \not\subseteq (A^c \cup B^c)$ $\therefore \exists x$ such that $x \in (A \cap B)^c$, $x \notin (A^c \cup B^c)[x \notin (A^c \cup B^c)] \Longrightarrow (x \notin A^c \wedge x \notin B^c) \Longrightarrow (x \in A \wedge x \in B) \Longrightarrow [x \in (A \cap B)] \Longrightarrow [x \notin (A \cap B)^c] \oplus$ (contradiction) $\therefore (A \cap B)^c \subseteq (A^c \cup B^c)$ ... (1)

(ii) $(A \cap B)^c \supseteq (A^c \cup B^c)$

Proof: (by contradiction) Assume: $(A^c \cup B^c) \not\subseteq (A \cap B)^c$ $\therefore \exists x$ such that $x \notin (A \cap B)^c$, $x \in (A^c \cup B^c)[x \notin (A \cap B)^c] \Longrightarrow [x \in (A \cap B)] \Longrightarrow (x \in A \wedge x \in B) \Longrightarrow [x \notin A^c \wedge x \notin B^c)] \Longrightarrow [x \notin (A^c \cup B^c)] \oplus$ (contradiction) $\therefore (A^c \cup B^c) \subseteq (A \cap B)^c$ ... (2)

By (1) and (2), $(A^c \cup B^c) = (A \cap B)^c$

5. Distributivity laws:

$$B \cap (A_1 \cup A_2 \cup \cdots \cup A_n) = (B \cap A_1) \cup (B \cap A_2) \cup \cdots \cup (B \cap A_n)$$

$$B \cup (A_1 \cap A_2 \cap \cdots \cap A_n) = (B \cup A_1) \cap (B \cup A_2) \cap \cdots \cap (B \cup A_n)$$

6. As a consequence of the definitions:

$$A \cup A = A, \qquad A \cap A = A,$$
$$A \cup \emptyset = A, \qquad A \cap \emptyset = \emptyset,$$
$$A \cup S = S, \qquad A \cap S = A,$$
$$A \cup A^c = S, \qquad A \cap A^c = \emptyset.$$

7. Some other important properties:

(I1)

$$A = (A \cap B) \cup (A \cap B^c) \tag{2.1}$$

proof.

$$\begin{aligned} A &= A \cap S \\ &= A \cap (B \cup B^c) \\ &= (A \cap B) \cup (A \cap B^c) \end{aligned}$$

(I2)

$$(A \cap B) \quad \text{and} \quad (A \cap B^c) \quad \text{are disjoint.} \tag{2.2}$$

proof.

$$\begin{aligned} (A \cap B) \cap (A \cap B^c) &= A \cap (B \cap B^c) \\ &= A \cap \emptyset \\ &= \emptyset \end{aligned}$$

(I3)

$$A \cup B = A \cup (A^c \cap B)$$

proof.

$$A \cup B = (A \cup B) \cap S = (A \cup B) \cap (A \cup A^c) = A \cup (A^c \cap B)$$

2.2 Probability Models

Probability Measure P is a probability measure on the state space S if all events $A \subseteq S$ are assigned numbers $P(A)$ satisfying

(a) $P(\emptyset) = 0$

(b) $P(S) = 1$

(c) $P(A) \geq 0$ for all $A \subseteq S$

(d) $P(A \cup B) = P(A) + P(B)$ for all disjoint $A, B \subseteq S$

The pair (S, P) is called a probability model.[1] [2]

ex: Die Roll $P(\{1\}) = P(\{2\}) = \cdots = P(\{6\}) = 1/6$. We can then use (d) to assign probabilities to other events. For instance,

1. $P(\{1\} \cup \{2\}) = P(\{1\}) + P(\{2\}) = 1/6 + 1/6 = 1/3$.

2.

$$\begin{aligned} P(\{\leq 2\} \cup \{\geq 4\}) &= P(\{\leq 2\}) + P(\{\geq 4\}) \\ &= P(\{1\} \cup \{2\}) + P(\{4\} \cup \{5\} \cup \{6\}) \\ &= P(\{1\}) + P(\{2\}) + P(\{4\}) + P(\{5\}) + P(\{6\}) \\ &= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 5/6 \end{aligned}$$

Other properties of probability models are implications of (a)-(d):

(p1) $P(A) + P(A^c) = 1$

(p2) $A \subseteq B$ implies that $P(A) \leq P(B)$

(p3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (law of addition)

Exercise 1. Prove (p1), (p2) and (p3) according to (a)-(d).

Anything we would want to call a probability measure should satisfy (a)-(d). This includes both "objective probability" (ex: die roll) and "subjective probability" (ex: your beliefs about baseball game). This distinction (objective v.s. subjective) will become important when we talk about statistics.

[1] This axiom approach to constructing the probability model was done by the famous Russian mathematician A.N. Kolmogorov (1903-1987) in his fundamental monograph *Grundbegriffe der Wahrscheinlichkeitsrechnung (Fundamental Ideas of the Probability Calculation)*, which appeared in 1933.

[2] A more thorough discussion of probability theory is based on measure theory by introducing the *probability space* $(S, \mathcal{F}, P)$. Where, $\mathcal{F}$ is the collection ($\sigma-$algebra) of measurable subsets of S. It is, however, beyond the scope of this course.

Likelihood Judgement Problem Linda is 31 years old, single, outspoken, and very bright. she majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and she participated in antinuclear demonstrations.

Rank order the following eight descriptions in terms of the probability that they describe Linda. write "1" in front of the most likely description, "2" in front of the next most likely, etc.

—— A. Linda is a teacher in an elementary school.

—— B. Linda works in a bookstore and takes yoga classes.

—— C. Linda is active in the feminist movement.

—— D. Linda is a psychiatric social worker.

—— E. Linda is a bank teller.

—— F. Linda is a member of the American Philosophical Association

—— G. Linda is an insurance salesperson.

—— H. Linda is a bank teller who is active in the feminist movement.

2.3 Conditional Probability

How should probability assessments change in the face of new information? The conditional probability $P(A|B)$ means "the probability of event A given (conditional on) event B". We can think of $P(\cdot|B)$ as a new probability measure on S.

ex: Die Roll $P(\{1\}|\text{Odd}) = P(\{1\}|\{1,3,5\}) = 1/3$.

We would like the conditional probability to satisfy two properties:

(C1) $P(B|B) = 1$

(C2) If $C, D \subseteq B$ and $P(D) \neq 0$, then

$$\frac{P(C|B)}{P(D|B)} = \frac{P(C)}{P(D)}$$

i.e. fixed relative probability for subsets of B. If D was twice as likely as C, it remains twice as likely as C while given B.

Example

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$B = \{\leq 5\} = \{1, 2, 3, 4, 5\}$$

$$C = \{1, 2\}$$

$$D = \{2, 3, 4\}$$

$$P(C) = 1/3 \qquad P(D) = 1/2$$

$$P(C|B) = 2/5 \qquad P(D|B) = 3/5$$

Definition (Conditional Probability)

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \text{ whenever} P(B) \neq 0.$$

Or

$$P(A \cap B) = P(A|B) \times P(B)$$

Therefore, according to the definition,

$$P(B|B) = \frac{P(B \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$$

$$\frac{P(C|B)}{P(D|B)} = \frac{\frac{P(C \cap B)}{P(B)}}{\frac{P(D \cap B)}{P(B)}} = \frac{P(C)}{P(D)}$$

In fact, the definition above is the only one which satisfy (C1) and (C2).

ex: Die Roll

$$P(\{1\}|\text{Odd}) = P(\{1\}|\{1,3,5\}) = \frac{P(\{1\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{P(\{1\})}{P(\{1,3,5\})} = \frac{1/6}{1/2} = 1/3$$

$$P(\text{Odd}|\text{Lowest 3}) = P(\{1,3,5\}|\{1,2,3\}) = \frac{P(\{1,3,5\} \cap \{1,2,3\})}{P(\{1,2,3\})} = \frac{P(\{1,3\})}{P(\{1,2,3\})} = \frac{1/3}{1/2} = 2/3$$

2.4 Conditional and Marginal Probabilities

Senate vote on a bill to remove financial regulations

	Y	N	
Democrat	3	48	51
Republican	35	14	49
	38	62	100

Suppose we select a senator at random.

1. divide by 100 to get probabilities
2. sum to get marginal probabilities: $P(D)$, $P(R)$, $P(Y)$, and $P(N)$.
3. after obtaining marginal probabilities, we can compute conditional probabilities. For instance,
   - $P(N|D) = \frac{P(N \cap D)}{P(D)} = \frac{0.48}{0.51} \approx 0.9412$
   - $P(R|Y) = \frac{P(R \cap Y)}{P(Y)} = \frac{0.35}{0.38} \approx 0.9211$

A General Representation: Let $A_1, \ldots, A_n \subseteq S_A$ be a partition of S_A, and $B_1, \ldots, B_n \subseteq S_B$ be a partition of S_B.

	B_1	$\cdots$	B_n	
A_1	$P(A_1 \cap B_1)$	$\cdots$	$P(A_1 \cap B_n)$	$P(A_1)$
$\vdots$	$\vdots$	$\ddots$	$\cdots$	$\vdots$
A_n	$P(A_n \cap B_1)$	$\cdots$	$P(A_n \cap B_n)$	$P(A_n)$
	$P(B_1)$	$\cdots$	$P(B_n)$	1

2.5 The Doctor's Problem Revisited

Let

- T={randomly selected person tests positive}
- D={randomly selected person has lovesickness}
- D^c={randomly selected person does not have lovesickness}

According to the information we have

- 1 in 1000 have disease $\Rightarrow P(D) = 0.001$
- Prob(test positive|disease) $= 0.99 \Rightarrow P(T|D) = 0.99$
- Prob(test negative|no disease) $= 0.95 \Rightarrow P(T^c|D^c) = 0.95$

We can further get

- $P(D^c) = 0.999$
- $P(T|D^c) = 0.05$

using the fact that

$$P(D) + P(D^c) = 1,$$

and

$$P(T|D^c) + P(T^c|D^c) = 1.$$

Here is the proof of the second equation:

$$\begin{aligned} P(T|D^c) + P(T^c|D^c) &= \frac{P(T \cap D^c)}{P(D^c)} + \frac{P(T^c \cap D^c)}{P(D^c)} \\ &= \frac{1}{P(D^c)}[P(T \cap D^c) + P(T^c \cap D^c)] \\ &= \frac{1}{P(D^c)}[P(D^c)] \quad \text{by equations (??) and (??)} \\ &= 1 \end{aligned}$$

What we want to know is

$$\text{Prob(disease|test positive)} = P(D|T)$$

First note that:

$$P(T) = P(T \cap D) + P(T \cap D^c)$$

by equations (**??**) and (**??**).

$$\begin{aligned} P(D|T) &= \frac{P(D \cap T)}{P(T)} \\ &= \frac{P(D \cap T)}{P(D \cap T) + P(D^c \cap T)} \\ &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|D^c)P(D^c)} \end{aligned} \tag{2.3}$$

Thus,

$$P(D|T) = \frac{(0.99) \cdot (0.001)}{(0.99) \cdot (0.001) + (0.05) \cdot (0.999)} = \frac{0.00099}{0.00099 + 0.09995} = \frac{0.00099}{0.05094} = 0.0194$$

Remarks: Equation (**??**) is a special case of *Bayes' Rule (Bayes' Formula).*

Bayes' Rule Let $T \subseteq S$ be a test event. Let $A_1, A_2, \ldots A_n \subseteq S$ be a group of categories. Suppose

- $P(T) > 0$ (the positive tests can happen)
- $A_1, A_2, \ldots, A_n$ be a partition of S (they are disjoint, $A_1 \cup A_2 \cup \cdots A_n = S$)

And suppose we know

- $P(A_1), P(A_2), \ldots, P(A_n)$ (the probability of each category)
- $P(T|A_1), P(T|A_2), \ldots, P(T|A_n)$ (how well each category performs on the test)

We want to know

- $P(A_i|T)$ how likely category i is if the test comes up positive)

Bayes' Rule:

$$P(A_i|T) = \frac{P(A_i \cap T)}{P(T)} = \frac{P(T|A_i)P(A_i)}{\sum_{j=1}^{n} P(T|A_j)P(A_j)}$$

Where

$$P(T) = \sum_{j=1}^{n} P(T|A_j)P(A_j)$$

is called **Law of Total Probability**.

Proof: The Bayes' Rule

Example A softball team has three pitchers, X and Y, with winning percentages of 0.4 and 0.8, respectively. They pitch with frequency 3, and 7 out of every 10 games, respectively. In other words, for a randomly selected game, $P(X) = 0.3$ and $P(Y) = 0.7$. Find:

1. P(team wins game)=$P(W)$.
2. P(X pitched game|team won)=$P(X|W)$.

$$P(W) = P(W|X)P(X) + P(W|Y)P(Y) = (0.4) \cdot (0.3) + (0.8) \cdot (0.7) = 0.68$$

$$P(X|W) = \frac{P(X \cap W)}{P(W)} = \frac{P(W|X)P(X)}{P(W)} = 0.12/0.68 = 3/17$$

2.6 Independence

Intuition Events are independent if the occurrence of some of them does not provide information about the occurrence of the others.

Definition Two events $A, B \subseteq S$ are <u>independent</u> if $P(A \cap B) = P(A)P(B)$.

Remarks

- If $P(B) \neq 0$, $P(A|B) = \frac{A \cap B}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$.
- As well, if $P(A) \neq 0$, $P(B|A) = P(B)$.

$\Longrightarrow$ Finding out that one event occurred does not change your assessment of the likelihood of the other event.

An Example of Independent Events Tossing a fair coin twice:

	Head	Tail	
Head	1/4	1/4	1/2
Tail	1/4	1/4	1/2
	1/2	1/2	1

However, in the example of Senate Vote, the event that the senator we selected at random is Democrat, D and the event that the senator voted for financial deregulation, Y are NOT independent. (Please Check!)

In many real world and experimental situations, we have good reasons to believe certain events are independent. Thus, knowing the probabilities of each individual event allows us to calculate all sorts of joint probabilities. For example, suppose $P(A) = 0.2$, $P(B) = 0.5$. What is $P(A \cup B)$?

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) - P(A)P(B) \\ &= 0.2 + 0.5 - 0.1 = 0.6 \end{aligned} \tag{2.4}$$

Definition The event $A_1, A_2, \ldots, A_n$ are <u>independent</u> if for all $I \subseteq \{1, 2, \ldots, n\}$

$$P(\bigcap_{i \in I} A_i) = \prod_{i \in I} P(A_i) \tag{2.5}$$

Remarks

1. Why ALL subsets? Because otherwise, if one event had probability zero, automatically have independence.

2. If equation (**??**) holds for all pairs only, the events are called *pairwise independent.*

2.7 Monty Hall's Paradox

Preliminary Reading

> "Behind Monty Hall's Doors: Puzzle, Debate and Answer?" The New York Times, July 21, 1991. By John Tierney, Special to The New York Times.

Monty Hall's Paradox Imagine that the set of Monty Hall's game show Let's Make a Deal has three closed doors, Door A, Door B and Door C. Behind one of these doors is a car; behind the other two are goats. The contestant does not know where the car is, but Monty Hall does.

The contestant picks a door and Monty opens one of the remaining doors, one he knows doesn't hide the car. If the contestant has already chosen the correct door, Monty opens either of the two remaining doors.

After Monty has shown a goat behind the door that he opens, the contestant is always given the option to switch doors. What is the probability of winning the car if she stays with her first choice? What if she decides to switch? The contestant, of course, wants to maximizing his chance of winning the car.

An Intuitive Argument The a priori probability that the car is behind Door X, $P(X) = 1/3$ for X={A,B,C}. When you chose Door A, the probability that you chose the car was 1/3 and the probability that it was behind one of the other doors was 2/3. By showing you which of Doors B and C does not hide the car (Door B, say), the Monty is giving you quite a bit of information about those two doors. The probability is still 2/3 that one of them hides the car, but now you know which of the two it would be: Door

Figure 2.1: Monty Hall's Paradox

C. So, the probability is still only 1/3 that the car is behind Door A, but 2/3 that it is behind Door C. So if we stick on Door A, we have probability 1/3 to win the car but if we switch, we will win the car with probability 2/3. Check the following table:

	Car Behind		
	Door A	Door B	Door C
Stick	1	0	0
Switch	0	1	1

But The Paradox Comes From... Think about the following argument. After you named Door A, suppose Monty showed you Door B. According to Bayes' Rule, we can get

$$P(A|B^c) = \frac{P(A \cap B^c)}{P(B^c)} = \frac{P(A)}{P(B^c)} = \frac{1/3}{2/3} = 1/2$$

and

$$P(C|B^c) = \frac{P(C \cap B^c)}{P(B^c)} = \frac{P(C)}{P(B^c)} = \frac{1/3}{2/3} = 1/2$$

That means it doesn't matter if you switch or not. Notice that we use the fact that $A \cap B^c = A$ when A, B disjoint.[3]

[3]The proof is as follows:

$$\begin{aligned} A &= (A \cap B) \cup (A \cap B^c) \quad \text{by equation (??)} \\ &= \emptyset \cup (A \cap B^c) \quad \text{by construction} \\ &= A \cap B^c \end{aligned}$$

A Solution to the Paradox If all we were told was "there is a goat behind Door B", the argument mentioned above would be right. We may update the probability to 1/2 for both $P(A|B^c)$ and $P(C|B^c)$.

BUT, we are told that "Monty opens Door B". Recall the procedures which Monty must follow: Monty has to show you the door that is

- not the door you picked
- not the door with the car

Given the procedures, therefore, "Monty opens Door B" means that

- a goat is behind Door B
- something more: you rule out the case that the car is behind Door A but Monty shows Door C.

So assume that after we have chosen Door A, Monty opens Door B or Door C with **equal** probability if we were right.[4] (This assumption is important for the result.) Denote the event that Monty shows Door B as SB. Then we have

The probability that Monty Hall opens door B if the prize were behind A,

$$P(SB|A) = 1/2$$

The probability that Monty Hall opens door B if the prize were behind B,

$$P(SB|B) = 0$$

The probability that Monty Hall opens door B if the prize were behind C,

$$P(SB|C) = 1$$

The probability that Month Hall opens door B is then

$$\begin{aligned} P(SB) &= P(A)P(SB|A) + P(B)P(SB|B) + P(C)P(SB|C) \\ &= 1/6 + 0 + 1/3 = 1/2 \end{aligned}$$

[4]That is, the car is behind A.

By Beyes' Rule,

$$\begin{aligned} P(A|SB) &= \frac{P(A)P(SB|A)}{P(SB)} \\ &= \frac{(1/6)}{(1/2)} \\ &= 1/3 \end{aligned}$$

and

$$\begin{aligned} P(C|SB) &= \frac{P(C)P(SB|C)}{P(SB)} \\ &= \frac{(1/3)}{(1/2)} \\ &= 2/3 \end{aligned}$$

Therefore, "switch" is more likely to win the car.

		Monty Shows			
		Door A	Door B	Door C	
Car	Door A	0	1/6	1/6	1/3
Behind	Door B	0	0	1/3	1/3
	Door C	0	1/3	0	1/3
		0	1/2	1/2	

Remark This argument only works if Monty is guaranteed to show you a bad door every time after you choose a door, something that was not assured in the original game show.

Exercise 2. Change the assumption that Monty opens Door B or Door C with **equal** probability when we were right. Check if you obtain a different result?

Chapter 3

Random Variables

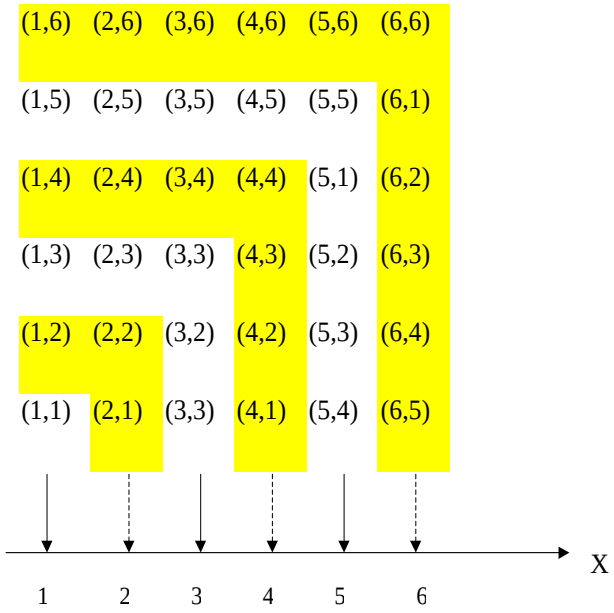
3.1 Random Variables

In general, one is not interested in events per se, but rather in some function of them. For instance, suppose one plays a game where the payoff depends on the number of dots on rolling a die twice. The score is the *maximum* of the two numbers that occur. Let $e = \{\text{first roll}, \text{second roll}\} = \{i, j\}$, and payoff $X(e) = max(i, j)$. That means if we get $e = 2, 3$, we will receive 3 units of money. Where, X is called a **random variable**.

Therefore, random variables are used to model uncertain numerical outcomes. We will talk about:

1. For 1 random variable
    - Full description: distribution
    - Average outcome: expected value
    - Dispersion of outcomes: variance, standard deviation
2. For 2 or more random variables
    - Full description: joint distribution
    - Comovement of outcomes: covariance, correlation
    - Generating new random variables out of old random variables (ex. stock returns $\Rightarrow$ portfolio returns)

Figure 3.1: Sample Space for Two Rolls of a Die



Definition (Random Variable) A random variable X is a function from the state space to the real numbers:[1]

$$X : S \longrightarrow \mathbb{R}$$

We call the numerical outcome that actually occurs the realization of the random variable, x. That is $X(e) = x$ with each possible outcome e in S.[2]

Idea X is used to describe the situation **before** the randomness is resolved. (ex ante) x is used **after** the randomness is resolved. (ex post)

Discrete Random Variable A random variable is discrete if the number of possible realizations is finite (or countable). For now, we only consider discrete random variables.

Distribution A single random variable is described by its distribution, which lists **ALL** of the possible realizations x and the probabilities of each, $P(X = x)$. It will be denoted

[1] Note that we denote random variables by capital letters.
[2] Note that we denote realizations by lowercase letters.

by $f(x)$ and called the **discrete probability density function** (discrete pdf). Another common terminology for $f(x)$ is **probability mass function** (pmf), and the possible realizations, x, are called **mass points** of X.

Suppose that X is a discrete random variable, taking values on some countable sample space $B \subseteq \mathbb{R}$. Then the probability mass function $f(x)$ for X is given by $f(x) : \mathbb{R} \mapsto [0, 1]$

$$f(x) = \begin{cases} P(X = x), & x \in B \\ 0, & x \in \mathbb{R} - B \end{cases} \tag{3.1}$$

Note that this explicitly defines $f(x)$ for all real numbers, including all values in $\mathbb{R}$ that X could never take; indeed, it assigns such values a probability of zero. (Alternatively, think of $P(X = x)$ as 0 when $x \in \mathbb{R} - B$.)

These probabilities must satisfy $\sum_{x \in B} P(X = x) = 1$. (why?)

Cumulative Distribution Function The cumulative distribution function (CDF) of a random variable X is defined for **any real** x by $F(x) : \mathbb{R} \mapsto [0, 1]$

$$F(x) = P(X \leq x)$$

- The function $F(x)$ often is referred to simply as the **distribution function** of X
- Let X be a random variable with pdf $f(x)$ and CDF $F(x)$. If the possible values of X are indexed in increasing order, $x_1 < x_2 < x_3 < \cdots$, then $f(x_1) = F(x_1)$, and for any $i > 1$,

  $$f(x_i) = F(x_i) - F(x_{i-1})$$

  Furthermore, if $x < x_1$ then $F(x) = 0$, and for any other real x

  $$F(x) = \sum_{x_i \leq x} f(x_i)$$

  where the summation is taken over all indices i such that $x_i \leq x$.
- Other properties

  $$\lim_{x \to -\infty} F(x) = 0 \tag{3.2}$$

$$\lim_{x\to\infty} F(x) = 1 \tag{3.3}$$

$$\lim_{h\to 0^+} F(x+h) = F(x) \tag{3.4}$$

$$a < b \quad \text{implies} \quad F(a) \le F(b) \tag{3.5}$$

$$P(a < X \le b) = F(b) - F(a) \tag{3.6}$$

The first two properties says that $F(x)$ can be made arbitrarily close to 0 or 1 by taking x arbitrarily (negatively or positively) large. Property (**??**) says that $F(x)$ is *continuous from the right*. Property (**??**) says that $F(x)$ is *nondecreasing*. For the last two properties, they follows from fact that

$$\{X \le b\} = \{X \le a\} \cup \{a < X \le b\}.$$

Clearly, $\{X \le a\}$ and $\{a < X \le b\}$ are disjoint. Thus, by definition, we have

$$\begin{aligned} F(b) &= F(a) + P(\{a < X \le b\}) \\ &\ge F(a), \end{aligned}$$

and

$$P(\{a < X \le b\}) = F(b) - F(a).$$

Remark In general, it is somewhat easier to understand the nature of a random variable and its probability distribution by considering the pdf directly, rather than the CDF, although the CDF will provide a good basis for defining *continuous* probability distribution. We will talk about this later.

ex. Die Roll: the discrete pdf, the CDF at its points of discontinuity, and the CDF

x	$f(x) = P(X = x)$	$F(x) = P(X \le x)$
1	1/6	1/6
2	1/6	2/6
3	1/6	3/6
4	1/6	4/6
5	1/6	5/6
6	1/6	1

Moreover, consider the CDF at every point rather than its points of discontinuity:

x	$P(X = x)$	$P(X \leq x)$
$x < 1$	0	0
$x = 1$	1/6	$P(X = 1) = 1/6$
$1 < x < 2$	0	$P(X = 1) = 1/6$
$x = 2$	1/6	$P(X = 1) + P(X = 2) = 2/6$
$2 < x < 3$	0	$P(X = 1) + P(X = 2) = 2/6$
$x = 3$	1/6	$P(X = 1) + \cdots + P(X = 3) = 3/6$
$3 < x < 4$	0	$P(X = 1) + \cdots + P(X = 3) = 3/6$
$x = 4$	1/6	$P(X = 1) + \cdots + P(X = 4) = 4/6$
$4 < x < 5$	0	$P(X = 1) + \cdots + P(X = 4) = 4/6$
$x = 5$	1/6	$P(X = 1) + \cdots + P(X = 5) = 5/6$
$5 < x < 6$	0	$P(X = 1) + \cdots + P(X = 5) = 5/6$
$x = 6$	1/6	$P(X = 1) + \cdots + P(X = 6) = 1$
$6 < x$	0	$P(X = 1) + \cdots + P(X = 6) = 1$

Thus the CDF is:

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < 1 \\ 1/6 & 1 \leq x < 2 \\ 2/6 & 2 \leq x < 3 \\ 3/6 & 3 \leq x < 4 \\ 4/6 & 4 \leq x < 5 \\ 5/6 & 5 \leq x < 6 \\ 1 & 6 \leq x \end{cases}$$

ex. Two Rolls of a Die: the discrete pdf and the CDF at its points of discontinuity Recall the example in Figure **??**. The discrete pdf of X, the maximum of two rolls of a die, is:

x	$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	1/36	1/36
2	3/36	4/36
3	5/36	9/36
4	7/36	16/36
5	9/36	25/36
6	11/36	1

$$F(x) = \begin{cases} 0 & x < 1 \\ 1/36 & 1 \leq x < 2 \\ 4/36 & 2 \leq x < 3 \\ 9/36 & 3 \leq x < 4 \\ 16/36 & 4 \leq x < 5 \\ 25/36 & 5 \leq x < 6 \\ 1 & 6 \leq x \end{cases}$$

Expected Value (Expectation, Mean) Expected value is the average realization.

$$\mu_X = E(X) = \sum_x xP(X = x)$$

Since $\sum_x P(X = x) = 1$, $E(X)$ is just a weighted average. It is the weighted average of the realizations, where the weights are the probabilities of each.

Variance Variance measures the dispersion of the realizations.

Q: How far do realizations typically deviate from the expectation?

A1: Mean Absolute Deviation

$$\sum_x |x - E(X)|P(X = x)$$

But...absolute value is difficult to work with because it is kinked.

A2: Variance

$$\sigma_X^2 = Var(X) = \sum_x (x - E(X))^2 P(X = x)$$

- the variance is the average squared deviation from the mean
- best thought of as a relative measure
- measured in squared units for original units, use standard deviation

$$\sigma_X = SD(X) = \sqrt{Var(X)}$$

ex. Die Roll

x	$P(X = x)$	$xP(X = x)$	$x - E(X)$	$(x - E(X))^2$	$(x - E(X))^2 P(X = x)$
1	1/6	1/6	-2.5	6.25	(6.25)/6
2	1/6	2/6	-1.5	2.25	(2.25)/6
3	1/6	3/6	-0.5	0.25	(0.25)/6
4	1/6	4/6	0.5	0.25	(0.25)/6
5	1/6	5/6	1.5	2.25	(2.25)/6
6	1/6	6/6	2.5	6.25	(6.25)/6
		$E(X) = 21/6 = 3.5$			$Var(X) = (17.5)/6 \approx 2.92$

Constants as Random Variables If c is a constant number, we can treat it like a random variable: $E(c) = c$, $Var(c) = 0$.

Moments

- r-th Moment

$$E(X^r) = \sum_x x^r P(X = x)$$

- r-th Central Moment

$$E(X - E(X))^r = \sum_x (x - E(X))^r P(X = x)$$

Remarks

1. $E(g(X)) = \sum_x g(x)P(X = x)$
2. $E(g(X)) \neq g(E(X))$ unless $g(\cdot)$ is a linear function. For instance,

$$E(\frac{1}{X}) \neq \frac{1}{E(X)}$$

Markov Inequality

$$P(|X| \geq k) \leq \frac{E(|X|^r)}{k^r} \tag{3.7}$$

for $r > 0$

Chebychev Inequality

$$P(|X - \mu| \geq k\sqrt{Var(X)}) \leq \frac{1}{k^2} \tag{3.8}$$

for $k > 0$

Or

$$P(|X - \mu| < k\sqrt{Var(X)}) > 1 - \frac{1}{k^2}$$

It means that with at least $(1-\frac{1}{k^2})$ chance that X locates within k standard deviations. For instance, if $k = 2$, with at least $(1 - \frac{1}{4}) = \frac{3}{4} = 75\%$ chance that X locates within 2 standard deviations.

Let $\varepsilon = k\sqrt{Var(X)}$, we can rewrite equation (**??**) as

$$P(|X - \mu| \geq \varepsilon) \leq \frac{Var(X)}{\varepsilon^2} \tag{3.9}$$

3.2 Multiple Random Variables

Joint Distributions Joint distributions specify the probability of every possible combination of outcomes.

ex. Joint Distribution of Stock Returns (I) Consider following joint distribution of the returns of Blue Chip stock (B)[3] and the returns of Technology stock (T):

Table 3.1: Joint Distribution of B and T

	Technology			
Blue Chip	30%	15%	0%	
20%	0.15	0.05	0	0.2
10%	0	0.3	0.3	0.6
5%	0.05	0.05	0.1	0.2
	0.2	0.4	0.4	

Marginal Distributions Marginal distributions are distributions of individual variables computed from joint distributions.

For instance,

$$\begin{aligned}\{B = 0.2\} &= \bigcup_{t=\{0.3,0.15,0\}} \{B = 0.2, T = t\} \\ &= \{B = 0.2, T = 0.3\} \cup \{B = 0.2, T = 0.15\} \cup \{B = 0.2, T = 0\}\end{aligned}$$

Since these events are disjoint,

$$P(B = 0.2) = P\left(\bigcup_t \{B = 0.2, T = t\}\right) = \sum_t P(B = 0.2, T = t) = 0.15 + 0.05 + 0 = 0.20$$

We can therefore obtain the marginal distribution of B and T:

Remark In this way we determine both marginal distributions for B and T. **BUT** the marginal distributions per se are NOT enough to determine the joint distribution. In particular, the marginal distributions do not tell us how the random variables are correlated.

[3]A "blue chip" is the nickname for a stock that is thought to be safe, in excellent financial shape and firmly entrenched as a leader in its field. Blue chips generally pay dividends and are favorably regarded by investors. A few examples of blue chips are Wal-Mart and Coca-Cola

Table 3.2: Discrete pdf (Marginal Distribution) of B

b	$f(b) = P(B = b)$
0.2	0.2
0.1	0.6
0.05	0.2

Table 3.3: Discrete pdf (Marginal Distribution) of T

t	$f(t) = P(T = t)$
0.3	0.2
0.15	0.4
0	0.4

Conditional Distributions The conditional distribution is the revised distribution of one random variable obtained after knowing the distribution of another random variable. We denote the distribution of X conditional on Y as $f(x|y)$.

Suppose we have learned that the Blue Chip stock returned 20%. What should our updated beliefs about the returns on the Technology stock be?

$$P(T = 0.3|B = 0.2) = \frac{P(T = 0.3, B = 0.2)}{P(B = 0.2)} = \frac{0.15}{0.2} = 0.75$$

$$P(T = 0.15|B = 0.2) = \frac{P(T = 0.15, B = 0.2)}{P(B = 0.2)} = \frac{0.05}{0.2} = 0.25$$

$$P(T = 0|B = 0.2) = \frac{P(T = 0, B = 0.2)}{P(B = 0.2)} = \frac{0}{0.2} = 0$$

That is

t	$f(t) = P(T = t)$	$f(t\|B = 0.2) = P(T = t\|B = 0.2)$
0.3	0.2	0.75
0.15	0.4	0.25
0	0.4	0

Exercise 3. Find the conditional distribution $f(b|T = 0)$.

3.3 Independence of Random Variables

ex. Joint Distribution of Stock Returns (II) Consider the new table of the joint distribution:

	$\tilde{T}$			
$\tilde{B}$	0.3	0.15	0	
0.2	0.04	0.08	0.08	0.2
0.1	0.12	0.24	0.24	0.6
0.05	0.04	0.08	0.08	0.2
	0.2	0.4	0.4	

What about the conditional distribution $f(\tilde{t}|\tilde{B} = 0.2)$ here?

$$P(\tilde{T} = 0.3|\tilde{B} = 0.2) = \frac{P(\tilde{T} = 0.3, \tilde{B} = 0.2)}{P(\tilde{B} = 0.2)} = \frac{0.04}{0.2} = 0.2$$

$$P(\tilde{T} = 0.15|\tilde{B} = 0.2) = \frac{P(\tilde{T} = 0.15, \tilde{B} = 0.2)}{P(\tilde{B} = 0.2)} = \frac{0.08}{0.2} = 0.4$$

$$P(\tilde{T} = 0|\tilde{B} = 0.2) = \frac{P(\tilde{T} = 0, \tilde{B} = 0.2)}{P(\tilde{B} = 0.2)} = \frac{0.08}{0.2} = 0.4$$

What did you find? It is clear that the conditional distribution is the same as the original marginal distribution. That is, $P(\tilde{T} = \tilde{t}|\tilde{B} = 0.2) = P(\tilde{T} = \tilde{t})$, for all $\tilde{t}$.

In fact, you can check that this is true of ALL distributions listed. Is this a coincidence? No! This example is intended to illustrate that the two random variables, $\tilde{B}$ and $\tilde{T}$, are **Independent**.

Idea: Learning the realizations of some random variables does not provide **new** information about the other random variables.

Definition For two random variables: X and Y are independent if for all realizations x and y,

$$P(X = x, Y = y) = P(X = x)P(Y = y) \tag{3.10}$$

If $P(Y = y) \neq 0$,

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

As well, if $P(X = x) \neq 0$,

$$P(Y = y|X = x) = P(Y = y)$$

Generally, knowing the (marginal) distributions of X and Y is NOT enough to determine their joint distribution (as shown in the case of stock returns of B and T).

However, if the random variables are independent, the marginal distributions are ENOUGH to determine their joint distribution (as shown in the case of stock returns of $\tilde{B}$ and $\tilde{T}$).

Definition For many random variables: X_1, X_2,...,X_n are independent for all realizations x_1, x_2,...,x_n,

$$P\left(\bigcap_{i=1}^{n}\{X_i = x_i\}\right) = \prod_{i=1}^{n} P(X_i = x_i) \tag{3.11}$$

Covariance Covariance is a measure of comovement of two random variables.

$$\begin{aligned}\sigma_{XY} = Cov(X, Y) &= E([X - E(X)][Y - E[Y]]) \\ &= \sum_x \sum_y (x - E(X))(y - E(Y))P(X = x, Y = y)\end{aligned}$$

Since $P(X = x, Y = y) > 0$,

$x - E(X)$	$y - E(Y)$	$Cov(X, Y)$
$+$	$+$	$+$
$-$	$-$	$+$
$+$	$-$	$-$
$-$	$+$	$-$

However, covariances are in strange units: (X unit $\times$ Y unit). So we introduce the concept of correlation between X and Y.

b	t	$P(B=b,T=t)$	$b-E(B)$	$t-E(T)$	$(b-E(B))(t-E(T))P(B=b,T=t)$
0.2	0.30	0.15	0.09	0.18	0.002430
0.2	0.15	0.05	0.09	0.03	0.000135
0.2	0	0	0.09	-0.12	0
0.1	0.30	0	-0.01	0.18	0
0.1	0.15	0.3	-0.01	0.03	-0.000090
0.1	0	0.3	-0.01	-0.12	0.000360
0.05	0.30	0.05	-0.06	0.18	-0.000540
0.05	0.15	0.05	-0.06	0.03	-0.000090
0.05	0	0.1	-0.06	-0.12	0.00720
					$Cov(B,T)=0.002925$

Correlation Coefficient

$$\rho_{XY}=Corr(X,Y)=\frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}$$

- unit-free
- Fact: $-1\le\rho_{XY}\le 1$
  1. $\rho_{XY}=1$: perfect correlation
  2. $\rho_{XY}=-1$: perfect negative correlation
  3. $\rho_{XY}=0$: zero correlation

Proof: $-1\le\rho_{XY}\le 1$ **(Will be shown later).**

ex. Joint Distribution of Stock Returns (I) From Tables **??** and **??**, we can obtain

$$E(B)=0.11 \qquad E(T)=0.12$$
$$Var(B)=0.0024 \qquad Var(T)=0.0126$$
$$SD(B)=\sqrt{Var(B)}=0.0490 \qquad SD(T)=\sqrt{Var(T)}=0.1122$$

$$Corr(B,T)=\frac{Cov(B,T)}{SD(B)SD(T)}=\frac{0.002925}{(0.0490)(0.1122)}\approx 0.53$$

Exercise 4. In the example: Joint Distribution of Stock Returns (II), compute $E(\tilde{B})$, $E(\tilde{T})$, $Var(\tilde{B})$, $Var(\tilde{T})$, $SD(\tilde{B})$, $SD(\tilde{T})$, $Cov(\tilde{B},\tilde{T})$, and $Corr(\tilde{B},\tilde{T})$.

Independence and Correlation

- X, Y independent $\Longrightarrow$ X, Y uncorrelated: $Cov(X,Y) = Corr(X,Y) = 0$ (Check!)
- "$\Longleftarrow$" is **false** since independence is a much more demanding condition
- Why? X, Y independent means: $P(X = x, Y = y) = P(X = x)P(Y = y)$ for <u>all</u> outcomes x, y (<u>many</u> equalities). **BUT** X, Y uncorrelated means $Corr(X,Y) = 0$ (only <u>one</u> equality).

Exercise 5. Prove that X, Y independent implies X, Y uncorrelated.

3.4 Functions of Random Variables

Let X= returns on UMC, Y=returns on eBay. The joint distribution of X and Y is

	Y	
X	0.20	0
0.15	0.4	0.1
0.05	0.2	0.3

We can easily obtain (check!)

$$E(X) = 0.10 \qquad E(Y) = 0.12$$
$$Var(X) = 0.0025 \qquad Var(Y) = 0.0096$$
$$Cov(X,Y) = 0.002$$

Suppose you invest \$ 1000 in UMC and \$ 500 in eBay. The dollar returns on this investment are $Z = 1000X + 500Y$: a new random variable. What is the distribution of returns, and the mean and variance of returns?

If $g(\cdot)$ is a function and $X_1, \ldots, X_n$ are random variables, then $Z = g(X_1, \ldots, X_n)$ is a new random variable. For instance,

- $z = g(x, y) = 1000x + 500y$ (ex post dollar returns, i.e. after the actual returns on UMC and eBay are known)
- $Z = g(X, Y) = 1000X + 500Y$ (ex ante dollar returns)

To compute the distribution of the new random variable:

1. Start with the original joint distribution of $X_1, \ldots, X_n$
2. Determine the possible realizations of $Z = g(X_1, \ldots, X_n)$
3. List them with their probability

x	y	$P(X = x, Y = y)$	$z = 1000x + 500y$
0.15	0.20	0.4	150+100=250
0.15	0	0.1	150
0.05	0.20	0.2	50+100=150
0.05	0	0.3	50

Therefore,

z	$P(Z = z)$
250	0.4
150	0.1+0.2=0.3
50	0.3

Next, to compute $E(Z)$ and $Var(Z)$, we could just make a table as usual. **BUT** since Z is a linear function of A and B, these can actually be computed from $E(X)$, $Var(X)$, etc by some basic properties of random variables.

Basic Properties of Random Variables

1. On random variables

$$E(aX+b) = aE(X) + b \tag{3.12}$$

$$Var(aX+b) = a^2 Var(X) \tag{3.13}$$

$$E(X+Y) = E(X) + E(Y) \tag{3.14}$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y) \tag{3.15}$$

$$Cov(aX+b, cY+d) = acCov(X,Y) \tag{3.16}$$

$$Cov(X,Y) = E(XY) - E(X)E(Y) \tag{3.17}$$

$$Var(X) = E(X^2) - [E(X)^2] \tag{3.18}$$

2. Sums of many random variables: Let $S_n^* = \sum_{i=1}^n X_i$

$$E(S_n^*) = E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i) \tag{3.19}$$

$$Var(S_n^*) = Var(\sum_{i=1}^n X_i) = \sum_{i=1}^n \sum_{j=1}^n Cov(X_i, X_j)$$

$$= \sum_{i=1}^n Var(X_i) + 2\sum_{i=2}^n \sum_{j=1}^{i-1} Cov(X_i, X_j) \tag{3.20}$$

3. Implications of independence

$$X \perp Y \Longrightarrow E(XY) = E(X)E(Y) \tag{3.21}$$

$$X \perp Y \Longrightarrow Cov(X,Y) = 0 \tag{3.22}$$

$$X \perp Y \Longrightarrow Var(X+Y) = Var(X) + Var(Y) \tag{3.23}$$

4. Additional Formulas

$$Cov(X, X) = Var(X) \tag{3.24}$$

$$Cov(X, c) = 0 \tag{3.25}$$

$$\begin{aligned} &Cov(X_1 + X_2, Y_1 + Y_2) \\ &= Cov(X_1, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_1) + Cov(X_2, Y_2) \end{aligned} \tag{3.26}$$

$$-1 \leq Corr(X, Y) \leq 1 \tag{3.27}$$

$$Cov(X, Y) = Cov(Y, X) \tag{3.28}$$

$$Corr(X, X) = 1 \tag{3.29}$$

5. Standardized Random Variable

   Suppose that $X \sim (\mu, \sigma^2)$. Let

$$Z = \frac{X - \mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}.$$

   Then $E(Z) = 0$, $Var(Z) = 1$. $Z \sim (0, 1)$, and is called **standardized random variable**.

Proof: Basic Properties of Random Variables

Example: UMC and eBay Now let's go back to the example of returns on UMC and eBay. Recall that:

$$Z = 1000X + 500Y$$

Then

$$\begin{aligned} E(Z) &= E(1000X + 500Y) \\ &= E(1000X) + E(500Y) \\ &= 1000E(X) + 500E(Y) \\ &= 1000(0.10) + 500(0.12) \\ &= 160 \quad \blacksquare \end{aligned}$$

$$
\begin{aligned}
Var(Z) &= Var(1000X + 500Y) \\
&= Var(1000X) + Var(500Y) + 2Cov(1000X, 500Y) \\
&= 1000^2 Var(X) + 500^2 Var(Y) + 2(1000)(500)Cov(X, Y) \\
&= 1000^2(0.0025) + 500^2(0.0096) + 1000000(0.002) \\
&= 6900 \quad \blacksquare
\end{aligned}
$$

$$
SD(Z) = SD(1000X + 500Y) = \sqrt{Var(Z)} = \sqrt{6900} = 83.06 \quad \blacksquare
$$

Conditional Expectation and Conditional Variance

- Conditional Expectation:

  Let X and Y be jointly distributed random variables. The **conditional expectation** of X given $Y = y$ is

  $$E(X|Y = y) = \sum_x xP(X = x|Y = y) = \sum_x xf(x|Y = y) = g(y)$$

  That is, $E(X|Y = y)$ is a function of y.

  In general,

  $$E(X|Y) = g(Y).$$

- Conditional Variance Let X and Y be jointly distributed random variables. The **conditional expectation** of X given $Y = y$ is

  $$Var(X|Y = y) = E[(X - E(X|Y = y))^2|Y = y)] = \sum_x (x - E(X|Y = y))^2 f(x|Y = y) = h(y).$$

  Thus, $Var(X|Y = y)$ is a function of y as well.

  In general,

  $$Var(X|Y) = h(Y).$$

  We can also write down the conditional expectation and conditional variance of Y given X:

  $$E(Y|X) = g(X),$$

  $$Var(Y|X) = h(X).$$

- Important Properity:

$$E[h(X)Y|X] = h(X)E[Y|X].$$

- Law of Iterated Expectations

$$E(E[Y|X]) = E(Y).$$

- Variance Decomposition

$$Var(Y) = E\left(Var(Y|X)\right) + Var\left(E(Y|X)\right)$$

Exercise 6. Show that $E(X|Y = y) = E(X)$ if X and Y are independent.

3.5 Descriptive Statistics: A Revisit

We have introduced descriptive statistics in Section 2 of Part I. In the previous sections of Part II, we offered many summary statistics for both single random variable and a pair of random variables: the mean $E(X)$, the variance $Var(X)$, the covariance $Cov(X, Y)$ and the correlation $Corr(X, Y)$.

Here we will show that all of these concepts used to characterize random variables have *counterparts* which are used to summarize lists of data. These summary statistics are called *descriptive statistics*. Moreover, while the formulas for random variables and for lists of data look slightly different, we will show that there are actually close connections between the two.

Let $\{x_i\}_{i=1}^N$ be a list of numbers, and $\{(y_i, z_i)\}_{i=1}^N$ be a list of paired numbers. For instance,

- x_i= numbers of cars owned by i-th Taipei resident

- (y_i, z_i)=(years of work experience, income) of i-th Taipei resident

And let X be a random variable representing the number of cars owned by **one randomly chosen** Taipei resident.

	Population	Random Sampling
Mean	$\mu_x = \frac{\sum_{i=1}^{N} x_i}{N}$	$\mu_X = E(X) = \sum_x xP(X = x)$
Variance	$\sigma_x^2 = \frac{\sum_{i=1}^{N}(x_i - \mu_x)^2}{N}$	$\sigma_X^2 = Var(X) - \sum_x (x - \mu_X)^2 P(X = x)$
S.D.	$\sigma_x = \sqrt{\sigma_x^2}$	$\sigma_X = \sqrt{\sigma_X^2}$
Covariance	$\sigma_{yz} = \frac{\sum_y \sum_z (y-\mu_y)(z-\mu_z)}{N}$	$\sigma_{YZ} = \sum_y \sum_z (y - \mu_Y)(z - \mu_Z) P(Y = y, Z = z)$
Correlation	$\rho_{yz} = \frac{\sigma_{yz}}{\sigma_y \sigma_z}$	$\rho_{YZ} = \frac{\sigma_{YZ}}{\sigma_Y \sigma_Z}$

Remarks

1. Be careful about the notations: **lowercase** letters for list of data; **CAPITAL** letters for random variables.

2. The important results are: $\mu_x = \mu_X$, $\sigma_x^2 = \sigma_X^2$, $\sigma_x = \sigma_X$, $\sigma_{yz} = \sigma_{YZ}$, and $\rho_{yz} = \rho_{YZ}$.

Chapter 4

Applications of Probability Theory

4.1 Applications: Diversification, Portfolio Selection, Risk and Return, and the CAPM

In this section, we will use the probability theory we have learned to study the relationship between the returns on a set of risky assets and the expected returns on portfolios created from these assets.

- Assets: numbered 1,2,...,n
- % Returns on assets: $R_1, R_2, \ldots, R_n$
- For typical asset i:

$$R_i = \frac{\text{unknown future price of } i - \text{present price of } i}{\text{present price of } i}$$

According to market data, we can get:

$$\mu_i = E(R_i), \qquad \sigma_i^2 = Var(R_i),$$
$$\sigma_i = SD(R_i), \qquad \sigma_{ij} = Cov(R_i, R_j),$$
$$\rho_{ij} = \frac{Cov(R_i, R_j)}{SD(R_i)SD(R_j)}.$$

We want to invest in a portfolio made up of these risky assets. It turns out to be simplest to think about our investment in *percentage* form: what % of our capital will be invested in each asset (what portion of each dollar will be invested in each asset).

Portfolio : $\vec{p} = (p_1, \ldots, p_n)$,[1] $\sum_i p_i = 1$. The % returns on portfolio:

$$R_p = \sum_{i=1}^{n} p_i R_i$$

Clearly, R_p is a new random variable with mean and variance, $E(R_p)$, $Var(R_p)$ respectively.

Some Stylized Facts about the Capital Market

1. Higher expected return, $E(R_i)$ is accompany with higher risk, $Var(R_i)$.
2. Diversifying has reduced risk. That is, $Var(R_i) \leq Var(R_p)$ for most i.

Two Asset Portfolios

1. $\vec{p} = (p_1, p_2)$
2. $R_p = p_1 R_1 + p_2 R_2$

Thus, the mean and variance of R_p are:

$$\begin{aligned} E(R_p) &= E(p_1 R_1 + p_2 R_2) \\ &= E(p_1 R_1) + E(p_2 R_2) \\ &= p_1 E(R_1) + p_2 E(R_2) \\ &= p_1 \mu_1 + p_2 \mu_2 \end{aligned} \tag{4.1}$$

$$\begin{aligned} Var(R_p) &= Var(p_1 R_1 + p_2 R_2) \\ &= Var(p_1 R_1) + Var(p_2 R_2) + 2Cov(p_1 R_1, p_2 R_2) \\ &= p_1^2 Var(R_1) + p_2^2 Var(R_2) + 2p_1 p_2 Cov(R_1, R_2) \\ &= p_1^2 Var(R_1) + p_2^2 Var(R_2) + 2p_1 p_2 SD(R_1) SD(R_2) Corr(R_1, R_2) \\ &= p_1^2 \sigma_1^2 + p_2^2 \sigma_2^2 + 2p_1 p_2 \sigma_1 \sigma_2 \rho_{12} \end{aligned} \tag{4.2}$$

[1] The arrow over the symbol $\vec{p}$ means that $\vec{p}$ is a **vector**, which is simply a list of numbers. This is just a shorthand which saves us from having to write out the entire list $(p_1, \ldots, p_n)$ again and again.

Example

- $R_S = \%$ returns on stock S
- $R_T = \%$ returns on stock T
- $\mu_S = 0.08$, $\sigma_S = 0.03$, $\mu_T = 0.16$, $\sigma_T = 0.06$, $\rho_{ST} = -0.25$
- Suppose our portfolio consists of 0.75 stock S and 0.25 stock T. That is, $\vec{p} = (p_S, p_T) = (0.75, 0.25)$

Therefore,

$$E(R_p) = (0.75)(0.08) + (0.25)(0.12) = 0.10$$
$$Var(R_p) = (0.75)^2(0.03)^2 + (0.25)^2(0.06)^2 + 2(0.75)(0.25)(0.03)(0.06)(-0.25) = 0.0005625$$
$$SD(R_p) = \sqrt{0.0005625} = 0.0237$$

Remark: $SD(R_p)$ is less than σ_S and σ_T. ***Diversifying has reduced risk***.

People may ignore the effects due to correlations in the asset returns and think about the risk of the portfolio in the following way:

$$SD_A = 0.75\sigma_S + 0.25\sigma_T = 0.75(0.03) + 0.25(0.06) = 0.0375$$

That is, we measure the risk of the portfolio as the weighted average of the risk of each asset in the portfolio. We may call SD_A as **naive estimate** of the risk. Clearly, $SD_A > SD(R_p)$: the naive estimate is greater than the actual standard deviation of the portfolio.

In our example, the correlation between the two securities is *negative*; this would appear to be the reason that: (1) diversification reduces risk; (2) our naive estimate was too high. Surprisingly, this intuition is *false*: the naive estimate will be too high so long as the assets' returns are not *perfectly* correlated. This can be shown in the following two propositions:

Proposition 1: Naive Estimate Overstate Risk Suppose that $p_1 > 0$, $p_2 > 0$ and $\rho_{12} < 1$. Then $SD(R_p) < SD_A$ and $Var(R_p) < Var_A$.

Proof: Naive Estimate Overstate Risk

This proposition simply shows that naive estimates overstate risk whenever the assets returns are NOT perfectly corrected.

First consider the following combinations of asset S and T :

(p_1, p_2)	$E(R_p)$	$SD(R_p)$
(1,0)	0.08	0.0300
(0.75,0.25)	0.10	0.0237
(0.5,0.5)	0.12	0.0300
(0.25,0.75)	0.14	0.0437
(0,1)	0.16	0.0600

A simple plot of these points can be easily shown.

Moreover, we may present a plot depicting the means and standard deviations of portfolio choices with the portfolio share in stock S, p_1, taking values from 0 to 1. In Figure **??** the point MV represents returns of minimum variance (MV) portfolio with $\vec{p} = (p_1, p_2) = (0.75, 0.25)$. Point (1,0) with $(E(R_p), SD(R_p)) = (0.08, 0.03)$ and point (0,1) with $(E(R_p), SD(R_p)) = (0.16, 0.06)$ represent all stock S portfolio and all stock T portfolio respectively.

Exercise 7. MV is the point on the curve ar which the standard deviation is minized. Show that at MV,

$$p_1 = \frac{\sigma - 2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

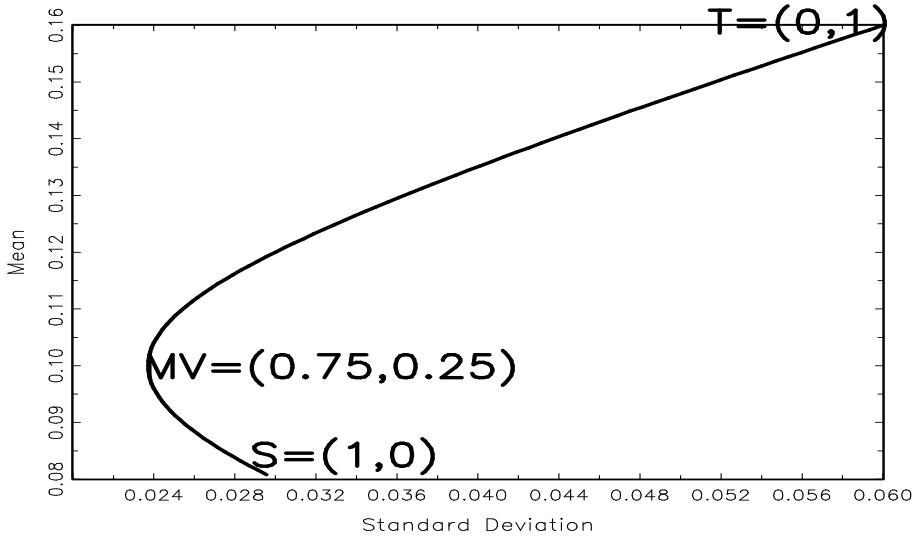
Exercise 8. Where do the naive estimates of the portfolio risk (standard deviations of the portfolios) lie in Figure **??**?

Before we present the Proposition 2, we first introduce the concept of *Efficient Set*.

Efficient Set

1. A portfolio is dominated if there is another portfolio with a higher mean return AND lower standard deviation.

Figure 4.1: Feasible Set of Portfolio Choices



2. The **Efficient Set** contains the returns of portfolios which are not dominated. Choice among these points on efficient set is determined by one's risk preference. It is also called *Efficient Frontier*.

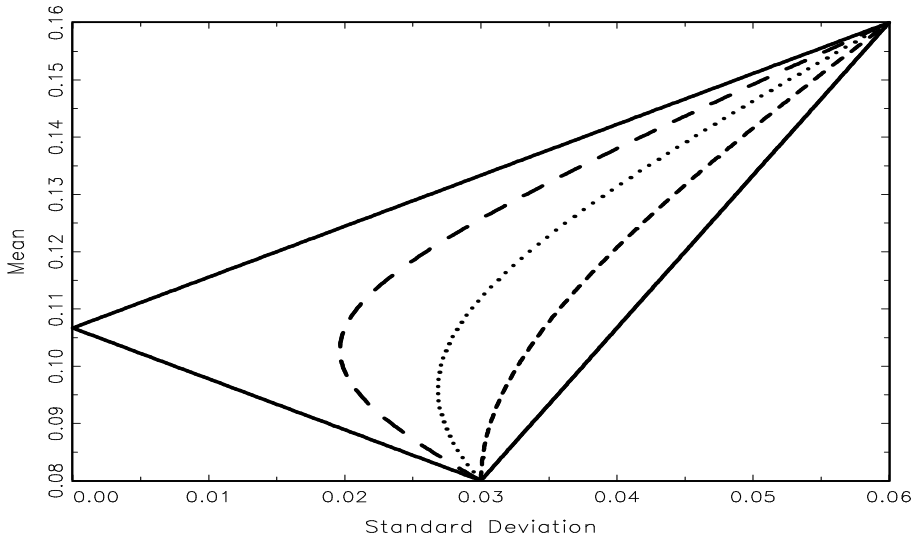
Remarks

- $\widehat{ST}$: Feasible Set
- $\widehat{MVT}$: Efficient Set
- $\widehat{MVS}$: Dominated Set

Proposition 2: Backward Bending Suppose that $\sigma_1 < \sigma_2$ and that $\mu_1 < \mu_2$. If $\rho_{12} < \frac{\sigma_1}{\sigma_2}$, then $SD(R_p) < \sigma_1 < \sigma_2$ **for some portfolio** $\vec{p}$, and the portfolio consisting solely of asset 1 (the asset with lower σ) is not in the efficient set.

Proof: Backward Bending

Figure 4.2: Feasible Sets with Cirrelation=-1, -0.5, 0, 0.5, and 1



Remarks

- For $\rho_{12} = 1$. This is the case that the two assets are perfectly correlated. Thus they can be viewed as the same asset. No diversification!

- For $\rho_{12} = -1$. This is the case that the two assets are perfectly uncorrelated. The variance of the portfolio , $Var(R_p) = (p_1\sigma_1 - (1-p_1)\sigma_2)^2$, can attend 0 while p_1 is set at $\frac{\sigma_2}{\sigma_1+\sigma_2}$. In addition,

$$\frac{dVar(R_p)}{dp_1} \gtreqless 0 \quad \text{when} \quad p_1 \gtreqless \frac{\sigma_2}{\sigma_1+\sigma_2}$$

and

$$SD(R_p) = \sqrt{Var(R_p)} \begin{cases} p_1\sigma_1 - (1-p_1)\sigma_2, & p_1 \geq \frac{\sigma_2}{\sigma_1+\sigma_2} \\ (1-p_1)\sigma_2 - p_1\sigma_1, & p_1 < \frac{\sigma_2}{\sigma_1+\sigma_2} \end{cases}$$

Since

$$\frac{dE(R_p)}{dp_1} = \mu_1 - \mu_2 < 0$$

and

$$\frac{dSD(R_p)}{dp_1} = \begin{cases} \sigma_2 + \sigma_2, & p_1 \geq \frac{\sigma_2}{\sigma_1+\sigma_2} \\ -(\sigma_1 + \sigma_2), & p_1 < \frac{\sigma_2}{\sigma_1+\sigma_2} \end{cases}$$

we have

$$\frac{dE(R_p)}{dSD(R_p)} = \begin{cases} \frac{\mu_1-\mu_2}{\sigma_2+\sigma_2} < 0, & p_1 \geq \frac{\sigma_2}{\sigma_1+\sigma_2} \\ \frac{\mu_1-\mu_2}{-(\sigma_1+\sigma_2)} > 0, & p_1 < \frac{\sigma_2}{\sigma_1+\sigma_2} \end{cases}$$

Many-Asset Portfolios

1. Assets: numbered 1,2,...,n
2. % Returns on assets: $R_1, R_2, \ldots, R_n$
3. $E(R_i) = \mu_i$, $Var(R_i) = \sigma_i^2$, $Cov(R_i, R_j) = \sigma_{ij}$
4. A portfolio $\vec{p} = (p_1, p_2, \ldots, p_n)$ with $\sum_{i=1}^n p_i = 1$
5. $R_p = \sum_{i=1}^n p_i R_i$

By repeatedly using the reasoning from the two asset case, it is not hard to show that:

$$E(R_p) = \sum_{i=1}^{n} p_i \mu_i \tag{4.3}$$

$$Var(R_p) = \sum_{i=1}^{n} p_i^2 \sigma_i^2 + 2\sum_{i=2}^{n}\sum_{j<i} p_i p_j \sigma_{ij} = \sum_{i=1}^{n} p_i^2 \sigma_i^2 + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} p_i p_j \sigma_{ij} \tag{4.4}$$

Risk-Free Borrowing and Lending Suppose we have decided to split our investment between a portfolio of risky assets $\vec{p}$ and some *risk free* asset with return r_f. ($E(r_f) = r_f$, $Var(r_f) = 0$)

Let

- s= percentage invested in portfolio $\vec{p}$
- $(1 - s)$= percentage invested in the risk free asset
- This combined portfolio has returns $Z_s = sR_p + (1 - s)r_f = s\sum_i p_i R_i + (1 - s)r_f$

We can get

- $Cov(R_p, r_f) = 0$
- $E(Z_s) = r_f + (E(R_p) - r_f)s$
- $Var(Z_s) = s^2(Var(R_p))$
- $SD(Z_s) = sSD(R_p)$

Remarks:

1. Both $E(Z_s)$ and $SD(Z_s)$ are *linear function* of s. (what does it mean?)
2. Since
$$\frac{dE(Z_s)}{ds} = E(R_p) - r_f > 0 \text{ (why?)}$$
and
$$\frac{dSD(Z_s)}{ds} = SD(R_p)$$
we have
$$\frac{dE(Z_s)}{dSD(Z_s)} = \frac{E(R_p) - r_f}{SD(R_p)} > 0$$
3. What if $s > 1$? $s > 1 \Rightarrow (1 - s) < 0$: Borrowing at the risk free rate

- *Question:* what do the optimal combinations of risky assets and the risk-free asset look like?
- *Answer:* Capital Market Line! (See Figure **??**)

Capital Market Line When a risk free asset is included in the model we can draw a line from the r_f rate to the efficient frontier, we have what is called a capital allocation line (CAL). We obviously want to take the highest CAL (the one with the highest return for a given level of risk). Thus, the optimal CAL will be just tangent to the efficient frontier. If the CAL is tangent at the market portfolio, then the CAL is called the capital market

line (CML). The point of tangency corresponds to a portfolio on the efficient frontier. That portfolio is called the **super-efficient portfolio**.

Using the risk-free asset, investors who hold the super-efficient portfolio may:

- leverage their position by shorting the risk-free asset and investing the proceeds in additional holdings in the super-efficient portfolio, or
- de-leverage their position by selling some of their holdings in the super-efficient portfolio and investing the proceeds in the risk-free asset.

The resulting portfolios all fall on the capital market line. Accordingly, portfolios which combine the risk free asset with the super-efficient portfolio are superior to the portfolios on the efficient frontier.

Separation Theorem ?)[2] shows that portfolio construction should be a two-step process. First, investors should determine the super-efficient portfolio. This should comprise the risky portion of their portfolio. Next, they should leverage or de-leverage the super-efficient portfolio to achieve whatever level of risk they desire. Significantly, the composition of the super-efficient portfolio is independent of the investor's appetite for risk. The two decisions:

1. the composition of the risky portion of the investor's portfolio, and
2. the amount of leverage to use,

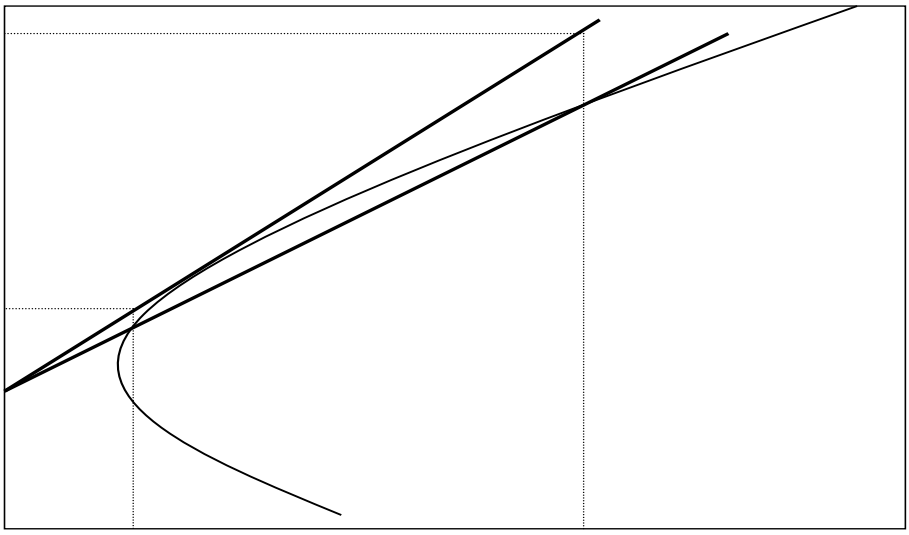
are entirely independent of one another. One decision has no effect on the other. This is called Tobin's **separation theorem**.

Risk and Return

"The market rewards the bearing of risk."

[2]Professor James Tobin, Yale University, USA. 1981 winner of The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel

Figure 4.3: The Capital Market Line



Since individual investors are risk averse, we should expect the investors who are willing to bear risks will be rewarded with high expected returns.

But what type of risk is one rewarded for bearing? Let R_m denotes the returns on the super-efficient portfolio ($\vec{m}$) defined above.

- absolute risk: $Var(R_i)$
- marginal risk:

$$\beta_i = \frac{Cov(R_i, R_m)}{Var(R_m)}$$

The market only rewards the bearing of marginal risk.

Interpretation of Beta

1. Roles of Risk-Free Assets

   Consider the following portfolio with risky-free asset

$$R_{rf} = \alpha r_f + (1 - \alpha) R_m,$$

$$Var(R_{rf}) = (1 - \alpha)^2 Var(R_m)$$

Suppose that initially we do not hold Risk-free Asset ($\alpha = 0$). Now we increase our holdings of Risk-free Asset a little bit and see how it affects the risk of our portfolio, $Var(R_{rf})$.

$$Var(R_{rf}) = (1-\alpha)^2 Var(R_m)$$

$$\frac{dVar(R_m)}{d\alpha} \approx \frac{dVar(R_{rf})}{d\alpha}\Big|_{\alpha=0} = [-2(1-\alpha)Var(R_m)]_{\alpha=0} = -2Var(R_m) < 0$$

2. Roles of Risky Assets

Consider the following portfolio with risky asset i

$$R = \alpha R_i + (1-\alpha)R_m,$$

$$Var(R) = (1-\alpha)^2 Var(R_m) + 2\alpha(1-\alpha)Cov(R_i, R_m) + \alpha^2 Var(R_i)$$

Suppose that initially we do not hold Asset i ($\alpha = 0$). Now we increase our holdings of Asset i a little bit and see how it affects the risk of our portfolio, $Var(R)$.

$$\begin{aligned}\frac{dVar(R_m)}{d\alpha} \approx \frac{dVar(R)}{d\alpha}\Big|_{\alpha=0} &= [-2(1-\alpha)Var(R_m) + 2(1-2\alpha)Cov(R_i, R_m) + 2\alpha Var(R_i)]_{\alpha=0} \\ &= 2[Cov(R_i, R_m) - Var(R_m)] \\ &= 2[\beta_i - 1]Var(R_m)\end{aligned}$$

Capital Portfolio Pricing Model (CPPM): The Beta Relation for pricing super-efficient portfolios.

$$E(R_i) = r_f + \beta_i[E(R_m) - r_f].$$

Or

$$\underbrace{E(R_i) - r_f}_{\text{risk premium of } i} = \beta_i \times \underbrace{[E(R_m) - r_f]}_{\text{risk premium of portfolio } m}$$

Proof: The Beta Relation

Table 4.1: Interpretation of Beta

Value of Beta	Effect on portfolio $\vec{m}$ of increasing the holdings of the asset i
$\beta_i > 1$	Adds risk to the overall portfolio
$\beta_i = 1$	Does not affect risk
$0 < \beta_i < 1$	Reduces risk, although not as well as the risk-free asset
$\beta_i = 0$	Reduces risk as well as the risk-free asset
$\beta_i < 0$	Reduce risk better than the risk-free asset

The Capital Asset Pricing Model (CAPM) **?**)[3] demonstrates that: given strong simplifying assumptions,[4] the super-efficient portfolio must be the market portfolio. All investors will hold the market portfolio, leveraging or de-leveraging it with positions in the risk-free asset in order to achieve a desired level of risk.

Instead of just talking about a single investor, the CAPM considers all investors in the economy. By homogenous expectations, all investors draw the same feasible return diagram. That means they have same super-efficient portfolio $\vec{m}$.

Therefore, $\vec{m}$ captures the asset holdings of the economy as a whole. That is the super-efficient portfolio $\vec{m}$ is also the market portfolio.

Example Two risk assets, $\vec{m} = (1/2, 1/2)$ for all investors.

Investor	Total \$	\$ in risk-free	Asset 1	Asset 2
1	100	90	5	5
2	200	80	60	60
3	250	50	100	100
			165	165

Thus, ratio=$(1/2, 1/2)$.

[3]Professor William F. Sharp, Stanford University Stanford, CA, USA. 1990 winner of The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel

[4](1) There are no taxes or transaction costs; (2) All investors have identical investment horizons; (3) All investors have homogenous expectations regarding the expected returns, volatilities and correlations of available risky investments.

Systematic and Idiosyncratic Risk CAPM divides the risk of holding risky assets into **systematic** and **idiosyncratic** risk.

Systematic risk (market risk) is the risk of holding the market portfolio. As the market moves, each individual asset is more or less affected. Systematic risk is a source of variance σ_i^2, and covariance $\sigma_{ij}, i \neq j$.

Idiosyncratic risk (firm specific risk) is the risk which is unique to an individual asset. It represents the component of an asset's return which is uncorrelated with general market moves. Idiosyncratic risk is a source of variance σ_i^2, but NOT covariance $\sigma_{ij}, i \neq j$.

According to CAPM, the marketplace compensates investors for taking systematic risk, but not for taking idiosyncratic risk. (i.e., high $E(R_i)$ iff. high β_i and β_i is proportional to $Cov(R_i, R_m)$.) This is because idiosyncratic risk can be diversified away. When an investor holds the market portfolio, each individual asset in that portfolio entails idiosyncratic risk, but through diversification, the investor's net exposure is just the systematic risk of the market portfolio.

Systematic risk can be measured using beta. According to CAPM, the expected return of a stock equals the risk-free rate plus the portfolio's beta multiplied by the expected excess return of the market portfolio.

Chapter 5

Some Useful Discrete Random Variables

Example: Sweepstakes

Bernoulli Trials Process

1. Repeatedly tossing a coin.
2. Performing a clinical trial for a new vaccine.
3. Conducting a yes/no opinion poll.
4. Controlling quality in a production process by determining whether or not pieces from the production line are defective.
5. Determining prize winners in the sweepstakes described above.

Each of these examples involves repeated random trials. What features do all of the examples have in common? We can point to three:

1. Each trial has two possible outcomes , 1 or 0. In the examples, these correspond to Heads or Tails, Ill or Not Ill, Yes or No, Defective or Acceptable, and Winner or Non-winner.
2. The probability p of outcome 1 is the same in every trial.

3. The trials are independent.

Any sequence of repeated observations satisfying these three properties defines a Bernoulli Trials Process, or a BTP for short.

5.1 Indicator Random Variables

Indicator random variables indicate whether a certain event occurs.

Definition: Let $A \subseteq \Omega$ be an event. The indicator random variable for the event A, denoted I_A, is

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \in A^c \end{cases}$$

ex.

$$I = \begin{cases} 1 & \text{if H} \\ 0 & \text{if T} \end{cases}$$

5.2 Bernoulli Random Variables

Definition: The random variable X has a Bernoulli distribution with parameter p (probability of success) if its distribution is

x	$P(X = x)$
1	p
0	$(1-p)$

We denote it as $X \sim$Bernoulli(p).

Clearly, an indicator random variable is always a Bernoulli random variable with parameter $p = P(A)$. That is,

$$I_A \sim \text{Bernoulli}(P(A))$$

ex. Tosses a Fair Coin To model one toss of a fair coin we could create a probability model with events H and $T = H^c$ with $P(H) = P(T) = 1/2$. We can also define a random variable X which indicates whether the toss come up heads.

$$X = \begin{cases} 1 & \text{if } H \text{ occurs} \\ 0 & \text{if } H^c = T \text{ occurs} \end{cases}$$

The the distribution of X is

x	$P(X = x)$
1	1/2
0	1/2

More generally, we might want to talk about whether some events with probability p occurs. (For instance, a coin with bias p).

If $X \sim$Bernoulli(p),

$$E(X) = 1p + 0(1-p) = p$$

$$Var(X) = E(X^2) - [E(X)]^2 = p - p^2 = p(1-p)$$

5.3 Bernoulli Trials Process

What if we have a sequence of tosses of a coin with bias (probability of heads) p?

Let $\{X_i\}_{i=1}^n$ be a sequence of random variables, where X_i indicates whether heads occurred on the i-th toss. Then

$$X_i \sim \text{Bernoulli}(p) \quad \forall i$$

$\{X_i\}_{i=1}^n$ is a sequence of **independent** random variables (learning the result of one toss does not provide information about the results of other tosses) We call such a sequence $X_{i i=1}^n$ as a **Bernoulli Trials Process**.

Definition: The sequence of random variables $\{X_i\}_{i=1}^n$ is a Bernoulli Trials Process with parameters n and p, denoted

$$\{X_i\}_{i=1}^n \sim \text{BTP}(n, p)$$

if the random variables, X_1, X_2,..., X_n are **independent**, and each has a Bernoulli distribution with parameter p.

Therefore, we can calculate the probability of any combination of outcomes using the product formula:

$$P(\bigcap_j \{X_j = x\}) = \prod_j P(X_j = x)$$

For instance,

	X_2	
X_1	1	0
1	p^2	$p(1-p)$
0	$(1-p)p$	$(1-p)^2$

ex. Coin Tossing Toss a biased coin with bias 0.3 ($p = P(X = 1) = 0.3$) n times.

$$\{X_i\}_{i=1}^n \sim \text{BTP}(n, 0.3)$$

- For instance, if $n = 2$,

$$X_1 \sim \text{Bernoulli}(0.3)$$

$$X_2 \sim \text{Bernoulli}(0.3)$$

$$\{X_i\}_{i=1}^2 \sim \text{BTP}(2, 0.3)$$

The joint distribution of X_1 and X_2 is as follows:

	X_2		
X_1	1	0	
1	0.09	0.21	0.3
0	0.21	0.49	0.7
	0.3	0.7	

Remarks

1. In general, we call sequences of random variables which are independent and share the same distribution **i.i.d. random variables**.

2. Two random variables with the same distribution (they are identically distributed) **need not** be identical! For instance,

$$X_1 \sim \text{Bernoulli}(0.3)$$

$$X_2 \sim \text{Bernoulli}(0.3)$$

The sample space is

$$\Omega = \{HH, HT, TH, TT\}$$

and

$$P(\{HH\}) = 0.09$$
$$P(\{HT\}) = 0.21$$
$$P(\{TH\}) = 0.21$$
$$P(\{TT\}) = 0.49$$

Consider following definition of random variables, X_1 and X_2:

$$X_1(\omega) = \begin{cases} 1 & \omega \in \{HH, HT\} \\ 0 & \omega \in \{TH, TT\} \end{cases}$$

$$X_2(\omega) = \begin{cases} 1 & \omega \in \{HH, TH\} \\ 0 & \omega \in \{HT, TT\} \end{cases}$$

To indicate the event $\omega = \{HT\}$, $X_1 = 1$, $X_2 = 0$. As well, to indicate the event $\omega = \{TH\}$, $X_1 = 0$, $X_2 = 1$.

That is, X_1 **and** X_2 **have same distribution but** X_1 **and** X_2 **are NOT identical**.

5.4 The Binomial Distribution

Question: Suppose I flip a coin with known bias ($P(\text{Heads}) = 0.6$) 3 times. What is the probability I get exactly 2 heads?

Answer: Let $H_i = \{\text{the i-th flip is heads}\}$, $T_j = \{\text{the j-th flip is tails}\}$.

$$\begin{aligned} P(H_1 \cap H_2 \cap T_3) &= P(H_1)P(H_2)P(T_3) \quad \text{by independence} \\ &= (0.6)^2(0.4) = 1.44 \end{aligned}$$

$$P(H_1 \cap T_2 \cap H_3) = P(H_1)P(T_2)P(H_3) = (0.6)(0.4)(0.6) = 1.44$$

$$P(T_1 \cap H_2 \cap H_3) = P(T_1)P(H_2)P(H_3) = (0.4)(0.6)^2 = 1.44$$

That is, each outcome with 2 heads and 1 tail has the same probability. If I could count how many such outcomes there are, by multiplying by this number I could get the result.

$$P(\text{Exactly 2 heads}) = P(H_1 \cap H_2 \cap T_3) + P(H_1 \cap T_2 \cap H_3) + P(T_1 \cap H_2 \cap H_3) = 0.432$$

More generally: Suppose I perform n flips of a coin with bias p. Let S be the random variable representing the number of heads. Then S is said to have a binomial distribution and denoted as $S \sim b(n, p)$

$$P(S = s) = \binom{n}{s} p^s (1-p)^{n-s}$$

where

$$\binom{n}{s} = \frac{n!}{s!(n-s)!}$$

Exercise 9. If we flip a fair coin 10 times, what is the probability of getting exactly 5 heads?

Remark What are the characteristics of a process for which a binomial random variable is an appropriate description?

1. Each trial has two possible outcomes
2. The probability of success is the same in each trial
3. The trials are independent

5.5 I.I.D. Random Variables

The heart of probability theory. I.I.D. is referred to "Independently and Identically Distributed." A sequence of random variables $\{X_i\}_{i=1}^n$ is i.i.d. if

- **Independence**: one provides no information about others, product rule for probabilities
- **Identically Distributed**: same marginal distribution

ex. 2 Random Variables X_1, X_2

	X_2		
X_1	10	20	
10	0.09	0.21	0.3
20	0.21	0.49	0.7
	0.3	0.7	

	X_2		
X_1	10	20	
10	0.18	0.42	0.6
20	0.12	0.28	0.4
	0.3	0.7	

	X_2		
X_1	10	20	
10	0.20	0.10	0.3
20	0.10	0.60	0.7
	0.3	0.7	

Remark: Independence and equality of distribution are two different ideas.

5.6 Functions of I.I.D. Random Variables

$$\{X_i\}_{i=1}^n \quad X_i \text{ i.i.d. with } E(X_i) = \mu, Var(X_i) = \sigma^2$$

$$S_n^* = \sum_{i=1}^n X_i \quad \text{(sum of random variables)}$$

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \quad \text{(average of random variables)}$$

Exercise 10. Show that

$$E(S_n^*) = n\mu$$

$$Var(S_n^*) = n\sigma^2$$

5.7 Binomial Random Variables Revisited

Let

$$S \sim b(n,p), \qquad P(S=s) = \binom{n}{s} p^s (1-p)^{n-s}$$

Where s represents number of heads in n tosses. What is

$$E(S) = \sum_s sP(S=s)$$

and

$$Var(S) = \sum_s (s - E(S))P(S=s)?$$

A tough calculation!

But we also know that the number of heads in n tosses can be represented as the sum $S_n^* = \sum_{i=1}^n X_i$, where X_i is i.i.d. Bernoulli random variables, $X_i \sim$ Bernoulli(p). Therefore, the binomial random variable S has the same distribution as sum of i.i.d. Bernoulli random variables $\sum_i^n X_i$. We can then compute $E(S)$ and $Var(S)$ as follows: recall that $E(X_i) = p$ and $Var(X_i) = p(1-p)$,

$$E(S) = n\mu = np$$

$$Var(S) = n\sigma^2 = np(1-p)$$

This is clearly the relationship between BTP and binomial distribution.

BTP v.s. Binomial Distribution Let

$$\{X_i\}_{i=1}^n \sim \text{BTP}(n, p)$$

and

$$S = \sum_{i=1}^n X_i$$

then

$$S \sim \text{binomial}(n, p).$$

Exercise 11. Show that[1]

$$\sum_{s=0}^{\infty} \binom{n}{s} p^s (1-p)^{n-s} = 1$$

5.8 Some Other Special Discrete Distributions

Negative Binomial Distribution Let X be the number of **tails** until k heads have been obtained.

$$P(X = x) = \binom{x+k-1}{k-1} p^{k-1}(1-p)^x \times p = \binom{x+k-1}{k-1} p^k (1-p)^x$$

[1]You need the Binomial Fomula: $(a+b)^n = \sum_{s=0}^n \binom{n}{s} b^s a^{n-s}$.

Negative Binomial Distribution (again!) Let X be the number of **trial** on which k heads have been obtained.

$$P(X = x) = \binom{x-1}{k-1} p^{k-1}(1-p)^{x-k} \times p = \binom{x-1}{k-1} p^k (1-p)^{x-k}$$

Geometric Distribution Let X be the number of **trial** at which the first heads occurs. That is, $k = 1$ in the above case.

$$P(X = x) = (1-p)^{x-1} p$$

Just looks like a Geometric series: $\sum_{r=0}^{\infty} ay^r = \sum_{r=1} ay^{r-1} = a(1 + y + y^2 + y^3 + \cdots) = \frac{a}{1-y}$. Hence,

$$\sum_{x=1}^{\infty} (1-p)^{x-1} p = \frac{p}{1-(1-p)} = 1.$$

Poisson Distribution Consider the binomial distribution $b(n, p)$ and let $n \to \infty$, $p \to 0$ but in such a way that $np = \lambda (> 0)$ for all n and p. That is, the probability of heads is very small and the number of trials is large.

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!},$$

where $e \approx 2.71829$.

A Poisson distributed random variable is often useful in estimating the number of occurrences over a specified interval of time or space It is a discrete random variable that may assume an infinite sequence of values ($x = 0, 1, 2, \ldots$). We denote it as $X \sim$ Poisson(λ). We will discuss the meaning of λ later.

Exercise 12. Show that

$$\sum_{x=0}^{\infty} \left(\frac{e^{-\lambda} \lambda^x}{x!} \right) = 1$$

Two Properties of a Poisson Experiment

1. The probability of an occurrence is the same for any two intervals of equal length.
2. The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.

Examples of Poisson Distribution

1. Number of arrivals at a McDonald's counter.
2. Number of repairs needed in 10 miles of highway.
3. Number of arrivals at the drive-up teller window of a bank during a 15-minute period on weekday mornings.

Important Facts If $X \sim \text{Poisson}(\lambda)$,

$$E(X) = Var(X) = \lambda$$

Proof: Remember that

$$e^{\lambda} = \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = \cdots$$

Hence, simply find out $E(X)$ and $E(X(X-1))$ by definition.

$$E(X) = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\lambda \sum_{x=1}^{\infty} \frac{\lambda^{x-1}}{(x-1)!} = \lambda$$

$$E(X(X-1)) = \sum_{x=0}^{\infty} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\lambda^2 \sum_{x=2}^{\infty} \frac{\lambda^{x-2}}{(x-2)!} = \lambda^2$$

Poisson Limit Theorem With $\lambda = np$,

$$\binom{n}{x} p^x (1-p)^{n-x} \longrightarrow \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{as} \quad n \to \infty, p \to 0$$

That is, a Binomial random variable converges to a Poisson random variable:

$$P(S_n = s) \longrightarrow P(X = x)$$

Proof: You need (1) $\lim_{n\to\infty}(1 + \frac{y}{n}) = e^y$, and (2) Taylor's expansion: $f(x) = f(c) + \frac{f'(c)}{1!}(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots$.

Sweepstakes Revisited

$$\binom{n}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{n-k} \longrightarrow \frac{e^{-1}1^k}{k!}$$

Hypergeometric Distribution

$$P(X = x) = \frac{\binom{r}{x}\binom{N-r}{n-x}}{\binom{N}{n}}$$

where:

$P(X = x)$ = probability of x successes in n trials

n = number of trials

N = number of elements in the population

r = number of elements in the population labeled success

The hypergeometric distribution is closely related to the binomial distribution. However, for the hypergeometric distribution:

1. the trials are not independent
2. the probability of success changes from trial to trial

Remark

1. $E(X) = n\left(\frac{r}{N}\right)$
2. $Var(X) = n\left(\frac{r}{N}\right)\left(1 - \frac{r}{n}\right)\left(\frac{N-n}{N-1}\right)$
3. When the population size is large, a hypergeometric distribution can be approximated by a binomial distribution with n trials and a probability of success $p = \frac{r}{N}$.

Chapter 6

Continuous Random Variables

6.1 Introduction

Motivation: Suppose we have a spinner along the outside of which we wrap the number line from zero to one. Thus "all results are equally likely". Suppose we would like to make probability statements about the result of a spin. Let X be the random variable representing the result of the spin. What we have done in the past is assign probabilities to each possible outcome. However, this is clearly **impossible** here: there are infinitely many outcomes, so problems will arise if we start trying to assign them all positive probability (in some uniform fashion). Thus "zero probability" no longer means impossible.

$$P(X = x) = 0 \quad \forall\, 0 \leq x < 1$$

The trick is to talk about events which have positive probability. For instance, $P(0 \leq x < 1)$ or $P(0 \leq x \leq 1/2)$. So it is useful to represent probability as areas under a curve.

Probability Density Function f is a probability density function (pdf) if $f(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f(x)dx = (\text{Area under } f) = 1$$

Continuous Random Variables X is a *continuous random variable* with pdf f if

$$P(a \leq X \leq b) = \int_a^b f(x)dx = (\text{Area under } f \text{ between } a \text{ and } b)$$

Remarks:

1. We do not care about f itself, only the areas underneath. It is these areas which represent probability, not the values of the function itself. In other word, density is not probability: $f(x) \neq P(X = x)$.

2. Consider the event $A = \{X = a\} = \{a \leq X \leq a\}$. Thus, $P(A) = P(\{X = a\}) = \int_a^a f(x)dx = 0$. That is, in the continuous case $P(X = x) = 0$ for every x. This means that the probability that X takes on a particular value x is zero. Do not confuse two distinct concepts: **zero probability** and **impossibility**. In the continuous case, a zero probability event is NOT an impossible event.

3. How can $P(X = x) = 0$ make sense? Suppose not, i.e., $P(X = x) = p > 0$. Let event A contains n distinct outcomes. Each outcome occurs with probability p, and $0 < p < 1$. Clearly, $p > \frac{1}{n}$ for n large. Hence,

$$P(X \in A) = \sum_{x \in A} P(X = x) = \sum_{x \in A} p = np > 1,$$

   a contradiction.

4. **Can many nothings make something?** The probability of an impossible event is zero, but the contrary is not true. For example, the probability that we hit the very centre of a target is zero, though it is not impossible. In fact, given any fixed point on the target, the probability that we hit that point is also zero. Does it mean that it is impossible to hit the target at all! Certainly not. If we take all the points on the target that we can hit and form an event, this event will have probability one. Therefore, many nothings make something. If a point on the real line has length zero, why is it that all the points from the interval $[a, b]$ together form a line segment of length $(b - a)$, which is non-zero? The answer to all these is related to the concept of countability. The sets of points, for example, the points on the interval $[a, b]$, are known as uncountable sets. Intuitively, it means that the number of points cannot be counted. When we have an uncountable number of zeroes, the

sum can be non-zero but when we have a countable number of zeroes, the sum will always remain zero.

5. While for a discrete random variable one could say that an event with probability zero is impossible, this can not be said in the case of a continuous random variable. The probability that X attains a value in an uncountable set (for example an interval) can not be found by adding the probabilities for individual values.

Exercise 13. For a continuous random variable X verify the following.[1]

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du \tag{6.1}$$

$$f(x) = \frac{dF(x)}{dx} \tag{6.2}$$

$$\int_{a}^{b} f(u)du = F(b) - F(a) \tag{6.3}$$

$$P(a < X \leq b) = P(a \leq X < b) = P(a < X < b) = P(a \leq X \leq b) = F(b) - F(a) \tag{6.4}$$

$$F(-\infty) = 0, \quad F(\infty) = 1 \tag{6.5}$$

$$F(\cdot) \text{ monotonically nondecreasing} \tag{6.6}$$

6.2 Functions of Continuous Random Variables

Suppose X is a continuous random variable with pdf $f(x)$ and CDF $F(x)$. The functions of X are defined as follows:

Expected Value

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Percentile if $0 < p < 1$, then a $100\times$ p-th percentile of the distribution of X is a solution x_p such that

$$F(x_p) = p$$

[1] Recall that $F(x)$ is the cumulative distribution function.

Median 50th percentile: $x_{0.5}$

$$F(x_{0.5}) = 0.5$$

Mode if the pdf has a **unique** maximum at $x = m_0$, then m_0 is called the mode of X

Variance

$$Var(X) = \int_{-\infty}^{\infty} (x - E(X))^2 f(x) dx$$

6.3 Uniform Random Variables

X is a uniform random variable on the interval $[l, h]$ (denoted $X \sim U[l, h]$) if the probability of a realization in any subinterval of $[l, h]$ is proportional to the length of the interval. It's density is given by

$$f(x) = \begin{cases} \frac{1}{h-l} & \text{if } x \in [l, h] \\ 0 & \text{otherwise} \end{cases}$$

Some Facts If $X \sim U[l, h]$, then

1.

$$\begin{aligned} F(x) &= \int_l^x f(u) du \\ &= \int_l^x \frac{1}{h-l} du \\ &= \frac{u}{h-l} \Big]_l^x \\ &= \frac{x}{h-l} - \frac{l}{h-l} \\ &= \frac{x-l}{h-l} \end{aligned}$$

2.

$$\begin{aligned} E(X) &= \int_l^h uf(u)du \\ &= \int_l^h u\frac{1}{h-l}du \\ &= \frac{1}{2}\frac{u^2}{h-l}\Big]_l^h \\ &= \frac{h^2-l^2}{2(h-l)} \\ &= \frac{(h+l)(h-l)}{2(h-l)} = \frac{h+l}{2} \end{aligned}$$

3.

$$Var(X) = \frac{(h-l)^2}{12}$$

Since

$$\begin{aligned} E(X^2) &= \int_l^h \frac{u^2}{h-l}du \\ &= \frac{1}{3}\frac{u^3}{h-l}\Big]_l^h \\ &= \frac{h^3-l^3}{3(h-l)} \\ &= \frac{(h-l)(h^2+hl+l^2)}{3(h-l)} = \frac{h^2+hl+l^2}{3}, \end{aligned}$$

$$\begin{aligned} Vat(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{h^2+hl+l^2}{3} - \left(\frac{h+l}{2}\right)^2 \\ &= \frac{(h-l)^2}{12} \end{aligned}$$

4. $x_{0.5} = E(X)$ (Median=Expectation)

$$\frac{1}{2} = \int_l^{x_{0.5}} \frac{1}{h-l}du = \frac{u}{h-l}\Big]_l^{x_{0.5}} = \frac{x_{0.5}-l}{h-l}$$

Thus,

$$x_{0.5} = \frac{h+l}{2} = E(X)$$

5. If $X \sim U[0,1]$, then

$$F(x) = P(X \leq x) = \frac{x-0}{1-0} = x$$

6.3.1 Invariance under Shifting and Scaling

1. If $X \sim U[0,1]$, and $Y = aX + b$ with $a > 0$, then

$$Y \sim U[b, a+b]$$

proof.

$$\begin{aligned} & P(X \leq x) = x \\ \iff & P(aX + b \leq ax + b) = x \\ \iff & P(Y \leq y) = x = \frac{y-b}{a} = \frac{y-b}{(a+b)-b} \\ \iff & Y \sim U[b, a+b] \end{aligned}$$

2. If $X \sim U[0,1]$, and $W = (h-l)X + l$, then

$$W \sim U[l, h]$$

3. If $W \sim U[l,h]$, and $Z = aW + b$ with $a > 0$, then

$$Z \sim U[al + b, ah + b]$$

proof.

$$Z = aW + b = \underbrace{a(h-l)}X + \underbrace{al+b} \sim U[al+b, a(h-l)+al+b] = U[al+b, ah+b]$$

Exercise 14. If $a < 0$, show that

1. If $X \sim U[0,1]$, and $Y = aX + b$ with $a < 0$, then

$$Y \sim U[a+b, b]$$

2. If $X \sim U[0,1]$, and $W = (l-h)X + h$, then

$$W \sim U[l, h]$$

3. If $W \sim U[l,h]$, and $Z = aW + b$ with $a < 0$, then

$$Z \sim U[ah + b, al + b]$$

6.3.2 A Revisit of General Uniform Random Variables

Let $X \sim U[0,1]$ with $E(X) = \frac{1}{2}$ and $Var(X) = \frac{1}{12}$. Hence, by Invariance Property under Shifting and Scaling any general uniform random variable $W \sim U[l,h]$ could be rewritten as

$$(1) \quad W = (h-l)X + l$$

or

$$(2) \quad W = (l-h)X + h$$

Use (1), we can easily obtain

$$E(W) = (h-l)E(X) + l = \frac{h-l}{2} + l = \frac{h+l}{2}$$

$$Var(W) = (h-l)^2 Var(X) = \frac{(h-l)^2}{12}$$

The same results can be reached by using (2).

6.4 The Exponential Distribution

Definition: The random variable T has an exponential distribution with **rate** $\lambda > 0$ if its density is described as follows

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & \text{if } t \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

denoted

$$T \sim exp(\lambda)$$

Remarks Note that $f(0) = \lambda e^0 = \lambda$. (The intercept $= \lambda$)

6.4.1 Some Properties

1.

$$\begin{aligned} P(T \le t) &= \int_0^t \lambda e^{-\lambda u} du \\ &= -e^{-\lambda u}\Big]_0^t \\ &= -e^{-\lambda t} - (-1) = 1 - e^{-\lambda t} \end{aligned}$$

2. $P(T > t) = 1 - P(T \le t) = e^{-\lambda t}$

3. $\int_0^\infty f(t)dt = \int_0^\infty \lambda e^{-\lambda t} dt = -e^{-\lambda t}\Big]_0^\infty = 0 - (-1) = 1$

4. Invariance under Positive Scaling

   If $W \sim exp(1)$ and $T = \frac{1}{\lambda} W$ with $\lambda > 0$, then $T \sim exp(\lambda)$.

   [proof]

$$\begin{aligned} & P(W > w) = e^{-w} \\ \iff & P(\lambda T > \lambda t) = e^{-w} \\ \iff & P(T > t) = e^{-\lambda t} \\ \iff & T \sim exp(\lambda) \end{aligned}$$

5. If $T \sim exp(\lambda)$, then $E(T) = \frac{1}{\lambda}$, $Var(T) = \frac{1}{\lambda^2}$.

   proof.

   Let $W \sim exp(1)$

$$E(W) = \int_0^\infty w e^{-w} dw$$

   Recall the Integration by Parts: $\int udv = uv - \int vdu$ Let $v = -e^{-w}$, $u = w$.

$$\begin{aligned} E(W) &= w(-e^{-w})\Big]_0^\infty - \int_0^\infty -e^{-w} dw \\ &= (0 - 0) - e^{-w}\Big]_0^\infty = 0 - (0 - 1) = 1 \end{aligned}$$

$$E(W^2) = \int_0^\infty w^2 e^{-w} dw$$

Let $v = -e^{-w}$, $u = w^2$.

$$\begin{aligned} E(W^2) &= w^2(-e^{-w})\Big]_0^\infty - \int_0^\infty -e^{-w}2wdw \\ &= (0-0) + 2\underbrace{\int_0^\infty we^{-w}dw}_{E(W)} \\ &= 2E(W) \end{aligned}$$

Thus,

$$Var(W) = E(W^2) - E(W)^2 = 2E(W) - [E(W)]^2 = 2 - 1 = 1$$

Let $T = \frac{1}{\lambda}W$. By Invariance under Positive Scaling,

$$T \sim exp(\lambda),$$

and

$$\begin{aligned} E(T) = E(\frac{1}{\lambda}W) = \frac{1}{\lambda}E(W) = \frac{1}{\lambda}, \\ Var(T) = Var(\frac{1}{\lambda}W) = \frac{1}{\lambda^2}Var(W) = \frac{1}{\lambda^2} \end{aligned}$$

Clearly, λ is the inverse of **expected life**:

$$\lambda = \frac{1}{E(T)}$$

6. No-Memory

$$P(T > m + n|T > m) = P(T > n)$$

7. Relationship between the Poisson and Exponential Distribution

6.5 The Normal Distribution

The random variable X is normal with mean μ and variance σ^2 (denoted as $X \sim N(\mu, \sigma^2)$) if its pdf is given by the bell shaped curve

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

with $-\infty < x < \infty$.

Standard Normal Distribution If $Z \sim N(0,1)$, we call it a standard normal random variable with density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

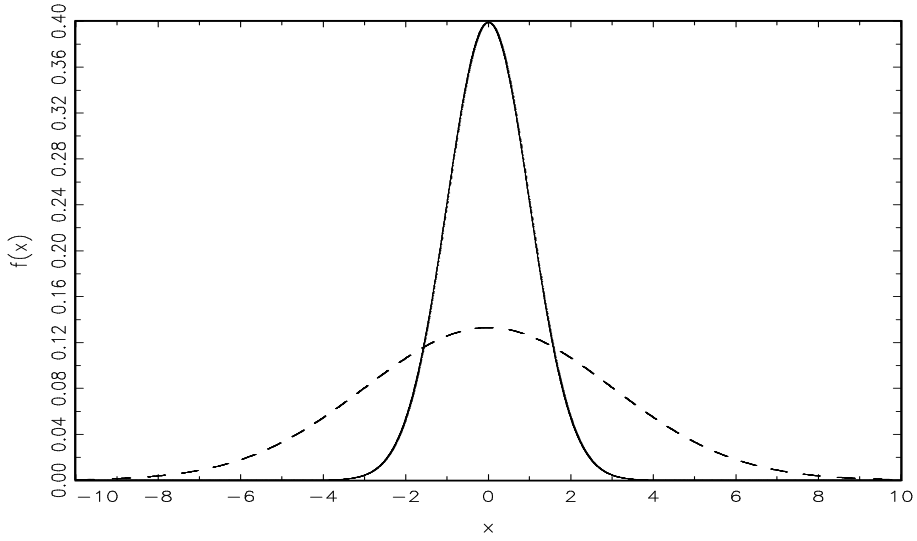
Some Properties of Normal Distributions

1. Bell shaped curve
2. Concave to convex at $\pm\sigma$
3. Highest point at mean (mode=mean)
4. Area under curve is 1
5. Symmetric about mean $\Rightarrow$ area on each side of mean is $\frac{1}{2}$ (median=mean)
6. Tails go out to $\pm\infty$
7. Changing μ shifts curve right or left
   (a) $\sigma^2 \uparrow \Rightarrow$ flattens curve
   (b) $\sigma^2 \downarrow \Rightarrow$ make it taller
8. If $X \sim N(\mu, \sigma^2)$
   (a) $P(X \in \mu \pm \sigma) = 0.6826 \approx \frac{2}{3}$
   (b) $P(X \in \mu \pm 2\sigma) = 0.9554$
   (c) $P(X \in \mu \pm 3\sigma) = 0.9972$

Two Important Facts about Normal Random Variables

1. If $X \sim N(\mu, \sigma^2)$, then $aX + b \sim N(a\mu + b, a^2\sigma^2)$ for $a \neq 0$.
2. If $\{X_i\}_{i=1}^n$ are normal and independent, then $\sum_{i=1}^n X_i$ is normal

Figure 6.1: Normal Distribution (with $\sigma = 1.0$ and $\sigma = 3.0$)



Computing Probability for Normal Random Variables For standard normal random variables $Z \sim N(0, 1)$, we have tables which tell us probabilities of the form $P(0 \leq Z \leq a)$, for $a > 0$. To figure out other sorts of probabilities, we use symmetry properties:

1. $P(Z \leq 0) = P(Z \geq 0) = 0.5$
2. $P(Z \leq -a) = P(Z \geq a)$
3. $P(-a \leq Z \leq 0) = P(0 \leq Z \leq a)$

Exercise 15. Compute the following probabilities.

1. $P(0 \leq Z \leq 1)$
2. $P(-2 \leq Z \leq 1)$
3. $P(-2.5 \leq Z \leq -0.5)$

An Important Property $X \sim N(\mu, \sigma^2) \Rightarrow X - \mu \sim N(0, \sigma^2) \Rightarrow \frac{X-\mu}{\sigma} \sim N(0, 1)$ Therefore, all probability statements about normal random variables can be converted to statements about standard normal random variables. For instance, suppose $X \sim N(5, 64)$, what is $P(X \geq 17)$?

$$\begin{aligned} P(X \geq 17) &= P(X - 5 \geq 12) \\ &= P(\frac{X-5}{8} \geq 1.5) \\ &= P(Z \geq 1.5) \\ &= P(Z \geq 0) - P(0 \leq Z \leq 1.5) \\ &= 0.5 - 0.4332 = 0.0668 \end{aligned}$$

6.6 The Continuity Correction

Sometimes we use a normal random variable X to **approximate** a random variable D which is actually discrete (typically integer valued). In this case, we approximate $P(D = d)$ by $P(d - 0.5 \leq X \leq d + 0.5)$. For example, $D \approx X \sim N(150, 100)$. What are $P(D = 150)$ and $P(160 \leq D \leq 175)$?

$$P(D = 150) = P(149.5 \leq X \leq 150.5)$$

$$P(160 \leq D \leq 175) = P(159.5 \leq X \leq 175.5)$$

6.7 Multivariate Continuous Random Variables

Joint Continuous Distribution A k-dimensional vector random variable $\vec{X} = (X_1, \ldots, X_k)$ is said to be continuous if there is a function $f(x_1, x_2, \ldots, x_k)$, called the *joint probability density function* (joint pdf), of $\vec{X}$, such that the joint CDF can be written as

$$F(x_1, \ldots, x_k) = \int_{-\infty}^{x_k} \cdots \int_{-\infty}^{x_1} f(u_1, \ldots, u_k) du_1 \cdots du_k$$

for all $\vec{x} = (x_1, \ldots, x_k)$.

As in the one-dimension case, joint pdf can be obtained from

$$f(x_1, x_2, \ldots, x_k) = \frac{\partial^k F(x_1, \ldots, x_k)}{\partial x_1 \cdots \partial x_k}$$

Further, any function $f(x_1, x_2, \ldots, x_k)$ is a joint pdf of a k-dimensional random variable iff.

$$f(x_1, x_2, \ldots, x_k) \geq 0 \ \ \forall x_1, x_2, \ldots, x_k$$
$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_k) dx_1 \cdots dx_k = 1$$

Marginal Density Function If the pair (X_1, X_2) of continuous random variables has joint pdf $f(x_1, x_2)$, then the marginal pdf's of X_1 and X_2 are

$$f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$$

$$f_2(x_2) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_1$$

Example: Let X and Y have the joint pdf

$$f(x, y) = 2, \quad 0 \leq x \leq y \leq 1$$

Hence,

$$f_X(x) = \int_x^1 2dy = 2y\Big]_x^1 = 2 - 2x = 2(1 - x), \quad 0 \leq x \leq 1$$
$$f_Y(y) = int_0^y 2dx = 2x\Big]_0^y = 2y, \quad 0 \leq y \leq 1$$

Independence Suppose the pair (X_1, X_2) of continuous random variables has joint pdf $f(x_1, x_2)$. X_1 and X_2 are independent if $f(x_1, x_2) = f_1(x_1) f_2(x_2)$

Conditional pdf If X_1 and X_2 are discrete **or** continuous random variables with joint pdf $f(x_1, x_2)$, then the conditional probability density function (conditional pdf) of X_2 given $X_1 = x_1$ is defined as

$$f(x_2|x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$$

for values x_1 such that $f_1(x_1) > 0$, and zero otherwise.

Similarly, the conditional pdf of X_1 given $X_2 = x_2$ is

$$f(x_1|x_2) = \frac{f(x_1, x_2)}{f_2(x_2)}$$

for values x_2 such that $f_2(x_2) > 0$, and zero otherwise.

Exercise 16. Show that if X_1 and X_2 are independent, then

$$f(x_2|x_1) = f_2(x_2)$$

$$f(x_1|x_2) = f_1(x_1)$$

Chapter 7

Special Topics: Moment Generating Functions and Transformations of Random variables

7.1 Support of a Random Variable

Definition The support of a random variable X, denoted by $supp(X)$, is the set of points where its density is positive.

$$supp(X) = \{x : f_X(x) > 0\}.$$

Note that

$$\sum_{x \in supp(X)} f(x) = 1, \quad \text{if } X \text{ discrete}$$

and

$$\int_{x \in supp(X)} f(x)dx = 1 \quad \text{if } X \text{ continuous}$$

Examples

X	$supp(X)$
Bernoulli	$\{0, 1\}$
Binomial	$\{0, 1, 2, 3, \ldots, n\}$
Poisson	$\{0, 1, 2, 3, \ldots\}$
Exponential	$[0, \infty)$
Normal	$(-\infty, \infty)$

7.2 The Moment Generating Function

Definition Let X be a random variable with (discrete) p.d.f. $f(x)$. If there is a positive number $h > 0$ such that

$$E(e^{tX}) = \begin{cases} \sum_{x \in supp(X)} e^{tx} f(x) = \sum_{x \in supp(X)} e^{tx} P(X = x) & X \text{ discrete} \\ \int_{x \in supp(X)} e^{tx} f(x) dx & X \text{ continuous} \end{cases}$$

exists and is finite for all t in $-h < t < h$, the the function of t defined by

$$M_X(t) = E(e^{tx})$$

is called the *moment generating function* (MGF) of X.

Example 1 $X \sim$ Bernoulli(p).

$$\begin{aligned} M_X(t) = E(e^{tx}) &= \sum_{x=0,1} e^{tx} P(X = x) \\ &= e^0(1-p) + e^t p \\ &= (1-p) + pe^t \end{aligned}$$

Exercise 17. $X \sim$ Binomial(n, p). Show that

$$M_X(t) = (pe^t + (1-p))^n.$$

Exercise 18. $X \sim$ Poisson(λ). Show that

$$M_X(t) = e^{\lambda(e^t - 1)}.$$

Exercise 19. $X \sim \exp(\lambda)$. Show that

$$M_X(t) = \frac{\lambda}{\lambda - t}, \ \ \lambda > t.$$

Exercise 20. $Z \sim$ N(0,1). Show that

$$M_Z(t) = e^{\frac{t^2}{2}}.$$

7.2.1 Some Facts

Fact 1: If $M_X(t) = M_Y(t)$ for all $t \in (-h, h)$, then X and Y have the same distribution. In the other wards, If the MGF exists, there is one and only one distribution associated with the MGF.

Fact 2:

$$E(X^n) = M_X^{(n)}(0) = M_X^{(n)}(t)|_{t=0},$$

where $M_X^{(n)}(t)$ denotes the n-th derivative of $M_X(t)$.

Examples

1. $Z \sim$ N(0,1), $M_Z(t) = e^{\frac{t^2}{2}}$.

$$M_Z'(0) = e^{\frac{t^2}{2}} t \Big]_{t=0} = 0$$

$$M_Z''(0) = e^{\frac{t^2}{2}} t^2 + e^{\frac{t^2}{2}} \Big]_{t=0} = 0 + 1 = 1$$

   Hence,

$$E(Z) = M_Z'(0) = 0$$

$$E(Z^2) = M_Z''(0) = 1$$

$$Var(Z) = E(Z^2) - E(Z)^2 = 1 - 0 = 1$$

2. $X \sim$ Poisson(λ), $M_X(t) = e^{\lambda(e^t - 1)}$.

$$E(X) = M_X'(0) = e^{\lambda(e^t-1)} \lambda e^t \Big]_{t=0} = \lambda$$

$$E(X^2) = M_X''(0) = (e^{\lambda(e^t-1)} \lambda e^t) \lambda e^t + e^{\lambda(e^t-1)} \lambda e^t \Big]_{t=0} = \lambda^2 + \lambda$$

Hence,

$$Var(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

3. $X \sim \exp(\lambda)$, $M_X(t) = \frac{\lambda}{\lambda - t}$.

$$E(X) = M_X'(0) = \frac{0 - \lambda \cdot (-1)}{(\lambda - t)^2} = \frac{\lambda}{(\lambda - t)^2}\Big]_{t=0} = \frac{1}{\lambda}$$

$$E(X^2) = M_X''(0) = \frac{0 - \lambda 2(\lambda - t)(-1)}{(\lambda - t)^4}\Big]_{t=0} = \frac{2\lambda(\lambda - t)}{(\lambda - t)^4}\Big]_{t=0} = \frac{2}{\lambda^2}$$

Hence,

$$Var(X) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$$

Fact 3: If $Y = aX + b$, then

$$M_Y(t) = e^{bt} M_X(at)$$

Exercise 21. Prove Fact 3.

Example Find out $M_X(t)$ when $X \sim$N(μ, σ^2).
Since $M_Z(t) = e^{\frac{t^2}{2}}$ with $Z \sim$N(0,1). Let

$$X = \sigma Z + \mu$$

By Fact 3, let $b = \mu$ and $a = \sigma$,

$$\begin{aligned} M_X(t) &= e^{\mu t} M_Z(\sigma t) \\ &= e^{\mu t} e^{\frac{\sigma^2 t^2}{2}} \\ &= e^{\mu t + \frac{\sigma^2 t^2}{2}} \end{aligned}$$

Fact 4: If X_1, X_2,..., X_n are independent with MGF $M_{X_1}(t)$, $M_{X_2}(t)$,..., $M_{X_n}(t)$ and if

$$Y = \sum_{i=1}^{n} X_i$$

then

$$M_Y(t) = M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t) = \Pi_{i=1}^{n} M_{X_i}(t)$$

Exercise 22. Prove Fact 4.

Exercise 23. Show that if $\{X_i\}_{i=1}^n \sim \text{BTP}(n,p)$ then $Y = \sum_{i=1}^n X_i \sim$Bernoulli$(n,p)$.

Exercise 24. Let $X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$ be independent random variables and $W = aX + bY$. Show that $W \sim N(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$.

Hence, the General case is

If

$$X_i \sim N(\mu_i, \sigma_i^2) \quad \text{independent},$$

and

$$W = a_1X_1 + a_2X_2 + \cdots a_nX_n$$

then

$$W \sim N(a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n, a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2).$$

7.3 Transformations of Random variables

In this section, we will discuss so called *one-to-one* transformation of random variables. Suppose that we have following information of a random variable X

- density function: $f_X(x)$
- distribution function: $F_X(x)$

If we have another random variable, Y, which is a one-to-one function of X. That is,

$$Y = u(X).$$

For instance, $Y = 7X$ or $Y = \ln X$. Since $u(\cdot)$ is one-to-one, we may rewrite X as

$$X = w(Y) = u^{-1}(Y).$$

For instance, $X = \frac{Y}{7}$ or $X = e^Y$.

Suppose that we would like to learn the density function of Y, $f_Y(y)$. There are two methods:

1. The CDF Technique (when $F_X(x)$ is known)

2. Transformation Method (when $F_X(x)$ is unknown)

7.3.1 The CDF Technique

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(u(X) \leq y) \\ &= P(X \leq u^{-1}(y)) \\ &= P(X \leq w((y)) \\ &= F_X(w(y)) \end{aligned}$$

Hence,

$$f_Y(y) = \frac{dF_Y}{dy}$$

Example

$$F_X(x) = 1 - e^{-2x}, \quad 0 < x < \infty$$

and

$$Y = e^X, \quad 1 < y < \infty$$

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(e^X \leq y) \\ &= P(X \leq \ln y) \\ &= F_X(\ln y) \\ &= 1 - e^{-2\ln y} \\ &= 1 - e^{\ln y^{-2}} \\ &= 1 - y^{-2} \end{aligned}$$

Hence,

$$f_Y(y) = \frac{dF_Y}{dy} = 2y^{-3}, \quad 1 < y < \infty$$

7.3.2 Transformation Methods

Theorem: Let X have p.d.f. $f_X(x)$ and let $Y = u(X)$, where $u(\cdot)$ is a one-to-one function. Then the p.d.f. of Y is given by

$$f_Y(y) = f_X(w(y)) \underbrace{\left| \frac{d}{dy} w(y) \right|}_{\text{Jacobian}}$$

Exercise 25. Prove the above Theorem.

Example:

$$f_X(x) = 2e^{-2x}, \quad 0 < x < \infty$$

and

$$Y = e^X, \quad 1 < y < \infty$$

Hence, $x = w(y) = \ln y$,

$$\begin{aligned} f_Y(y) &= f_X(w(y)) \left| \frac{d}{dy} w(y) \right| \\ &= f_X(\ln y) \left| \frac{d}{dy} \ln y \right| \\ &= f_X(\ln y) \left| \frac{1}{y} \right| \\ &= 2e^{-2 \ln y} \frac{1}{y} \\ &= 2e^{\ln y^{-2}} \frac{1}{y} \\ &= 2y^{-2} y^{-1} = 2y^{-3}, \quad 1 < y < \infty \end{aligned}$$

Chapter 8

Sampling and Asymptotic Theory

8.1 Sampling Theory

Population Any well defined set of objects about which a statistical enquiry is being made is called a population. The total number of objects in a population is known as the size of the population which may be finite or infinite. In general, it is difficult to study the whole population due to lack of resources (expensive, time-consuming). Thus, what we can do is to study a part or small section selected from the population, which is called a sample.

Random Samples The set of *random variables* $\{X_i\}_{i=1}^n$. is said to be a *random sample* of size n from a population with density function $f(x)$ if the joint pdf has the form

$$f(x_1, \ldots, x_n) = f(x_1) \cdots f(x_n)$$

That is, random sampling assumes that the sample is taken in such a way that the random variables for each trial are independent and follow the common population density function. So we also call the random samples as **i.i.d.** samples. The term "i.i.d." is referred to "**i**ndependently and **i**dentically **d**istributed".

Remark It is common practice to refer to the set of observed values, or data, $x_1, \ldots, x_n$ as a random sample. That is, although an observation is simply a measured attribute, conceptually it can be treated as a random variable because of the *uncertainty* in its value.

Thus, if we select a household at random and obtain its annual income in a given year, then we will treat the observed value (say x_1) as a random variable also because if we choose another household, we will obtain a different value, thus attesting to the **inherent random process**. The concept of "random" is **ex ante**.

8.2 Asymptotic Theory

8.2.1 Weak Law of Large Numbers

ex1. Betting on a Fair Coin $\{X_i\}_{i=1}^n$ i.i.d.

x	$P(X_i = x)$
1	1/2
-1	1/2

$E(X_i) = 0, \quad Var(X_i) = 1$

ex2. Betting on Roulette $\{X_i\}_{i=1}^n$ i.i.d.

x	$P(X_i = x)$
1	9/19
-1	10/19

$E(X_i) = -1/19 \approx -0.05263, \quad Var(X_i) = 360/361 \approx 0.9973$

Let

$$S_n^* = \sum_{i=1}^n X_i$$

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

Thus S_n^* is total winnings, and $\bar{X}_n$ is winnings per bet.

- **Question:** What happens to winnings per trial *in the long run*?
- **Answer:** Weak Law of Large Numbers (WLLN)

Limit of a Sequence Given a sequence of real numbers, $b_1, \ldots, b_n$. If there exists a real number b such that for every $\varepsilon > 0$, there exists an integer $N(\varepsilon)$ with the property that for all $n > N(\varepsilon)$, we have $|b_n - b| < \varepsilon$, then we say that b is the **limit** of the sequence $\{b_n\}$ and denote it as

$$\lim_{n\to\infty} b_n = b$$

Convergence in Probability The sequence of random variable X_n is said to *converge in probability* to the real number x if for each $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - x| > \varepsilon) = 0.$$

We write, $X_n \xrightarrow{p} x$

Thus it becomes less and less likely that the random variable $X_n - x$ lies outside the interval $(-\varepsilon, +\varepsilon)$.

An equivalent definition is:

Convergence in Probability II Given $\varepsilon > 0$ and $\delta > 0$, there exists $N(\varepsilon, \delta)$ such that

$$P(|X_n - x| > \varepsilon) < \delta$$

for all $n > N(\varepsilon, \delta)$.

Weak Law of Large Numbers Given $\{X_i\}_{i=1}^n$ i.i.d. sample, and $\bar{X}_n = \sum_{i=1}^n X_i$. Then

$$\bar{X}_n \xrightarrow{p} E(\bar{X}_n)$$

That is, the WLLN says that sample average must approach the mean.

Proof of Weak Law of Large Numbers

Convergence in Distribution Given a sequence of random variable X_n whose CDF is $F_n(x)$, and a CDF $F_X(x)$ corresponding to the random variable X, we say that X_n *converges in distribution* to X if

$$\lim_{n\to\infty} F_n(x) = F_X(x)$$

at all points x at which $F_X(x)$ is continuous. We denote it as

$$X_n \xrightarrow{d} X$$

The Central Limit Theorem (CLT) Let $\{X_i\}_{i=1}^n$ i.i.d. with $E(X_i) = \mu$, and $Var(X_i) = \sigma^2 < \infty$, then

$$Z_n = \frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}} \xrightarrow{d} N(0,1)$$

We can also write it as follows

$$\frac{\bar{X}_n - E(\bar{X}_n)}{\sqrt{Var(\bar{X}_n)}} \overset{A}{\sim} N(0,1)$$

or

$$\bar{X}_n - E(\bar{X}_n) \overset{A}{\sim} N(0, Var(\bar{X}_n))$$

or

$$\bar{X}_n \overset{A}{\sim} N(E(\bar{X}_n), Var(\bar{X}_n))$$

That is, suppose n is not too small, the sample average of i.i.d. random variables is approximately normally distributed with the mean and variance that it ought to have.

Important Facts

- (Continuous Map Theorem) If $X_n \xrightarrow{p} X$, then $g(X_n) \xrightarrow{p} g(X)$ where $g(\cdot)$ is a continuous function.
- If $X_n \xrightarrow{p} X$, then $X_n \xrightarrow{d} X$
- (Slutsky's Theorem) If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then
  1. $X_n + Y_n \xrightarrow{d} X + c$
  2. $X_n Y_n \xrightarrow{d} cX$
  3. $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$ for $c \neq 0$

  The proofs are beyond the scope of the course and omitted here.

8.3 Applications

8.3.1 The Hot Hand in Basketball

Readings

- Tversky and Gilovich, "The Cold Facts About the 'Hot Hand' in Basketball."
- Gilovich, Vallone, and Tversky, "The Hot Hand in Basketball: On the Misperception of Chance Sequence."

1. Fan Survey:
   - 91% agree that "A player has a better chance of making a shot after just made his last two or three shots than after having missed his last two or three shots."
   - 84% agree that "It is important to pass the ball to someone who has just made several shots in a row."
2. Field Goal Data: Philadelphia 76ers, 1980-1981 season

	Probability hit after
3 misses	0.56
2 misses	0.53
1 misses	0.54
1 hits	0.51
2 hits	0.50
3 hits	0.46

   Slight negative relationship, not statistically significant. No evidence of correlation between results of shots.
3. Free Throw Data: Shot in pairs. As well, no evidence of correlation between first shot and the probability of second shot.
4. Controlled Shooting Experiment: Cornell's men's and women's basketball teams

	Probability hit after
3 misses	0.45
2 misses	0.47
1 misses	0.47
1 hits	0.48
2 hits	0.49
3 hits	0.49

Slightly positive correlation, not statistically significant.

5. Conclusion: The statistical evidence very strongly supports the theory that **the players' shooting follows an i.i.d. process.** $\Rightarrow$ There is no hot hand.

Question: If there is no hard hand, why is belief in it so prevalent? Consider the following two sequences of heads and tails, each of which has 11 heads and 10 tails:

1. T T T H T T H T T H H H H H T H T H H H T
2. H T H T H T T T H H T H T H T T H H H T H

Which of these sequences look like it was created by a random process?

Explanation: "belief in the 'law of small numbers'." Recall that according to "law of large numbers," after a large enough number of tosses of a fair coin, the observed proportion of heads is near 1/2. The "law of small numbers" represents the **mistaken belief** that the proportion of heads should be near 1/2 in small sequences of tosses as well.

Thus, the "law of small numbers" explains

1. beliefs in the hot hands
2. the gambler's fallacy
3. beliefs in technological trading strategies (investment decisions based on past price movements)

8.3.2 Financial Market Efficiency

The Weak Form Efficiency The Efficient Market Hypothesis states that at any given time, stock prices fully reflect all available information. Most individuals that buy and sell stocks, do so under the assumption that the stocks they are buying are worth more than the price that they are paying, while stocks that they are selling are worth less than the selling price. But if markets are efficient and current prices fully reflect all information, then buying and selling stocks in an attempt to outperform the market will effectively be a game of chance rather than skill. **The Weak Form Efficiency** asserts that there is no information about future stock price changes contained in past stock prices. All past market prices and data are fully reflected in stock prices.[1]

Intuition: if investors could predict tomorrow price change today, the price would change today instead. If all information about price changes is utilized, price changes will be random.

Technical Trading People who base buy and sell decisions on past price movements are called technical analysts or chartists. They believe that weak form efficiency is false.

Comments The empirical studies on developed markets support the weak-form efficiency in general. However, the research results on emerging markets are controversial.[2]

A Random Walk Model of Stock Price Movements

$$P_t = P_{t-1} + X_t \tag{8.1}$$

[1]There are three forms of the efficient market hypothesis: (1) The "Weak" form as mentioned here. In other words, technical analysis is of no use. (2) The "Semi-strong" form: all publicly available information is fully reflected in stock prices. In other words, fundamental analysis is of no use. (3) The "Strong" form: all (public and private) information is fully reflected in stock prices. In other words, even insider information is of no use.

[2]Empirical studies of weak-form efficiency can be divided into two types of test: (a) Statistical tests for patterns in price movements over time. In general a number of studies suggest that successive price changes are unrelated. (b) Tests of trading strategies based on past price movements which can earn abnormal returns. In general, trading rules such as the filter rule are not found to generate large abnormal returns after taking account of risk and transaction costs.

where $\{X_t\}$ i.i.d. with $E(X_t) = 0$. That is, prices follow a **random walk**. Equation (**??**) states that the best forecast of the price of a security at time $t + 1$ is the price at time t, which in turn implies that the expected gain or loss for any holding period is zero. (i.e. $E(P_t - P_{t-1}) = E(X_t) = 0$)

Chapter 9

Statistics and Sampling Distributions

9.1 Statistics

Statistics A function of observable random variables, $T = t(X_1, \ldots, X_n)$, which does not depend on any unknown parameters, is called a *statistic*. For instance,

1. Sample Mean:
$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$$

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

3. Sample Standard Deviation: S

4. Sample Covariance:
$$S_{XY} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{n-1}$$

5. Sample Correlation Coefficient
$$r_{XY} = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n)}{\sqrt{\sum_{i=1}^n (X_i - \bar{X}_n)}\sqrt{\sum_{i=1}^n (Y_i - \bar{Y}_n)}}$$

Exercise 26. If $X_1, \ldots, X_n$ denotes a random sample of size n from $f(x)$ with $E(X) = \mu$, $Var(X) = \sigma^2$. Show that

$$E(\bar{X}_n) = \mu \tag{9.1}$$

$$Var(\bar{X}_n) = \frac{\sigma^2}{n} \tag{9.2}$$

$$E(S^2) = \sigma^2 \tag{9.3}$$

WLLN and CLT Revisited Since $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \frac{\sigma^2}{n}$, we have

$$\bar{X}_n \xrightarrow{p} \mu \tag{9.4}$$

$$\bar{X}_n \overset{A}{\sim} N(\mu, \frac{\sigma^2}{n}) \tag{9.5}$$

Use of CLT in Statistics $\{X_i\}$ i.i.d. indicators.

$$X_i = \begin{cases} 1 & \text{if vote for Gore} \\ 0 & \text{otherwise} \end{cases}$$

Therefore, $X_i \sim I(p)$ where p is proportion who plan to vote for Gore. Recall that $E(X_i) = p$, $Var(X_i) = p(1-p)$. Then according to CLT,

$$\bar{X}_n \overset{A}{\sim} N(p, \frac{p(1-p)}{n})$$

But we do not know p. How do we use the result of the poll to draw inferences about the actual proportion who plan to vote for Bush? That's **Statistic Inference**!

9.2 Sampling Distributions

A statistic is also a random variable, the distribution of which depends on the distribution of a random sample and on the form of the function $t(X_1, \ldots, X_n)$. The distribution of a statistic sometimes is referred to as a *sampling distribution*, in contrast to the population distribution.

9.2.1 Linear Combinations of Normal Variables

- (T1) If $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ denote n independent normal variables, then

$$\sum_{i=1}^{n} \alpha_i X_i \sim N(\sum_{i=1}^{n} \alpha_i \mu_i, \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2) \tag{9.6}$$

- (T2) If $X_1, \ldots, X_n$ denotes a random sample from $N(\mu, \sigma^2)$, then

$$\bar{X}_n \sim N(\mu, \frac{\sigma^2}{n}) \tag{9.7}$$

Exercise 27. Prove (T1) and (T2).

Exercise 28. What is the difference between equation (**??**) and equation (**??**)?

Exercise 29. Consider two independent random samples $X_1, \ldots, X_n$, and $Y_1, \ldots, Y_m$, with respective sample sizes n and m, from normally distributed populations, $X_i \sim N(\mu_X, \sigma_X^2)$ and $Y_i \sim N(\mu_Y, \sigma_Y^2)$. What is the distribution of $\bar{X}_n - \bar{Y}_m$?

9.2.2 Chi-Square Distribution

If $Z_1, \ldots, Z_k$ are **independent** $N(0, 1)$ variables, and $W = \sum_{i=1}^{k} Z_i^2$, then the pdf of W is

$$f(w) = \frac{\frac{1}{2}(\frac{w}{2})^{\frac{k}{2}-1}}{\Gamma(\frac{k}{2})} e^{-\frac{w}{2}}$$

for $w > 0$. Here $\Gamma(n)$ is the gamma function:

$$\Gamma(n) = \int_0^\infty y_{n-1} e^y dy,$$

where

$$\Gamma(1/2) = \sqrt{\pi}, \quad \Gamma(1) = 1, \quad \Gamma(n+1) = n\Gamma(n).$$

We denote the Chi-Square distribution as

$$W \sim \chi^2(k)$$

An important fact is

$$E(W^r) = \frac{2^r \Gamma(r + \frac{k}{2})}{\Gamma(\frac{k}{2})}$$

Hence, we have

$$E(W) = \sum_{i=1}^{k} E(Z_i^2) = k, \quad Var(W) = \sum_{i=1}^{k} Var(Z_i^2) = 2k.$$

The parameter k, traditionally called "the degree of freedom,"[1]

is simply the expectation of the variable W.

For future reference, we also report that

$$E\left(\frac{1}{W}\right) = \frac{1}{k-2} \quad \text{for } k > 2$$

$$E\left(\frac{1}{W^2}\right) = \frac{1}{(k-2)(k-4)} \quad \text{for } k > 4$$

Remarks

1. The derivation reverses: if $W \sim \chi^2(k)$, then W can be expressed as the sum of squares of k independent $N(0,1)$ variables.

2. (Additive Property) If $X \sim \chi^2(m)$, $Y \sim \chi^2(n)$, and X and Y are independent, then $X+Y \sim \chi^2(m+n)$. Thus the sum of independent chi-square is also chi-square with degree of freedom as the sum of the degree of freedom.

9.2.3 Student's t-distribution

If $Z \sim N(0,1)$, $W \sim \chi^2(k)$, with Z and W being **independent**, and

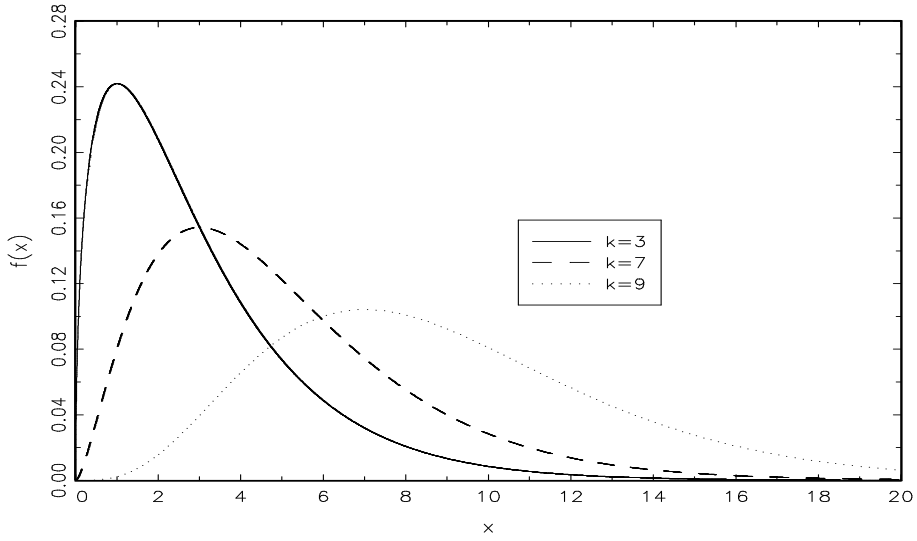
$$U = \frac{Z}{\sqrt{\frac{W}{k}}},$$

the the pdf of U is

$$\frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})} \frac{1}{\sqrt{k\pi}} \left(1 + \frac{u^2}{k}\right)^{-\frac{k+1}{2}}.$$

[1]Chi-square distributed random variables are sums of squares (or quadratic forms), and can be represented as the squared lengths of vectors. The dimension of the subspace in which the vector is free to roam is exactly the degrees of freedom. For examples, Let $Z_1, \ldots, Z_k$ be independent $N(0,1)$ variables, and let Z be the column vector whose ith element is Z_i. Then Z can roam all over Euclidean k-space. Its squared length, $Z'Z = Z_1^2 + \ldots + Z_n^2$, has a chi-square distribution with k degrees of freedom.

Figure 9.1: Chi-Squared Distribution (with $k = 3, 5, 9$)



This pdf defines the Student's t-distribution and we denote it as

$$U \sim t(k).$$

The parameter k is again called "the degree of freedom." The pdf is symmetric, centered at zero, and similar in shape to a standard normal pdf.

Remark The derivation reverses: if $U \sim t(k)$, then U can be expressed as

$$\frac{Z}{\sqrt{\frac{W}{k}}} = \frac{\sqrt{k}Z}{\sqrt{W}},$$

where $Z \sim N(0,1)$ is independent of $W \sim \chi^2(k)$.

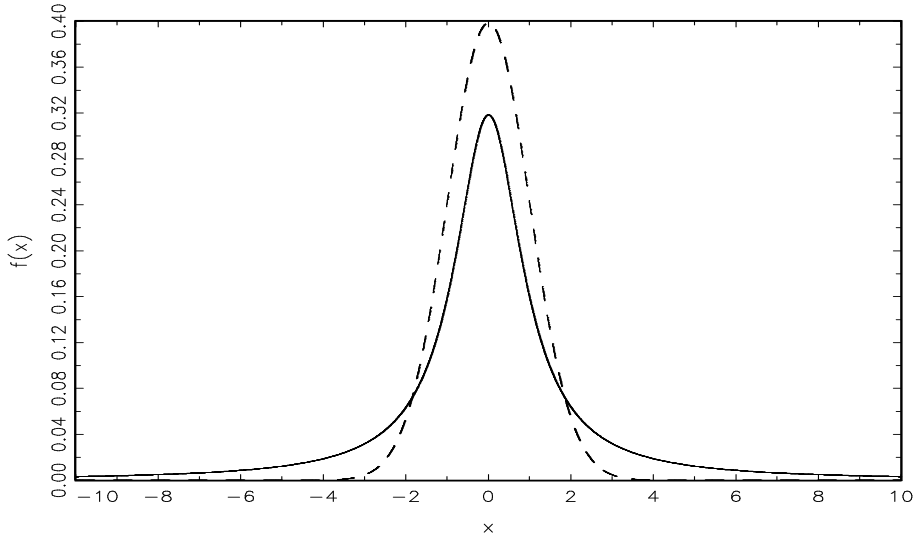
Exercise 30. Show that

$$E(U) = 0, \qquad Var(U) = \frac{k}{k-2} \quad \text{for } k > 2$$

We can find the fcat from Figure **??** that

$$t(k) \longrightarrow N(0,1) \text{ as } k \longrightarrow \infty$$

Figure 9.2: T Distribution (with $k = 1$ and $k = 300$)



9.2.4 F Distribution

If X_1 and X_2 are two **independent** chi-squared variables with degree of freedom parameters n_1 and n_2, respectively, then the ratio

$$F = \frac{X_1/n_1}{X_2/n_2} \sim F(n_1, n_2)$$

has the F distribution with n_1 and n_2 degrees of freedom. The pdf of F is given by

$$\frac{\Gamma(\frac{n_1+n_2}{2})}{\Gamma(\frac{n_1}{2})\Gamma(\frac{n_2}{2})} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} f^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2} f\right)^{-\frac{n_1+n_2}{2}}$$

Remarks:

1. If $X \sim F(n_1, n_2)$ and $Y = \frac{1}{X}$, then $Y \sim F(n_2, n_1)$.

2. $n_1 F(n_1, n_2) \to \chi^2(n_1)$ as $n_2 \to \infty$.

3. If $t \sim t(k)$, then $t^2 \sim F(1, k)$.

According to Figure **??**, it is clear that the F-distribution appears to be more symmetric as the degree of freedom increases.

Figure 9.3: F Distribution (with $n_1 = 10$ and $n_2 = 10$)

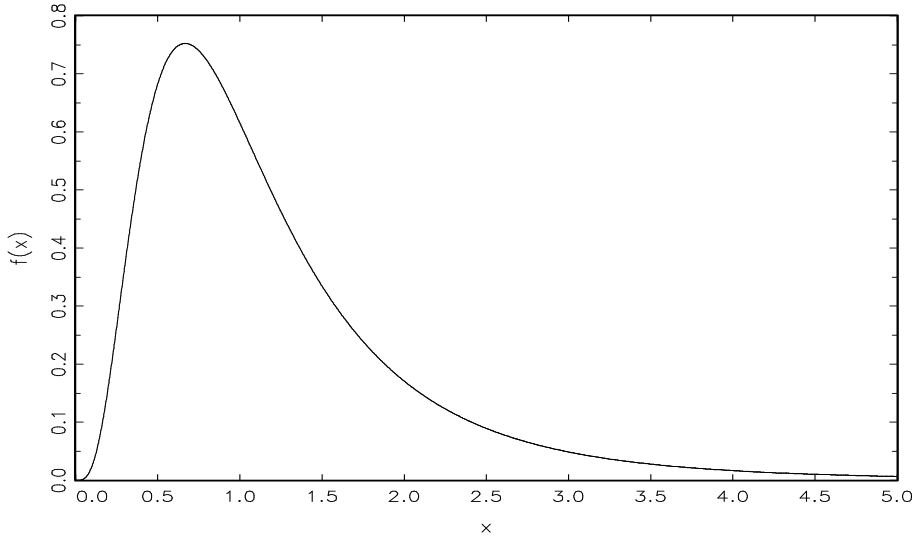
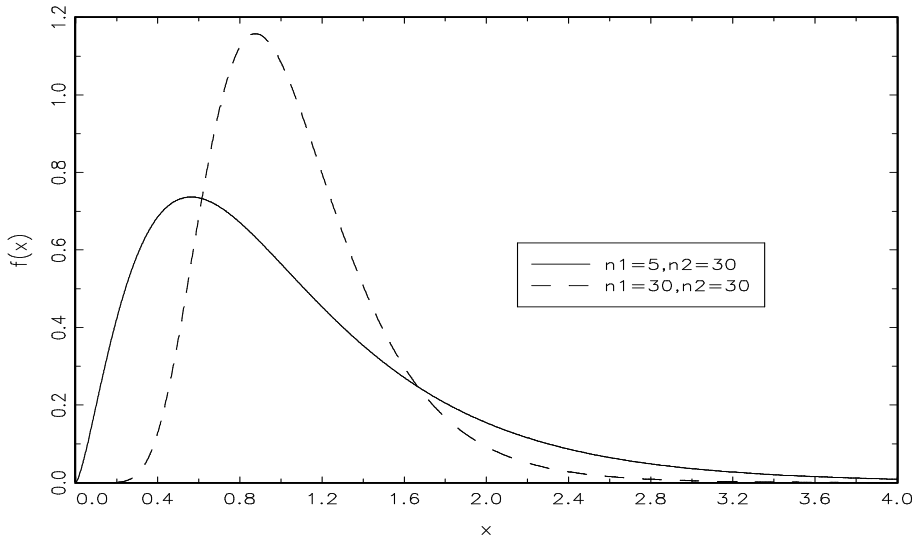


Figure 9.4: F Distribution (with $n_1 = 5, n_2 = 30$ and $n_1 = 30, n_2 = 30$)



9.3 Sampling from a Normal Population

Now suppose $X \sim N(\mu, \sigma^2)$, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, and $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$. We have

- (F1):

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \tag{9.8}$$

- (F2):

$$\bar{X} \text{ and } S^2 \text{ are independent} \tag{9.9}$$

- (F3):

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \tag{9.10}$$

- (F4):

$$\frac{\sqrt{n}(\bar{X}-\mu)}{S} \sim t(n-1) \tag{9.11}$$

Two Random Samples Let $U_1, \ldots, U_m$ and $V_1, \ldots, V_n$ be random sample from normal populations with variance σ_U^2 and σ_V^2, respectively. Then

- (F5)

$$\frac{S_U^2/\sigma_U^2}{S_V^2/\sigma_V^2} \sim F(m-1, n-1) \tag{9.12}$$

Part III

Proofs

Chapter 2

Proof: $P(A) + P(A^c) = 1$**.** Since A and A^c are disjoint,

$$\begin{aligned} P(A) + P(A^c) &= P(A \cup A^c) \quad \text{by (d)} \\ &= P(S) \\ &= 1. \end{aligned}$$

Proof: $A \subseteq B$ **implies that** $P(A) \leq P(B)$**.** First, $A \subseteq B$ implies that $B = A \cup (B - A)$. Since A and $B - A$ are disjoint, Thus, $P(B) = P(A \cup (B - A)) = P(A) + P(B - A) \geq P(A)$.

Proof: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$**.** Since $A \cup B = A \cup (B - A)$, where A, $(B - A)$ disjoint, $P(A \cup B) = P(A) + P(B - A)$.

Recall that

$$\begin{aligned} B &= B \cap S \\ &= B \cap (A \cup A^c) \\ &= (B \cap A) \cup (B \cap A^c) \\ &= (B \cap A) \cup (B - A) \end{aligned}$$

where $(B \cap A)$ and $(B - A)$ are disjoint. Thus, $P(B) = P(B \cap A) + P(B - A)$.

We then have

$$P(A \cup B) = P(A) + P(B - A) = P(A) + P(B) - P(A \cap B).$$

The Proof of Bayes' Rule By definition, whenever $P(A_i) > 0$,

$$P(T|A_i) = \frac{P(T \cap A_i)}{P(A_i)}$$

Therefore, we can get

$$P(T \cap A_i) = P(T|A_i)P(A_i) \tag{9.13}$$

Notice that since A_i form a partition of S (i.e. they are mutually exclusive and their union is S),

$$P(T) = P(T \cap S) = P(T \cap (A_1 \cup A_2 \cup \cdots \cup A_n)) = \sum_{i=1}^{n} P(T \cap A_i). \tag{9.14}$$

Therefore,

$$\begin{aligned} P(A_i|T) &= \frac{P(T \cap A_i)}{P(T)} \quad \text{by definition} \\ &= \frac{P(T \cap A_i)}{\sum_{i=1}^{n} P(T \cap A_i)} \quad \text{by equation (??)} \\ &= \frac{P(T \cap A_i)}{\sum_{i=1}^{n} P(T|A_i)P(A_i)} \quad \text{by equation (??)} \end{aligned}$$

■

Chapter 3

The Proof of $P(u(X) \geq c) \leq \frac{E[u(X)]}{c}$ First we show that if X is a random variable and $u(x)$ is a **nonnegative** real-valued function, then for any positive constant $c > 0$,

$$P(u(X) \geq c) \leq \frac{E[u(X)]}{c}. \tag{9.15}$$

proof: If $A = \{x|u(x) \geq c\}$, then

$$\begin{aligned} E[u(X)] &= \sum_{x} u(x)P(X = x) \\ &= \sum_{x \in A} u(x)P(X = x) + \sum_{x \in A^c} u(x)P(X = x) \\ &\geq \sum_{x \in A} u(x)P(X = x) \\ &\geq \sum_{x \in A} cP(X = x) \\ &= cP(X \in A) \\ &= cP(u(x) \geq c) \end{aligned}$$

The Proof of Markov Inequality If $u(x) = |x|^r$ for $r > 0$, then using equation (**??**):

$$\begin{aligned} &P(|x|^r \geq c) \\ &= P(|x| \geq c^{1/r}) \leq \frac{E[|X|^r]}{c} \end{aligned}$$

Let $c = k^r$, we obtain equation (**??**), the Markov Inequality. ■

The Proof of Chebychev Inequality Let $u(X) = (X - \mu)^2$, and $c = k^2\sigma^2$. Using equation (**??**) gives us equation (**??**), the Chebychev Inequality. ■

Proof: $-1 \leq \rho_{XY} \leq 1$ Since for any λ,

$$E[((x - \mu_X) - \lambda(y - \mu_Y))^2] \geq 0 \tag{9.16}$$

We can rewrite equation (**??**) as:

$$\begin{aligned} E[((x - \mu_X) - \lambda(y - \mu_Y))^2] &= E[(x - \mu_X)^2] + \lambda^2 E[(y - \mu_Y)^2] - 2\lambda E(x - \mu_X)(y - \mu_Y) \\ &= Var(X) + \lambda^2 Var(Y) - 2\lambda Cov(X, Y) \geq 0. \end{aligned}$$

Let

$$\lambda = \frac{Cov(X, Y)}{Var(Y)},$$

we have

$$Var(X) + \frac{(Cov(X, Y))^2}{Var(Y)} - 2\frac{(Cov(X, Y))^2}{Var(Y)} \geq 0.$$

That is

$$Var(X) \geq \frac{(Cov(X, Y))^2}{Var(Y)},$$

or

$$1 \geq \frac{(Cov(X, Y))^2}{Var(X)Var(Y)}.$$

The proof is complete by definition of ρ_{XY}. ■

Proof: Basic Properties of Random Variables Note: The proofs for equations (**??**), (**??**), and (**??**) through (**??**) are given first since they are useful in many of the other proofs.

(**??**) $E(aX + b) = aE(X) + b$

Proof: Let $Z = aX + b$, $z = ax + b$

$$\begin{aligned} E(aX+b) = E(Z) &= \sum_z zP(Z=z) \\ &= \sum_z zP(aX+b=z) \\ &= \sum_z zP(X=\frac{z-b}{a}) \\ &= \sum_x (ax+b)P(X=x) \\ &= a\sum_x xP(X=x) + b\sum_x P(X=x) \\ &= aE(X)+b \end{aligned}$$

■

(**??**) $E(X+Y) = E(X) + E(Y)$

Proof:

$$\begin{aligned} E(X+Y) = &= \sum_x \sum_y (x+y)P(X=x,Y=y) \\ &= \sum_x \sum_y xP(X=x,Y=y) \sum_x \sum_y yP(X=x,Y=y) \\ &= \sum_x x\sum_y P(X=x,Y=y) + \sum_y y\sum_x P(X=x,Y=y) \\ &= \sum_x xP(X=x) + \sum_y yP(Y=y) \\ &= E(X)+E(Y) \end{aligned}$$

■

(**??**) $Cov(X,X) = Var(X)$

Proof:

$$Cov(X,X) = E(X-E(X))(X-E(X)) = E(X-E(X))^2 = Var(X)$$

■

(**??**) $Cov(X, c) = 0$

Proof:

$$Cov(X, c) = E(X - E(X))(c - E(c)) = E(X - E(X))(c - c) = 0E(X - E(X)) = 0$$

■

(**??**) $Cov(X_1 + X_2, Y_1 + Y_2) = Cov(X_1, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_1) + Cov(X_2, Y_2)$

Proof: Let $\hat{X_1} = X_1 - E(X_1)$, $\hat{X_2} = X_2 - E(X_2)$, $\hat{Y_1} = Y_1 - E(Y_1)$, $\hat{Y_2} = Y_2 - E(Y_1)$

$$\begin{aligned} Cov(X_1 + X_2, Y_1 + Y_2) &= E((X_1 + X_2) - E(X_1 + X_2))((Y_1 + Y_2) - E(Y_1 + Y_2)) \\ &= E(\hat{X_1} + \hat{X_2})(\hat{Y_1} + \hat{Y_2}) \\ &= E(\hat{X_1}\hat{Y_1} + \hat{X_1}\hat{Y_2} + \hat{X_2}\hat{Y_1} + \hat{X_2}\hat{Y_2}) \\ &= E(\hat{X_1}\hat{Y_1}) + E(\hat{X_1}\hat{Y_2}) + E(\hat{X_2}\hat{Y_1}) + E(\hat{X_2}\hat{Y_2}) \\ &= Cov(X_1, Y_1) + Cov(X_1, Y_2) + Cov(X_2, Y_1) + Cov(X_2, Y_2) \end{aligned}$$

■

(**??**) $Cov(X, Y) = E(XY) - E(X)E(Y)$

Proof:

$$\begin{aligned} Cov(X, Y) &= E(X - E(X))(Y - E(Y)) \\ &= E(XY - XE(Y) - YE(X) + E(X)E(Y)) \\ &= E(XY) - E(Y)E(X) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

■

(**??**) $Var(X) = E(X^2) - [E(X)]^2$

Proof:

$$Var(X) = Cov(X, X) = E(XX) - E(X)E(X) = E(X^2) - [E(X)]^2$$

■

(**??**) $Var(aX+b) = a^2 Var(X)$

Proof:

$$\begin{aligned} Var(aX+b) &= E((aX+b)^2) - [E(aX+b)]^2 \\ &= E(a^2X^2 + 2abX + b^2) - [aE(X)+b]^2 \\ &= (a^2E(X^2) + 2abE(X) + b^2) - (a^2[E(X)]^2 + 2abE(X) + b^2) \\ &= a^2(E(X^2) - [E(X)]^2) \\ &= a^2 Var(X) \end{aligned}$$

■

(**??**) $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$

Proof:

$$\begin{aligned} Var(X+Y) &= E((X+Y)^2) - [E(X+Y)]^2 \\ &= E(X^2 + 2XY + Y^2) - [E(X) + E(Y)]^2 \\ &= (E(X^2) + 2E(XY) + E(Y^2)) - ([E(X)]^2 + 2E(X)E(Y) + [E(Y)]^2) \\ &= (E(X^2) - [E(X)]^2) + (E(Y^2) - [E(Y)]^2) + 2(E(XY) - E(X)E(Y)) \\ &= Var(X) + Var(Y) + 2Cov(X,Y) \end{aligned}$$

■

(**??**) $Cov(aX+b, cY+d) = acCov(X,Y)$

Proof:

$$\begin{aligned} &Cov(aX+b, cY+d) = E(aX+b)(cY+d) - E(aX+b)E(cY+d) \\ &= E(acXY + adX + bcY + bd) - (aE(X)+b)(cE(Y)+d) \\ &= acE(XY) + adE(X) + bcE(Y) + bd - (acE(X)E(Y) + adE(X) + bcE(Y) + bd) \\ &= ac(E(XY) - E(X)E(Y)) \\ &= acCov(X,Y) \end{aligned}$$

■

(**??**) $E(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} E(X_i)$

(**??**) $Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j) = \sum_{i=1}^{n} Var(X_i) + 2 \sum_{i=1}^{n} \sum_{j=1}^{i-1} Cov(X_i, X_j)$

Proof: These statements are proved by repeatedly applying (**??**) and (**??**). ■

(**??**) $X \perp Y \Rightarrow E(XY) = E(X)E(Y)$

Proof:

$$\begin{aligned} E(XY) &= \sum_x \sum_y xyP(X = x, Y = y) \\ &= \sum_x \sum_y xyP(X = x)P(Y = y) \\ &= (\sum_x xP(X = x))(\sum_y yP(Y = y)) \\ &= E(X)E(Y) \end{aligned}$$

■

(**??**) $X \perp Y \Rightarrow Cov(X, Y) = 0$

Proof:

$$Cov(X, Y) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0$$

■

(**??**) $X \perp Y \Rightarrow Var(X + Y) = Var(X) + Var(Y)$

Proof:

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y) = Var(X) + Var(Y)$$

■

Chapter 4

Proof: Naive Estimate Overstate Risk To begin, observe that

$$Var_A = p_1\sigma_1^2 + p_2\sigma_2^2 \tag{9.17}$$

$$\begin{aligned}(SD_A)^2 &= (p_1\sigma_1 + p_2\sigma_2)^2 \\ &= p_1^2\sigma_1^2 + p_2^2\sigma_2^2 + 2p_1p_2\sigma_1\sigma_2 \end{aligned} \tag{9.18}$$

$$Var(R_p) = p_1^2\sigma_1^2 + p_2^2\sigma_2^2 + 2\rho_{12}p_1p_2\sigma_1\sigma_2 \tag{9.19}$$

First we prove that

$$Var(R_p) < (SD_A)^2. \tag{9.20}$$

Note that

$$(SD_A)^2 - Var(R_p) = (1-\rho_{12})(2p_1p_2\sigma_1\sigma_2)$$

Since $p_1, p_2, \sigma_1\sigma_2$ and $1-\rho_{12}$ are all strictly positive, we see that $(SD_A)^2 - Var(R_p) > 0$.
Second, we prove that

$$(SD_A)^2 \leq Var_A. \tag{9.21}$$

$$\begin{aligned}Var_A - (SD_A)^2 &= (p_1 - p_1^2)\sigma_1^2 + (p_2 - p_2^2)\sigma_2^2 - 2p_1p_2\sigma_1\sigma_2 \\ &= (p_1 - p_1^2)\sigma_1^2 + ((1-p_1) - (1-p_1)^2)\sigma_2^2 - 2p_1(1-p_1)\sigma_1\sigma_2 \\ &= p_1(1-p_1)(\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2) \\ &= p_1(1-p_1)(\sigma_1 - \sigma_2)^2 \geq 0.\end{aligned}$$

Where we use $p_2 = 1 - p_1$. That is, we have shown that $Var_A - (SD_A)^2 \geq 0$.
According to equations (**??**) and (**??**), the proof is complete. ■

Proof: Backward Bending Since $p_2 = 1 - p_1$, $Var(R_p)$ can be rewritten as

$$Var(R_p) = p_1^2\sigma_1^2 + (1-p_1)^2\sigma_2^2 + 2\rho_{12}p_1(1-p_1)\sigma_1\sigma_2.$$

Differentiate with respect to p_1:

$$\frac{dVar(R_p)}{dp_1} = 2p_1\sigma_1^2 - 2(1-p_1)\sigma_2^2 + 2\rho_{12}\sigma_1\sigma_2(1-2p_1).$$

Evaluate at $p_1 = 1$ to obtain

$$\left.\frac{dVar(R_p)}{dp_1}\right|_{p_1=1} = 2\sigma_1^2 - 2\rho_{12}\sigma_1\sigma_2. \tag{9.22}$$

The interpretation of equation (**??**) is as follows: suppose that we start with a portfolio consisting solely of asset 1 and the slightly reduce holdings in asset 1 (spending the extra money on asset 2), the variance of the portfolio will decrease and the mean return increases (since $\mu_1 < \mu_2$) if $2\sigma_1^2 - 2\rho_{12}\sigma_1\sigma_2 > 0$. This is true precisely when $\rho_{12} < \frac{\sigma_1}{\sigma_2}$. ■

Proof: The Beta Relation

$$E(R_i) = r_f + \beta_i[E(R_m) - r_f]$$

That is,

$$E(R_i) = \begin{cases} E(R_m) & \beta_i = 1 \\ r_f + \beta_i[E(R_m) - r_f] & \beta_i \neq 0 \\ r_f & \beta_i = 0 \end{cases}$$

1. **Case 1**: $\beta_i = 1$

$$\beta_i = \frac{Cov(R_i, R_m)}{Var(R_m)} = 1 \Longrightarrow \underbrace{Cov(R_i, R_m) = Var(R_m)}_{(A)}$$

Consider $R = \alpha R_i + (1-\alpha)R_m$

$$E(R) = \alpha E(R_i) + (1-\alpha)E(R_m) \Longrightarrow E(R) - E(R_m) = \alpha[E(R_i) - E(R_m)]$$

By (A)

$$\begin{aligned} Var(R) &= (1-\alpha)^2 Var(R_m) + 2\alpha(1-\alpha)Cov(R_i, R_m) + \alpha^2 Var(R_i) \\ &= (1-\alpha^2)Var(R_m) + \alpha^2 Var(R_i) \end{aligned}$$

Since

$$\frac{dVar(R)}{d\alpha}\Big|_{\alpha=0} = 2\alpha(Var(R_i) - Var(R_m))|_{\alpha=0} = 0,$$

we have $Var(R) \approx Var(R_m)$.

That is, $Var(R)$ and $Var(R_m)$ are nearly the same give a tiny change (increase/decrease) of α.

- If $E(R_i) > E(R_m)$ and $\alpha > 0$ small, then $E(R) > E(R_m)$ with $Var(R) \approx Var(R_m)$, contradict to that R_m is efficient.
- If $E(R_i) < E(R_m)$ and $\alpha < 0$ small, then $E(R) > E(R_m)$ with $Var(R) \approx Var(R_m)$, contradict to that R_m is efficient.

Therefore, $E(R_i) = E(R_m)$ for $\beta_i = 1$.

2. **Case 2**: $\beta_i \neq 0$

   Consider

$$Z = \frac{1}{\beta_i}R_i + \left(1 - \frac{1}{\beta_i}\right) r_f$$

   Hence,

$$\frac{Cov(Z, R_m)}{Var(R_m)} = \frac{\frac{1}{\beta_i}Cov(R_i, R_m)}{Var(R_m)} = \frac{\frac{1}{\beta_i}Var(R_m)\beta_i}{Var(R_m)} = 1$$

   This is exactly **Case 1**. Therefore, $E(Z) = E(R_m)$ from **Case 1**. That is

$$\frac{1}{\beta_i}E(R_i) + \left(1 - \frac{1}{\beta_i}\right) r_f = E(R_m)$$

   Rearrange,

$$E(R_i) = r_f + \beta_i[E(R_m) - r_f]$$

3. **Case 3**: $\beta_i = 0$

$$\beta_i = 0 \Longrightarrow Cov(R_i, R_m) = 0$$

   Consider

$$Y = \frac{1}{2}R_i + \frac{1}{2}R_m$$

Since

$$Cov(Y, R_m) = Cov(\frac{1}{2}R_i + \frac{1}{2}R_m, R_m) = \frac{1}{2}Cov(R_i, R_m) + \frac{1}{2}Var(R_m) = \frac{1}{2}Var(R_m)$$

we know that $\beta_Y = \frac{1}{2}$ for Y (The same as **Case 2**).

Thus,

$$E(Y) = r_f + \frac{1}{2}[E(R_m) - r_f]$$

$$E\left[\frac{1}{2}R_i + \frac{1}{2}R_m\right] = r_f + \frac{1}{2}[E(R_m) - r_f]$$

$$E(R_i) = r_f \quad \text{for } \beta_i = 0.$$

Chapter 5

Proof of Poisson Distribution

$$\begin{aligned}\binom{n}{x}p^x(1-p)^{n-x} &= \frac{n(n-1)\cdots(n-x+1)}{x!}p^x(1-p)^{n-x} \\ &= \frac{n(n-1)\cdots(n-x+1)}{x!}\frac{\lambda^x}{n^x}\left(1-\frac{\lambda}{n}\right)^{n-x} \\ &= \frac{1(1-\frac{1}{n})(1-\frac{2}{n})\cdots(1-\frac{x-1}{n})}{x!}\lambda^x\left[\left(1-\frac{\lambda}{n}\right)^n\right]^{\frac{n-x}{n}}\end{aligned}$$

Since

$$\lim_{n\to\infty}(1-\frac{\lambda}{n})^n = e^{-\lambda},$$

we have

$$\binom{n}{x}p^x(1-p)^{n-x} \longrightarrow \frac{e^{-\lambda}\lambda^x}{x!} \quad \text{as} \quad n \to \infty, p \to 0.$$

Thus the pdf of Poisson Distribution is given by

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, 2, \ldots$$

■

Additionally,

$$\begin{aligned}\sum_{x=0}^{\infty} f(x;\lambda) &= \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots\right) \\ &= e^{-\lambda}e^{\lambda} \\ &= 1\end{aligned}$$

■

Chapter 8

Proof: Weak Law of Large Numbers Recall that $E(\bar{X}) = \mu$ and that $Var(\bar{X}) = \frac{\sigma^2}{n}$. Therefore, for any fixed $\varepsilon > 0$, Chebyshev Inequality (equation (**??**)) tells us that

$$P(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{Var(\bar{X})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2}.$$

Hence, as $n \to \infty$,

$$P(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

■

Chapter 16

Proof: E(MSE) and E(MSF) in ANOVA

$$\begin{aligned}\text{MSE} &= \frac{\sum_i \sum_j (y_{ij} - \bar{y}_i)^2}{N - k} \\ &= \frac{1}{N-k}[sum_j(y_{1j} - \bar{y}_1)^2 + \sum_j (y_{1j} - \bar{y}_1)^2 + \cdots + \sum_j (y_{kj} - \bar{y}_k)^2] \\ &= \frac{1}{N-k}[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2 + \cdots + (n_k - 1)S_k^2].\end{aligned}$$

Thus,

$$\begin{aligned}
E(\text{MSE}) &= \frac{1}{N-k}E[(n_1-1)S_1^2+(n_2-1)S_2^2+\cdots+(n_k-1)S_k^2] \\
&= \frac{1}{N-k}[(n_1-1)\sigma^2+(n_2-1)\sigma^2+\cdots+(n_k-1)\sigma^2] \\
&= \frac{\sigma^2}{N-k}[\sum_{i=1}^{k} n_i - k] \\
&= \frac{\sigma^2}{N-k}(N-k) \\
&= \sigma^2.
\end{aligned}$$

■

$$\text{MSF} = \frac{1}{k-1}\underbrace{(\sum_i \sum_j (\bar{y}_i - \bar{\bar{y}})^2)}_{(A)}$$

$$\begin{aligned}
(A) &= \sum_i \sum_j (\bar{y}_i - \bar{\bar{y}})^2 \\
&= \sum_i \sum_j [(\bar{y}-\mu) - (\bar{\bar{y}}-\mu)]^2 \\
&= \sum_i \sum_j (\bar{y}-\mu)^2 + \underbrace{\sum_i \sum_j (\bar{\bar{y}}-\mu)^2}_{(B)} \\
&\quad - 2\underbrace{\sum_i \sum_j (\bar{y}_i - \mu)(\bar{\bar{y}}-\mu)}_{(C)}
\end{aligned}$$

$$(B) = \sum_i \sum_j (\bar{\bar{y}}-\mu)^2 = (\bar{\bar{y}}-\mu)^2 \sum_i \sum_j 1 = (\bar{\bar{y}}-\mu)^2 \sum_i n_i.$$

$$
\begin{aligned}
(C) &= \sum_i \sum_j (\bar{y}_i - \mu)(\bar{\bar{y}} - \mu) \\
&= (\bar{\bar{y}} - \mu) \sum_i (\bar{y}_i - \mu) \sum_j 1 \\
&= (\bar{\bar{y}} - \mu) \sum_i (\bar{y}_i - \mu) n_i \\
&= (\bar{\bar{y}} - \mu) [\sum_i n_i \bar{y}_i - \mu \sum_i n_i] \\
&= (\bar{\bar{y}} - \mu) \sum_i n_i \left[\frac{\sum_i n_i \bar{y}_i}{\sum_i n_i} - \mu \right] \\
&= (\bar{\bar{y}} - \mu) \sum_i n_i (\bar{\bar{y}} - \mu) \\
&= (\bar{\bar{y}} - \mu)^2 \sum_i n_i.
\end{aligned}
$$

That is,

$$
\text{MSF} = \frac{1}{k-1} (\underbrace{\sum_i \sum_j (\bar{y}_i - \mu)^2}_{(D)} - (\bar{\bar{y}} - \mu)^2 \sum_i n_i) \tag{9.23}
$$

$$
\begin{aligned}
(D) &= \sum_i \sum_j (\bar{y}_i - \mu)^2 \\
&= \sum_i \sum_j [(\bar{y}_i - \mu_i) + (\mu_i - \mu)]^2 \\
&= \sum_i \sum_j (\bar{y}_i - \mu_i)^2 + \sum_i \sum_j (\mu_i - \mu)^2 + 2 \sum_i \sum_j (\bar{y}_i - \mu_i)(\mu_i - \mu).
\end{aligned}
$$

$$
\text{MSF} = \frac{1}{k-1} (\underbrace{\sum_i \sum_j (\bar{y}_i - \mu_i)^2}_{(i)} + \underbrace{\sum_i \sum_j (\mu_i - \mu)^2}_{(ii)} + 2 \underbrace{\sum_i \sum_j (\bar{y}_i - \mu_i)(\mu_i - \mu)}_{(iii)} - \underbrace{(\bar{\bar{y}} - \mu)^2 \sum_i n_i}_{(iv)}).
$$

Thus,

$$
E(\text{MSF}) = \frac{1}{k-1} (E(i) + E(ii) + 2E(iii) - E(iv)).
$$

$$
\begin{aligned}
E(i) &= \sum_i \sum_j E(\bar{y}_i - \mu_i)^2 \\
&= \sum_i \sum_j Var(\bar{y}_i) \\
&= \sum_i \sum_j \frac{\sigma_i^2}{n_i} \\
&= \sigma^2 \sum_i \sum_j \frac{1}{n_i} \\
&= \sigma^2 \sum_i \frac{1}{n_i} \sum_j 1 \\
&= \sigma^2 \sum_i \frac{1}{n_i} n_i \\
&= k\sigma^2.
\end{aligned}
$$

$$
\begin{aligned}
E(ii) &= \sum_i \sum_j E(\mu_i - \mu)^2 \\
&= \sum_i \sum_j (\mu_i - \mu)^2 \\
&= \sum_i (\mu_i - \mu)^2 \sum_j 1 \\
&= \sum_i n_i (\mu_i - \mu)^2.
\end{aligned}
$$

$$
\begin{aligned}
E(iii) &= \sum_i \sum_j E(\bar{y}_i - \mu_i)(\mu_i - \mu) \\
&= \sum_i \sum_j E(\bar{y}_i \mu_i - \mu_i^2 - \bar{y}_i \mu + \mu_i \mu) \\
&= \sum_i \sum_j (\mu_i^2 - \mu_i^2 - \mu \mu_i + \mu_i \mu) \\
&= 0.
\end{aligned}
$$

Since

$$y_i \sim N(\mu_i, \sigma^2),$$

we have

$$\bar{y}_i = \frac{\sum_j y_{ij}}{n_i} \sim N(\mu_i, \frac{\sigma^2}{n_i})$$

and

$$\bar{\bar{y}} = \frac{1}{\sum_i n_i} \sum_i n_i \bar{y}_i \sim N(\frac{\sum_i n_i \mu_i}{\sum_i n_i}, \frac{\sigma^2}{\sum_i n_i}).$$

Thus,

$$\begin{aligned} E(iv) &= \sum_i \sum_j E(\bar{\bar{y}} - \mu)^2 \\ &= \sum_i \sum_j \frac{\sigma^2}{\sum_i n_i} \\ &= \frac{\sigma^2}{\sum_i n_i} \sum_i \sum_j 1 \\ &= \sigma^2. \end{aligned}$$

Thus,

$$\begin{aligned} E(\text{MSF}) &= \frac{1}{k-1}(k\sigma^2 + \sum_i n_i(\mu_i - \mu)^2 - \sigma^2) \\ &= \frac{1}{k-1}[(k-1)\sigma^2 + \sum_i n_i(\mu_i - \mu)^2] \\ &= \sigma^2 + \frac{\sum_i n_i(\mu_i - \mu)^2}{k-1}. \end{aligned}$$

■

Chapter 17

Properties of LS Estimators

- The properties of $\hat{\beta}$ and α.

  Since

$$\begin{aligned} \hat{\beta} &= \frac{\sum_i (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} \end{aligned}$$

Then

$$\begin{aligned}
E(\beta) &= \frac{1}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x}) E(Y_i) \\
&= \frac{1}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x})(\alpha + \beta x_i) \\
&= 0 + \frac{1}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x}) \beta x_i \\
&= \frac{\beta}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x}) x_i \\
&= \frac{\beta}{\sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x})(x_i - \bar{x}) \\
&= \beta.
\end{aligned}$$

■

$$\begin{aligned}
Var(\hat{\beta}) &= Var\left(\frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i (x_i - \bar{x})^2} \right) \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \sum_i (x_i - \bar{x})^2 Var(Y_i) \\
&= \frac{1}{(\sum_i (x_i - \bar{x})^2)^2} \sum_i (x_i - \bar{x})^2 \sigma^2 \\
&= \frac{\sigma^2}{(\sum_i (x_i - \bar{x})^2)^2} \sum_i (x_i - \bar{x})^2 \\
&= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}
\end{aligned}$$

■

$$E(\hat{\alpha}) = E(\bar{Y} - \hat{\beta}\bar{x}) = \alpha + \beta\mu_x - \beta\mu_x = \alpha.$$

■

Notice that

$$\begin{aligned}\hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{x} \\ &= \frac{1}{n}\sum_i Y_i - \bar{x}\frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2} \\ &= \sum_i \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \right) Y_i\end{aligned}$$

$$\begin{aligned}Var(\hat{\alpha}) &= \sum_i \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \right)^2 Var(Y_i) \\ &= \sigma^2 \sum_i \left(\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \right)^2 \\ &= \sigma^2 \sum_i \left(\frac{1}{n^2} + \frac{\bar{x}^2 (x_i - \bar{x})^2}{(\sum_i (x_i - \bar{x})^2)^2} - \frac{2}{n}\frac{\bar{x}(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} - 0 \right) \\ &= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right).\end{aligned}$$

■

$$\begin{aligned}Cov(\hat{\alpha}, \hat{\beta}) &= Cov(\bar{Y} - \hat{\beta}\bar{x}, \hat{\beta}) \\ &= Cov(\bar{Y}, \hat{\beta}) - \bar{x} Cov(\hat{\beta}, \hat{\beta}) \\ &= Cov(\frac{\sum_i Y_i}{n}, \frac{\sum_i (x_i - \bar{x})Y_i}{\sum_i (x_i - \bar{x})^2}) - \bar{x} Cov(\hat{\beta}, \hat{\beta}) \\ &= \frac{\sigma^2}{n \sum_i (x_i - \bar{x})^2} \sum_i (x_i - \bar{x}) - \bar{x} Cov(\hat{\beta}, \hat{\beta}) \\ &= 0 - \bar{x} Var(\hat{\beta}) \\ &= -\frac{\bar{x}\sigma^2}{\sum_i (x_i - \bar{x})^2}\end{aligned}$$

■

Alternatively, using

$$Var(\hat{\alpha} + \hat{\beta}) = Var(\hat{\alpha}) + Var(\hat{\beta}) + 2Cov(\hat{\alpha}, \hat{\beta})$$

and

$$\hat{\alpha} + \hat{\beta} = \bar{Y} - \hat{\beta}\bar{x} + \hat{\beta} = \bar{Y} + (1 - \bar{x})\hat{\beta}$$

the above results can be obtained by trivial algebra. ■

- The property of S^2.

$$S^2 = \frac{\sum_i \hat{e}_i^2}{n-2}$$

Recall that

$$Y_i = \alpha + \beta x_i + e_i = E(Y_i) + e_i.$$

Hence

$$\begin{aligned}
\sum_i e_i^2 &= \sum_i (Y_i - E(Y_i))^2 \\
&= \sum_i [(Y_i - \hat{Y}_i) + (\hat{Y}_i - E(Y_i))]^2 \\
&= \sum_i (Y_i - \hat{Y}_i)^2 + 2\sum_i (Y_i - \hat{Y}_i)(\hat{Y}_i - E(Y_i)) + \sum_i (\hat{Y}_i - E(Y_i))^2 \\
&= \sum_i \hat{e}_i^2 + 2\sum_i \hat{e}_i(\hat{Y}_i - E(Y_i)) + \sum_i [(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)x_i]^2 \\
&= \sum_i \hat{e}_i^2 + 2[\sum_i \hat{e}_i \hat{Y}_i - \sum_i \hat{e}_i(\alpha + \beta x_i)] + \sum_i [(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)x_i]^2
\end{aligned}$$

Since

$$\sum_i \hat{e}_i = 0, \quad \text{by equation (??)} \tag{9.24}$$

$$\sum_i \hat{e}_i x_i = 0, \quad \text{by equation (??)} \tag{9.25}$$

and

$$\sum_i \hat{e}_i \hat{Y}_i = 0, \quad \text{by equations (??) and (??)}, \tag{9.26}$$

we have

$$\sum_i e_i^2 = \sum_i \hat{e}_i^2 + \sum_i [(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)x_i]^2 \tag{9.27}$$

Now we have TWO ways to reach our goal.

Proof 1 Since

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x},$$
$$\alpha = \mu_Y - \beta\mu_x = \mu_Y - \beta\bar{x},$$

equation (**??**) can be rewritten as

$$\begin{aligned}
\sum_i e_i^2 &= \sum_i \hat{e}_i^2 + \sum_i [(\hat{\alpha} - \alpha) + (\hat{\beta} - \beta)x_i]^2 \\
&= \sum_i \hat{e}_i^2 + \sum_i [(\bar{Y} - \mu_Y) - (\hat{\beta} - \beta)\bar{x} + (\hat{\beta} - \beta)x_i]^2 \\
&= \sum_i \hat{e}_i^2 + \sum_i [(\bar{Y} - \mu_Y) - (\hat{\beta} - \beta)(x_i - \bar{x})]^2 \\
&= \sum_i \hat{e}_i^2 + \sum_i [(\bar{Y} - \mu_Y)^2 + (\hat{\beta} - \beta)^2(x_i - \bar{x})^2 - 2(\bar{Y} - \mu_Y)(\hat{\beta} - \beta)(x_i - \bar{x})] \\
&= \sum_i \hat{e}_i^2 + \sum_i (\bar{Y} - \mu_Y)^2 + (\hat{\beta} - \beta)^2 \sum_i (x_i - \bar{x})^2
\end{aligned}$$

That is,

$$\sum_i e_i^2 = \sum_i \hat{e}_i^2 + n(\bar{Y} - \mu_Y)^2 + (\hat{\beta} - \beta)^2 \sum_i (x_i - \bar{x})^2. \tag{9.28}$$

Taking expectation,

$$\underbrace{E(\sum_i e_i^2)}_{n\sigma^2} = E(\sum_i \hat{e}_i^2) + n\underbrace{E(\bar{Y} - \mu_Y)^2}_{\frac{\sigma^2}{n}} + \underbrace{E(\hat{\beta} - \beta)^2}_{\frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}} \sum_i (x_i - \bar{x})^2.$$

We the have

$$n\sigma^2 = E(\sum_i \hat{e}_i^2) + \sigma^2 + \sigma^2$$

and

$$E(\sum_i \hat{e}_i^2) = (n-2)\sigma^2$$

■

An alternative treatment is to divide equation (**??**) by σ^2:

$$\frac{\sum_i e_i^2}{\sigma^2} = \frac{\sum_i \hat{e}_i^2}{\sigma^2} + \frac{n(\bar{Y}-\mu_Y)^2}{\sigma^2} + \frac{(\hat{\beta}-\beta)^2 \sum_i (x_i-\bar{x})^2}{\sigma^2}.$$

And then rewrite it as

$$\sum_i \left(\frac{e_i - 0}{\sigma}\right)^2 = \frac{\sum_i \hat{e}_i^2}{\sigma^2} + \left(\frac{\bar{Y}-\mu_Y}{\frac{\sigma}{\sqrt{n}}}\right)^2 + \left(\frac{\hat{\beta}-\beta}{\frac{\sigma}{\sqrt{\sum_i (x_i-\bar{x})^2}}}\right)^2$$

That is

$$\frac{\sum_i \hat{e}_i^2}{\sigma^2} = \underbrace{\sum_i \left(\frac{e_i - 0}{\sigma}\right)^2}_{\chi^2(n)} - \underbrace{\left(\frac{\bar{Y}-\mu_Y}{\frac{\sigma}{\sqrt{n}}}\right)^2}_{\chi^2(1)} - \underbrace{\left(\frac{\hat{\beta}-\beta}{\frac{\sigma}{\sqrt{\sum_i (x_i-\bar{x})^2}}}\right)^2}_{\chi^2(1)},$$

or

$$\frac{\sum_i \hat{e}_i^2}{\sigma^2} \sim \chi^2(n-2).$$

Hence, $E(\sum_i \hat{e}_i^2) = (n-2)\sigma^2$, and

$$E(\hat{\sigma}^2) = E(\frac{\sum_i \hat{e}_i^2}{n-2}) = \sigma^2.$$

■

Proof 2 We can rewrite equation (**??**) as

$$\begin{aligned}\sum_i e_i^2 &= \sum_i \hat{e}_i^2 + [n(\hat{\alpha}-\alpha)^2 + 2(\hat{\alpha}-\alpha)(\hat{\beta}-\beta)\sum_i x_i + (\hat{\beta}-\beta)^2 \sum_i x_i^2] \\ &= \sum_i \hat{e}_i^2 + \Phi.\end{aligned}$$

And

$$\begin{aligned}E(\Phi) &= nE(\hat{\alpha}-\alpha)^2 + 2(\sum_i x_i)E(\hat{\alpha}-\alpha)(\hat{\beta}-\beta) + (\sum_i x_i^2)E(\hat{\beta}-\beta)^2 \\ &= nVar(\hat{\alpha}) + 2(\sum_i x_i)Cov(\hat{\alpha},\hat{\beta}) + (\sum_i x_i^2)Var(\hat{\beta}) \\ &= n\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i-\bar{x})^2}\right)\sigma^2 - \frac{2\bar{x}\sum_i x_i}{\sum_i (x_i-\bar{x})^2}\sigma^2 + \frac{\sum_i x_i^2}{\sum_i (x_i-\bar{x})^2}\sigma^2 \\ &= \left(1 + \frac{n\bar{x}^2}{\sum_i (x_i-\bar{x})^2} - \frac{2n\bar{x}^2}{\sum_i (x_i-\bar{x})^2} + \frac{\sum_i x_i^2}{\sum_i (x_i-\bar{x})^2}\right)\sigma^2 \\ &= 2\sigma^2.\end{aligned}$$

Thus,

$$E(\sum_i e_i^2) = E(\sum_i \hat{e}_i^2) + E(\Phi) \implies n\sigma^2 = E(\sum_i \hat{e}_i^2) + 2\sigma^2$$

$$E(\sum_i \hat{e}_i^2) = (n-2)\sigma^2 \implies E(\hat{\sigma}^2) = \frac{E(\sum_i \hat{e}_i^2)}{n-2} = \sigma^2$$

■

Gauss-Markov Theorem Let $b = \sum_i c_i Y_i$ be a linear unbiased estimator of β. Thus,

$$\begin{aligned}b &= \sum_i c_i Y_i \\ &= \sum_i c_i(\alpha + \beta x_i + e_i) \\ &= \alpha \sum_i c_i + \beta \sum_i c_i x_i + \sum_i c_i e_i\end{aligned}$$

Since b is unbiased, $E(b) = \alpha \sum_i c_i + \beta \sum_i c_i x_i + \sum_i c_i \underbrace{E(e_i)}_{0} = 0$. That is,

$$\sum_i c_i = 0,$$

and

$$\sum_i c_i x_i = 1.$$

When these conditions are satisfied,

$$b = \beta + \sum_i c_i e_i.$$

Therefore,

$$Var(b) = E(b-\beta)^2 = E(\sum_i c_i e_i)^2 = \sigma^2 \sum_i c_i^2$$

Note that

$$c_i = w_i + (c_i - w_i),$$

where $w_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2}$ with $\sum_i w_i = 0$ and $\sum_i w_i^2 = \frac{1}{\sum_i (x_i - \bar{x})^2}$.

Accordingly,

$$\sum_i c_i^2 = \sum_i w_i^2 + \sum_i (c_i - w_i)^2 + 2\sum_i w_i (c_i - w_i).$$

Obviously,

$$\sum_i w_i (c_i - w_i) = 0,$$

since

$$\sum_i w_i c_i = \sum_i \frac{(x_i - \bar{x}) c_i}{\sum_i (x_i - \bar{x})^2} = \frac{1}{\sum_i (x_i - \bar{x})^2}$$

and

$$\sum_i w_i^2 = \frac{1}{\sum_i (x_i - \bar{x})^2}.$$

Thus

$$\begin{aligned} &\sum_i c_i^2 = \sum_i w_i^2 + \sum_i (c_i - w_i)^2 \\ &\Leftrightarrow \underbrace{\sigma^2 \sum_i c_i^2}_{Var(b)} = \underbrace{\sigma^2 \sum_i w_i^2}_{Var(\hat{\beta})} + \sigma^2 \sum_i (c_i - w_i)^2 \end{aligned}$$

That is, $Var(b) = Var(\hat{\beta}) + \sigma^2 \sum_i (c_i - w_i)^2$ ■

Answers to Exercises

1 (p1) $P(A) + P(A^c) = 1$

Since $A \cap A^c = \emptyset$ and

$$A \cup A^c = S$$

From (d)

$$1 = P(S) = P(A \cup A^c) = P(A) + P(A^c)$$

(p2) $A \subseteq B$ implies that $P(A) \leq P(B)$

$$\begin{aligned} B &= B \cap S \\ &= B \cap (A \cup A^c) \\ &= (B \cap A) \cup (B \cap A^c) \\ &= A \cup (B - A) \end{aligned}$$

Clearly, A, $(B - A)$ disjoint. Hence by (d),

$$P(B) = P(A) + P(B - A) \geq P(A)$$

(p3) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (law of addition)

Since

$$A \cup B = (A \cup B) \cap S = (A \cup B) \cap (A \cup A^c) = A \cup (A^c \cap B)$$

Moreover,

$$B = (A \cap B) \cup (A^c \cap B)$$

with $(A \cap B)$, $(A^c \cap B)$ disjoint. Hence,

$$P(B) = P(A \cap B) + P(A^c \cap B) = P(A \cap B) + [P(A \cup B) - P(A)]$$

That is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

3

b	$f(b\|T=0)$
0.20	0
0.10	0.75
0.05	0.25

5

$$\begin{aligned}
Cov(X,Y) &= E(XY) - E(X)E(Y) \\
&= \sum_x \sum_y xyP(X=x, Y=y) - \sum_x xP(X=x) \sum_y yP(Y=y) \\
&= \sum_x \sum_y xyP(X=x)P(Y=y) - \sum_x xP(X=x) \sum_y yP(Y=y) \\
&= \sum_x xP(X=x) \sum_y yP(Y=y) - \sum_x xP(X=x) \sum_y yP(Y=y) \\
&= 0
\end{aligned}$$

6

$$\begin{aligned}
E(X|Y=y) &= \sum_x xP(X=x|Y=y) \\
&= \sum_x x\frac{P(X=x, Y=y)}{P(Y=y)} \\
&= \sum_x x\frac{P(X=x)P(Y=y)}{P(Y=y)} \\
&= \sum_x xP(X=x) \\
&= E(X)
\end{aligned}$$

7

$$\min Var(R_p) = p_1^2\sigma_1^2 + p_2^2\sigma_2^2 + 2p_1p_2\sigma_{12} \quad \text{s.t. } p_1 + p_2 = 1.$$

Hence, the minimization problem without constraint is

$$\min Var(R_p) = p_1^2\sigma_1^2 + (1-p_1)^2\sigma_2^2 + 2p_1(1-p_1)\sigma_{12}$$

$$2p_1\sigma_1^2 - 2(1-p_1)\sigma_2^2 + 2(1-2p_1)\sigma_{12} = 0$$

$$(2\sigma_1^2 + 2\sigma_2^2 - 4\sigma_{12})p_1 = 2\sigma_2^2 - 2\sigma_{12}$$

$$p_1 = \frac{\sigma - 2^2 - \sigma_{12}}{\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}}$$

8 On $\overline{ST}$.

9

$$\binom{10}{5}(0.5)^5(0.5)^5.$$

10

$$E(S_n^*) = E(\sum_{i=1}^{n} X_i) = \sum_i (E(X_i)) = \sum_i \mu = n\mu.$$

$$Var(S_n^*) = \frac{Var(\sum_{i=1}^{n} X_i)}{n^2} = \sum_i Var(X_i) = n\sigma^2$$

11 Recall that

$$(a+b)^n = \sum_{s=0}^{n} \binom{n}{s} b^s a^{n-s}$$

Let $b = p$, $a = 1 - p$,

$$\sum_{s=0}^{\infty} \binom{n}{s} p^s (1-p)^{n-s} = [(1-p) + p]^n = 1$$

12

$$\begin{aligned}
\sum_{x=0}^{\infty} \left(\frac{e^{-\lambda}\lambda^x}{x!} \right) &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{\lambda^x}{x!} \\
&= e^{-\lambda} \underbrace{\left[\frac{1}{0!} + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \cdots \right]}_{e^{\lambda}} \\
&= 1
\end{aligned}$$

13 1.

$$F(x) = P(X \leq x) = P(-\infty \leq X \leq x) = \int_{-\infty}^{x} f(u)du$$

2. According to Leibniz's Formula (or Fundamental Theorem of Calculus),

$$\frac{dF(x)}{dx} = \frac{d}{dx} \int_{-\infty}^{x} f(u)du = f(x)$$

3.

$$\int_{-\infty}^{\infty} f(u)du = \int_{-\infty}^{a} f(u)du + \int_{a}^{b} f(u)du + \int_{b}^{\infty} f(u)du$$

Thus,

$$\begin{aligned}\int_{a}^{b} f(u)du &= \Big[\int_{-\infty}^{\infty} f(u)du - \int_{b}^{\infty} f(u)du\Big] - \int_{-\infty}^{a} f(u)du \\ &= \int_{-\infty}^{b} f(u)du - \int_{-\infty}^{a} f(u)du = F(b) - F(a)\end{aligned}$$

14 1. If $X \sim U[0,1]$, and $Y = aX + b$ with $a < 0$, then

$$Y \sim U[a+b, b]$$

proof.

$$\begin{aligned}&P(X \le x) = x \\ &\iff P(aX + b \ge ax + b) = x \\ &\iff P(Y \ge y) = x \\ &\iff P(Y \le y) = 1 - P(Y \ge y) = 1 - x \\ &\iff P(Y \le y) = 1 - \frac{y-b}{a} = \frac{y-(b+a)}{-a} = \frac{y-(b+a)}{b-(b+1)} \\ &\iff Y \sim U[a+b, b]\end{aligned}$$

2. If $X \sim U[0,1]$, and $W = (l-h)X + h$, then

$$W \sim U[l, h]$$

3. If $W \sim U[l,h]$, and $Z = aW + b$ with $a < 0$, then

$$Z \sim U[ah+b, al+b]$$

proof.

$$Z = aW + b = \underbrace{a(l-h)}_{>0} X + \underbrace{ah+b} \sim U[ah+b, a(l-h)+(ah+b)] = U[ah+b, al+b]$$

17

$$\begin{aligned} M_X(t) = E(e^{tx}) &= \sum_{x=0}^{n} e^{tx} P(X = x) \\ &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= (pe^t + (1-p))^n \end{aligned}$$

Since $\sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x} = (a+b)^n$ (binomial formula).

18

$$\begin{aligned} M_X(t) = E(e^{tx}) &= \sum_{x=0}^{\infty} e^{tx} P(X = x) \\ &= \sum_{x=0}^{\infty} (e^t)^x \frac{e^{-\lambda} \lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} \\ &= e^{-\lambda} \left(\frac{1}{0!} + \frac{\lambda e^t}{1!} + \frac{(\lambda e^t)^2}{2!} + \cdots \right) \\ &= e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t - 1)} \end{aligned}$$

Since $e^y = \frac{1}{0!} + \frac{y}{1!} + \frac{y^2}{2!} + \cdots$ (Taylor expansion).

19

$$\begin{aligned} M_X(t) = E(e^{tx}) &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\ &= \lambda \int_0^{\infty} e^{-(\lambda - t)x} dx \\ &= \frac{-\lambda}{\lambda - t} e^{-(\lambda - t)x} \Big]_0^{\infty} \\ &= 0 - \left(\frac{-\lambda}{\lambda - t} \right) \\ &= \frac{\lambda}{\lambda - t}. \end{aligned}$$

Since $e^y = \frac{1}{0!} + \frac{y}{1!} + \frac{y^2}{2!} + \cdots$ (Taylor expansion).

20

$$\begin{aligned}
M_Z(t) = E(e^{tz}) &= \int_{-\infty}^{\infty} e^{tz} f(z)dz \\
&= \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{\frac{-z^2}{2}} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(z^2-2tz)}{2}} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\frac{-(z^2-2tz+t^2)}{2}} e^{\frac{t^2}{2}} dz \\
&= e^{\frac{t^2}{2}} \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}} e^{\frac{-(z-t)^2}{2}}}_{N(t,1)} dz \\
&= e^{\frac{t^2}{2}}
\end{aligned}$$

21

$$\begin{aligned}
M_Y(t) &= E(e^{tY}) \\
&= E\left[e^{t(aX+b)}\right] \\
&= E\left[e^{atX} e^{bt}\right] \\
&= e^{bt} E(e^{atX}) \\
&= e^{bt} M_X(at)
\end{aligned}$$

22 Using an example of continuous random variables,

$$\begin{aligned}
M_Y(t) &= E(e^{tY}) \\
&= E(e^{t(X_1+X_2+\cdots+X_n)}) \\
&= \int_{supp(X_1)} \cdots \int_{supp(X_n)} e^{t(x_1+x_2+\cdots+x_n)} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \\
&= \int_{supp(X_1)} \cdots \int_{supp(X_n)} e^{t(x_1+x_2+\cdots+x_n)} f(x_1) f(x_2) \ldots f(x_n) dx_1 dx_2 \cdots dx_n \\
&= \int_{supp(X_1)} e^{tx_1} f(x_1) dx_1 \int_{supp(X_2)} e^{tx_2} f(x_2) dx_2 \cdots \int_{supp(X_n)} e^{tx_n} f(x_n) dx_n \\
&= M_{X_1}(t) M_{X_2}(t) \cdots M_{X_n}(t)
\end{aligned}$$

23 We know that

$$M_{X_i}(t) = (1-p) + pe^t$$

Hence, by Fact 4,

$$M_Y(t) = \underbrace{[(1-p)+pe^t]\cdot[(1-p)+pe^t]\cdots[(1-p)+pe^t]}_{\text{n times}} = [(1-p)+pe^t]^n$$

That is, the MGF of a binomial random variable.

24 Since

$$M_X(t) = e^{t\mu_X + \frac{t^2\sigma_X^2}{2}}$$

and

$$M_Y(t) = e^{t\mu_Y + \frac{t^2\sigma_Y^2}{2}}$$

By Fact 3,

$$M_{aX}(t) = M_X(at) = e^{at\mu_X + \frac{a^2t^2\sigma_X^2}{2}}$$

and

$$M_{bY}(t) = M_Y(bt) = e^{bt\mu_Y + \frac{b^2t^2\sigma_Y^2}{2}}$$

By Fact 4,

$$\begin{aligned} M_W(t) = M_{aX}(t)M_{bY}(t) &= e^{at\mu_X + \frac{a^2t^2\sigma_X^2}{2}} e^{bt\mu_Y + \frac{b^2t^2\sigma_Y^2}{2}} \\ &= e^{t(a\mu_X + b\mu_Y) + \frac{t^2(a^2\sigma_X^2 + b^2\sigma_Y^2)}{2}}. \end{aligned}$$

That is, the MGF of a $\mathrm{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$ random variable.

25 Since $y = u(x)$ is one-to-one, it is either monotonic increasing or monotonic decreasing.

Case 1: monotonic increasing $u(x) \le a$ if and only if $x \le u^{-1}(a) = w(a)$.

That is

$u(x) \le y$ if and only if $x \le w(y)$.

Thus

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(u(X) \le y) \\ &= P(X \le w(y)) \\ &= F_X(w(y)) \end{aligned}$$

$$f_Y(y) = \frac{d}{dy}F_X(w(y)) = \frac{dF_X(w(y))}{dw(y)}\frac{dw(y)}{dy}$$
$$= f_X(w(y))\frac{d}{dy}w(y) = f_X(w(y))\left|\frac{d}{dy}w(y)\right|$$

because $\frac{d}{dy}w(y) > 0$.

Case 2: monotonic decreasing $u(x) \leq a$ if and only if $x \geq u^{-1}(a) = w(a)$. That is

$u(x) \leq y$ if and only if $x \geq w(y)$.

Thus

$$F_Y(y) = P(Y \leq y)$$
$$= P(u(X) \leq y)$$
$$= P(X \geq w(y))$$
$$= 1 - F_X(w(y))$$

$$f_Y(y) = -\frac{d}{dy}F_X(w(y)) = -\frac{dF_X(w(y))}{dw(y)}\frac{dw(y)}{dy}$$
$$= f_X(w(y))(-\frac{d}{dy}w(y)) = f_X(w(y))\left|\frac{d}{dy}w(y)\right|$$

because $\frac{d}{dy}w(y) < 0$.

26

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

$$\begin{aligned} Var(\bar{X}_n) &= Var\left[\frac{\sum X_i}{n}\right] \\ &= \frac{1}{n^2} Var(\sum X_i) \\ &= \frac{1}{n^2} \sum Var(X_i) \\ &= \frac{1}{n^2} \sum \sigma^2 \\ &= \frac{\sigma^2}{n} \end{aligned}$$

$$S^2 = \frac{1}{n-1} \underbrace{\sum (X_i - \bar{X}_n)^2}_{(A)}$$

$$\begin{aligned} E((A)) &= E(\sum (X_i - \bar{X}_n)^2) \\ &= E(\sum X_i^2 - n\bar{X}_n^2) \\ &= \sum E(X_i^2) - nE(\bar{X}_n^2) \\ &= nE(X_1^2) - n\left[Var(\bar{X}_n) + E(\bar{X}_n)^2\right] \\ &= n(\sigma^2 + \mu^2) - \sigma^2 - n\mu^2 \\ &= (n-1)\sigma^2 \end{aligned}$$

Hence,

$$E(S^2) = \frac{1}{n-1} E((A)) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

27 (T1) The MGF of X_i is

$$M_{X_i}(t) = E(e^{X_i t}) = e^{\mu_i t + \frac{\sigma_i^2 t^2}{2}}$$

Moreover,

$$\begin{aligned} M_{\alpha_i X_i}(t) &= E(e^{\alpha_i X_i t}) \\ &= M_{X_i}(\alpha_i t) \\ &= e^{\alpha_i \mu_i t + \frac{\sigma_i^2 \alpha_i^2 t^2}{2}} \end{aligned}$$

Hence,

$$\begin{aligned} M_{\sum_i \alpha_i X_i}(t) &= \Pi_i M_{\alpha_i X_i}(t) \\ &= \Pi_i e^{\alpha_i \mu_i t + \frac{\sigma_i^2 \alpha_i^2 t^2}{2}} \\ &= e^{(\sum_i \alpha_i \mu_i) t + \frac{1}{2}(\sum_i \alpha_i^2 \sigma_i^2) t^2} \end{aligned}$$

That is,

$$\sum_i \alpha_i X_i \sim N(\sum_i \alpha_i \mu_i, \sum_i \alpha_i^2 \sigma_i^2).$$

(T2) Since $\bar{X}_n = \frac{1}{n} \sum_i X_i$. Let

$$\alpha_i = \alpha = \frac{1}{n}, \quad \mu_i = \mu, \quad \sigma_i = \sigma, \quad \forall\, i$$

Hence,

$$\sum_i \alpha_i X_i = \bar{X}_n$$

$$\sum_i \alpha_i \mu_i = \sum_i \frac{1}{n} \mu = \frac{1}{n} n \mu = \mu$$

$$\sum_i \alpha_i^2 \sigma_i^2 = \sum_i \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{1}{n^2} n \sigma^2 = \frac{\sigma^2}{n}$$

According to (T1),

$$\underbrace{\sum_i \alpha_i X_i}_{\bar{X}_n} \sim \underbrace{N(\sum_i \alpha_i \mu_i, \sum_i \alpha_i^2 \sigma_i^2)}_{N(\mu, \frac{\sigma^2}{n})}.$$

29 By (T2),

$$\bar{X}_n \sim N(\mu_X, \frac{\sigma_X^2}{n})$$

and

$$\bar{Y}_m \sim N(\mu_Y, \frac{\sigma_Y^2}{m})$$

Hence, take $\alpha_X = 1$ and $\alpha_Y = -1$, by (T1)

$$\bar{X}_n - \bar{Y}_m \sim N(\mu_X - \mu_Y, \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m})$$

30

$$E(U) = \sqrt{k}E[Z]E[1/\sqrt{W}] = \sqrt{k} \cdot 0 \cdot E[1/\sqrt{W}] = 0.$$

$$\begin{aligned} Var(U) &= E(U^2) \\ &= E(kZ^2/W) \\ &= kE(Z^2)E(1/W) = \frac{k}{k-2}. \end{aligned}$$