Chapter 2

Sampling distribution

2.1 The Parameter and the Statistic

When we have collected the data, we have a whole set of numbers or descriptions written down on a paper or stored on a computer file. We try to summarize important information in the **sample** into one or two numbers, called **a statistics**. For each statistic, there is a corresponding summary number in the **population**, called **a parameter**.

2.2 Measures for quantitative variables

There are two basic statistics when we have a quantitative variables, the mean and the variance. The mean measures the center of data while the variance measures the spread out of data from its mean.

The population mean is denoted by μ

The population variance is denoted by σ^2

The standard deviation of the population is denoted by σ

The sample mean is denoted by $\bar{x}$

The sample variance is denoted by s^2

The standard deviation of the sample is denoted by s

2.3 Measures for qualitative variables

The measure for the qualitative variables(the characteristic which we want to study in the population) is called the proportion , thus we have :

1- the population proportion

$$p = \frac{\textit{the number of the characteristic in the population}}{N}$$

2- the sample proportion

$$r = \frac{\textit{the number of the characteristic in the sample}}{n}$$

Note that the proportion must be numbers between 0 and 1. A percentage may be obtained by multiplying the proportion by 100.

2.4 Sampling Methods

There are two basic types of sampling:

1- **Probability sampling:**

every population element has a chance of being chosen for the sample with known probability.

2- Non probability sampling:

Not every population element has a chance of being chosen for the sample or the probability of choosing an element is unknown .

For the statistical methods, it is useful to use the first type. Thus, we do not consider the second type.

The simple random sampling is the basic kind of probability sampling. We have two cases for sampling:

1- Sampling with replacement

The number of all possible samples of size n from a population of size N with replacement is

$$k = N^n$$

2- Sampling without replacement

The number of all possible samples of size n from a population of size N without replacement is

$$k = C_n^N = \binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

Ex(1)

If we have a population of size 5 and we want to choose samples of size 2.

a) With replacement

b) Without replacement

Solu.

The number of all possible samples is:

a) $k = 5^2 = 25$

b) $k = \binom{5}{2} = \frac{5!}{2!\,3!} = 10$

2.5 Sampling distribution of the sample mean $\bar{x}$

We can use the following steps to obtain the sampling distribution of the sample mean $\bar{x}$:

1- Find all possible samples of size n from a population of size N.

2- Calculate $\bar{x}$ for each sample, where $\bar{x} = \frac{\sum x}{n}$.

3- Construct the frequency table, for all different values of $\bar{x}$ and also the frequency of each value(the total of frequencies =k).

To study the sampling distribution of the sample mean and its relationship with the population parameters (μ, σ^2), we need to find its mean $E(\bar{x}) = \mu_{\bar{x}}$ and its variance $V(\bar{x}) = \sigma_{\bar{x}}^2$, as follows:

$$\mu_{\bar{x}} = \frac{\sum \bar{x} f}{k} \quad \text{and} \quad \sigma_{\bar{x}}^2 = \frac{\sum \bar{x}^2 f - k\mu_{\bar{x}}^2}{k}$$

4- Its relationship with population parameters (μ, σ^2) is:

a) $\mu_{\bar{x}} = \mu$

b) $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ (sampling with replacement) and

$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$(Sampling without replacement)

Where the fraction $\left(\frac{N-n}{N-1}\right)$ is called *correction factor.* We may ignore this correction factor in practice if $n \leq 0.05\ N$ *or* $\frac{n}{N} \leq 0.05$, that is the sample size is less than or equal to 5% of the population size N because that factor will approaches to 1.

5-Determine the form or shape of the sampling distribution. Since it depends on the distribution of the variable in the population(Normal or not normal), thus we have two cases for its form:

a) The population and the variable X in that population is normally distributed (I.e., $X \approx N(\mu, \sigma^2)$.Then, (with or without replacement)

$\bar{x} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$

b) The population and the variable X in that population has any distribution other than normal distribution with mean μ and variance σ^2, thus by applying the **Central Limit Theorem** :

(i) With replacement $\bar{x} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$

(ii) Without replacement $\bar{x} \approx N\left(\mu, \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)\right)$

2.5.1 Central Limit Theorem

If the sample size large enough, then some statistics (such as the sample mean and the sample proportion) have an approximate normal distribution with the same mean and variance as that obtained for the sampling distribution of that statistics. Whenever we can ignore the correction factor for the variance of $\bar{x}$,as n becomes larger $(n > 30)$, then $\bar{x}$ has an approximate normal distribution $\bar{x} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \rightarrow \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$ .This result is very important in the next chapter.

EX(2)

Consider a small population of five children with the following weights:

$$X_1 = 13.6, X_2 = 14.7,\ X_3 = 13.4, X_4 = 15, X_5 = 14.2$$

(1)Find the sampling distribution of the sample mean for samples of size 2 with and without replacement.

(2) can we ignore the correction factor?

(3) can we apply the central limit theorem?

(4) what is the form(type) of the sampling distribution of the mean

Solu.

1-(a)With replacement, the number of all possible samples of size 2 is:

$$k = 5^2 = 25$$

$\bar{x}$	F	$\bar{x}\,f$	$\bar{x}^2\,f$
13.4	1	13.4	179.56
13.5	2	27	364.5
13.6	1	13.6	184.96
13.8	2	27.6	380.88
13.9	2	27.8	386.42
14.05	2	28.1	394.805
14.15	2	28.3	400.445
14.2	3	42.6	604.92
14.3	2	28.6	408.98
14.45	2	28.9	417.605
14.6	2	29.2	426.32
14.7	1	14.7	216.09
14.85	2	29.7	441.045
15	1	15	225
Total	25	354.5	5031.53

Thus: $\mu_{\bar{x}} = \frac{\sum \bar{x}\,f}{k} = \frac{354.5}{25} = 14.18$

and $\sigma_{\bar{x}}^2 = \frac{\sum \bar{x}^2\,f - k\mu_{\bar{x}}^2}{k} = \frac{5031.53 - 25(14.18)^2}{25} = 0.1888$

(b)Without replacement, the number of all possible samples of size 2 is:

$$k = \binom{5}{2} = \frac{5!}{2!\ 3!} = 10$$

$\bar{x}$	F	$\bar{x}\,f$	$\bar{x}^2\,f$
13.5	1	13.5	182.25
13.8	1	13.8	190.44
13.9	1	13.9	193.21
14.05	1	14.05	197.4025
14.15	1	14.15	200.2225
14.2	1	14.2	201.64
14.3	1	14.3	204.49
14.45	1	14.45	208.8025
14.6	1	14.6	213.16
14.85	1	14.85	220.5225
Total	10	141.8	2012.14

Thus :

$$\mu_{\bar{x}} = \frac{\sum \bar{x}\,f}{k} = \frac{141.8}{10} = 14.18$$

and $\sigma_{\bar{x}}^2 = \frac{\sum \bar{x}^2 f - k\mu_{\bar{x}}^2}{k} = \frac{2012.14 - 10(14.18)^2}{10} = 0.1416$

Note that:

$$\mu = \frac{\sum x}{N} = \frac{70.9}{5} = 14.18$$

and $\sigma^2 = \frac{\sum X^2 - N\mu^2}{N} = \frac{1007.25 - 5(14.18)^2}{5} = 0.3776$

we can verify the relationship between population parameters and the sampling distributions of the mean as follows:

1) $\mu_{\bar{x}} = 14.18 = \mu$ (With or without replacement)

2) $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{0.3776}{2} = 0.1888$ (With replacement)

$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = 0.1888(0.75) = 0.1416$ (without replacement)

(2) The correction factor cannot be ignored because

$$\frac{n}{N} = \frac{2}{5} = 0.4 > 0.05$$

(3) Also, we cannot apply the central limit theorem, since the population is not normal and $n < 30$.

(4) we cannot know the form of this sampling distribution.

2.6 Sampling distribution of the sample proportion

Another statistics that will be seen is the sample proportion *r* that has one of two possible values of a qualitative variable. The sampling distribution for the sample proportion can be deduced by the previous steps for finding the sampling distribution for the sample mean, i.e., For each

possible sample, we will find the value of the sample proportion *r,* and then prepare a frequency table which gives the sampling distribution for the sample proportion.

1- the population proportion

$$\boldsymbol{p = \frac{the\ number of\ units\ in\ the\ population\ with\ \ the\ characteristic}{N} = \frac{A}{N}}$$

2- the sample proportion

$$\boldsymbol{r = \frac{the\ number of\ units\ in\ the\ sample\ with\ \ the\ characteristic}{n} = \frac{a}{n}}$$

2.6.1 The sampling distribution for the sample proportion without replacement

1- Find all possible samples of size *n* from a population of size *N.*

2- Calculate r for each sample, where $r = \frac{a}{n}$.

3- Construct the frequency table, for all different values of r and also the frequency of each value(the total of frequencies =k). This table is the sampling distribution for the sample proportion r.

To study the sampling distribution of the sample proportion and its relationship with the population proportion p, we need to find its mean μ_r and its variance σ_r^2, as follows:

$$\mu_r = \frac{\sum r\, f}{k} \quad \text{and} \quad \sigma_r^2 = \frac{\sum r^2\, f - k\mu_r{}^2}{k}$$

4- Its relationship with population proportion p is:

c) $\mu_r = p$ (with and without replacement)

d) $\sigma_r^2 = \frac{p(1-p)}{n}$ (sampling with replacement)

and $\sigma_r^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$ (Sampling without replacement)

Also, We may ignore ***the correction factor*** in practice if $n \leq 0.05\ N$.

Ex(3)

A population consists of five children , we asked each child if he like milk or not. we get the following results:

$$X_1 = No\ , X_2 = Yes,\ X_3 = Yes, X_4 = No\ , X_5 = Yes$$

1- Find the population proportion of the children who liked milk.

2- Find the sampling distribution for the sample proportion of size 3 without replacement for children who liked milk.

3- Deduce the mean and the variance of the sampling distribution of the sample proportion r.

4- Verifying the relation between the population proportion P,and the mean and the variance of the sample proportion r.

Solu.

(1) P = $\frac{A}{N} = \frac{3}{5} = 0.6$

(2) K = $c_n^N = c_3^5 = 10$

Samples	r = $\frac{a}{n}$	r	F
Nyy, NyN, Nyy,	$\frac{2}{3}, \frac{1}{3}, \frac{2}{3}$	$\frac{1}{3}$	3
NyN, Nyy, NNy,	$\frac{1}{3}, \frac{2}{3}, \frac{1}{3}$	$\frac{2}{3}$	6
yyN, yyy, yNy,	$\frac{2}{3}, \frac{3}{3}, \frac{2}{3}$	$\frac{3}{3}$	1
yNy	$\frac{2}{3}$		10

(3) $\mu_r = \frac{\sum rf}{K} = \frac{6}{10} = 0.6$ and $\sigma_r^2 = \frac{\sum r^2 f - k\mu_r^2}{k} = \frac{4 - 10(0.6)^2}{10} = 0.04$

(4) $\mu_r = P = 0.6$ and $\sigma_r^2 = \frac{P(1-P)}{n}\left(\frac{N-n}{N-1}\right) = \frac{(0.6)(.4)}{3}\left(\frac{5-3}{5-1}\right) = 0.04$

2.6.2 The sampling distribution for the sample proportion with replacement

In this case, we can use the Binomial distribution to find the values of the sample proportion r and their frequencies by using the following table:

A	r = $\frac{a}{n}$	f = $c_a^n (A)^a (N-A)^{n-a}$
0	$\frac{0}{n} = 0$	$c_0^n (A)^0 (N-A)^n$
1	$\frac{1}{n}$	$c_1^n (A)(N-A)^{n-1}$
2	$\frac{2}{n}$	$c_1^n (A)^2 (N-A)^{n-2}$
:		. . .
N	$\frac{n}{n} = 1$	$c_n^n (A)^n (N-A)^0$

Ex(4)

A population consists of 10 persons, two of them have influenza virus.

Find:

1- The population proportion of persons who have influenza virus.
2- The sampling distribution for the sample proportion of size 4 with replacement for persons who have influenza virus.
3- The mean and the variance of the sampling distribution for the sample proportion r.
4- Apply the relation between the population proportion P,and the mean and the variance of the sample proportion r.

Solu.

N = 10 A = 2 N-A = 8

(1) $P = \frac{A}{N} = \frac{2}{10} = 0.2$

(2) $K = N^n = 10^4 = 10000$

A	$r = \frac{a}{n}$	$f = C_a^4 (2)^a (8)^{4-a}$
0	0	$C_0^4 (2)^o (8)^4 = 4096$
1	$\frac{1}{4} = 0.25$	$C_1^4 (2)(8)^3 = 4096$
2	$\frac{2}{4} = 0.5$	$C_2^4 (2)^2 (8)^2 = 1536$
3	$\frac{3}{4} = 0.75$	$C_3^4 (2)^3 (8) = 256$
4	$\frac{4}{4} = 1$	$C_4^4 (2)^4 (8)^0 = 16$
	Total	10000

(3) $\mu_r = \frac{\sum rf}{K} = 0.2$ and $\sigma_r^2 = \frac{\sum r^2 f - k\mu_r^2}{k} = 0.04$

(4) $\mu_r = P = 0.2$ and $\sigma_r^2 = \frac{P(1-P)}{n} = \frac{0.2x0.8}{4} = 0.04$

2.6.3 Determining the form of the sampling distribution for the sample proportion

To Determine the form or shape of the sampling distribution for the proportion r we have two cases:

(1) If the sample size is large $(n > 30)$, thus we can apply the **Central Limit Theorem** as follows:

a) With replacement $r \approx N\left(p, \frac{p(1-p)}{n}\right)$

b) Without replacement $r \approx N\left(p, \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)\right)$

where ***the correction factor*** $\left(\frac{N-n}{N-1}\right)$ approaches to 1 and thus can be ignored if $n \leq 0.05\ N$.

(2) If the sample size is small ($n < 30$), thus we can't apply the **Central Limit Theorem** and also, we can't determine the form of the sampling distribution for the proportion r.

Ex(5):

A population of 500 women, 100 of them have a high blood pressure. If we made a sampling of size 25 with replacement, find:

1- The population proportion of women with high blood pressure.

2- $\mu_r \ and \ \sigma_r^2$

3- The form of the sampling distribution of the proportion r.

Solu.

(1) $P = \frac{100}{500} = 0.2$

(2) $\mu_r = \frac{100}{500} = 0.2$

$$\sigma_r^2 = \frac{P(1-P)}{n} = \frac{0.2(0.8)}{25} = 0.0064$$

(3) since, $n < 30$,we can't apply the central limit theorem and thus, we can't determine the distribution for the sample proportion r.

Ex(6):

In a city, we made a study about youth and marriage, we obtain that 80% of them married. If we take a sample of size 50 with replacement. Answer the following questions:

1- what is the value of the population proportion of not married?

2- μ_r *and* σ_r ?

3- can we apply the central limit theorem?

4- What is the type of the distribution of the proportion ?

Solu.

(1) P = 0.20

(2) $\mu_r = P = 0.2$ and $\sigma_r^2 = \frac{P(1-P)}{n} = \frac{0.2 \times 0.8}{50} = 0.0032$

thus $\sigma_r = 0.0566$

(3) Yes, we can apply the central limit theorem.

(4) since, $n > 30$,we can apply the central limit theorem and thus, the distribution for the sample proportion r is:

$r \approx N(\mu_r = 0.2,\ \sigma_r^2 = 0.0032)$

Chapter 3
Estimation of Confidence Intervals

3.1 A Point Estimate:

A point estimate of some population parameter θ , is a single value $\hat{\theta}$ of a statistic $\hat{\theta}$. For example, $\bar{x}$ of the statistic $\bar{X}$ computed from a sample of size **n** is a point estimate of the population parameter μ.Similarly $r = \frac{a}{n}$ is a point estimate of the true proportion P for a binomial experiment. $\bar{x}$ is an estimator of the population parameter μ, but the value of $\bar{x}$ is an estimate of μ.

The statistic that one uses to obtain a point estimate is called an estimator or a decision function.

Hence the decision function $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ which is a function of the random sample is an estimator of the population variance σ^2.The single numerical value that results from evaluating this formula is called an estimate of the parameter σ^2. An estimator is not expected to estimate the population parameter without error.

3.2 Interval Estimate:

An interval estimate of a population parameter θ is an interval of the form $\hat{\theta}_L < \theta < \hat{\theta}_U$ where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the value of the statistic $\hat{\theta}$ for a particular sample and also on the sampling distribution of $\hat{\theta}$.Since different samples will generally yield different values of $\hat{\theta}$ and therefore, different

values of $\hat{\theta}_L$ and $\hat{\theta}_U$. From the sampling distribution of $\hat{\theta}$ we shall be able to determine $\hat{\theta}_L$ and $\hat{\theta}_U$ such that the $P(\hat{\theta}_L < \theta < \hat{\theta}_U)$ is equal to any positive fractional value we care to specify. If for instance we find $\hat{\theta}_L$ and $\hat{\theta}_U$ such that:

$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1-\alpha \quad for \quad 0 < \alpha < 1$, then we have a probability of $(1-\alpha)$ of selecting a random sample that will produce an interval containing θ.

The interval $\hat{\theta}_L < \theta < \hat{\theta}_U$ computed from the selected sample, is then called a $(1-\alpha)100\%$ confidence interval, the fraction $(1-\alpha)$ is called confidence coefficient or the degree of confidence and the end points $\hat{\theta}_L$ and $\hat{\theta}_U$ are called the lower and upper confidence limits.

Thus when $\alpha = 0.05$ we have a 95% confidence interval and so on, that we are 95% confident that θ is between $\hat{\theta}_L$, $\hat{\theta}_U$

3.3 Estimating the Mean:

3.3.1 Confidence Interval of μ when σ is Known:

If $\bar{X}$ is the mean of a random sample of size **n** from a population with known variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (1)$$

where $Z_{1-\alpha/2}$ is the Z-value leaving an area of $\frac{\alpha}{2}$ to the right.

$$\hat{\theta}_L = \bar{X} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \quad , \quad \hat{\theta}_U = \bar{X} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad (2)$$

EX (1):

The mean of the quality point averages of a random sample of **36** college seniors is calculated to be **2.6**.

Find the 95% and 99% confidence intervals for the mean of the entire senior class. Assume that the population standard deviation is **0.3**.

Solu:

$n = 36$,$\bar{X} = 2.6$, $\sigma = 0.3$

95% confidence interval for the mean μ :

---

at $1-\alpha = 0.95 \to \alpha = 0.05 \to \frac{\alpha}{2} = 0.025 \to Z_{1-\frac{\alpha}{2}} = 1.96$,

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$2.6 \pm (1.96)(\frac{0.3}{\sqrt{36}}) \to 2.6 \pm 0.098$$

$$2.502 < \mu < 2.698 \to P(2.502 < \mu < 2.698) = 0.95$$

We are 95% confident that μ lies between 2.502, 2.698

99% confidence interval for the mean μ :

At $1-\alpha = 0.99 \to \alpha = 0.01 \to \frac{\alpha}{2} = 0.005 \to Z_{1-\frac{\alpha}{2}} = 2.57$,

$$\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$2.6 \pm (2.57)(\frac{0.3}{\sqrt{36}}) \to 2.6 \pm 0.1285$$

$$2.4715 < \mu < 2.7285 \to P(2.4715 < \mu < 2.7285) = 0.99$$

we are 99% confident that μ lies between 2.4713,2.7288

Theorem (1):

If $\bar{X}$ is used as an estimate of μ, we can then be $(1-\alpha)100\%$ confident that the error will not be exceed $Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

For example (1): d= (1.96) (0.3/6)=0.098 or

$$d = (2.575)\ (0.3/6) = 0.1288$$

Theorem (2):

If $\bar{X}$ is used as an estimate of μ, we can be $(1-\alpha)100\%$ confident that the error will not exceed a specified amount, d ,when the sample size is:

$$n = \left(\frac{Z_{1-\frac{\alpha}{2}}.\sigma}{d}\right)^2$$

The fraction of **n** is rounded up to next whole number.

EX (2):

How large a sample is required in Ex. (1) if we want to be 95% confident that our estimate of μ is off by less than **0.05** (the error is 0.05)**?**

Solu:

$$n = \left(\frac{Z_{1-\frac{\alpha}{2}}.\sigma}{d}\right)^2 = \left(\frac{(1.96)(0.3)}{0.05}\right)^2 = 138.2976 \cong 139$$

n is rounded up to whole number.

3.3.2 Confidence Interval of μ when σ is Unknown $(n < 30)$:

If $\bar{X}$ and S are the mean and standard deviation of a random sample from a normal population with unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is given by:

$$\bar{X} - t_{1-\frac{\alpha}{2},\, n-1}\frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{1-\frac{\alpha}{2},\, n-1}\frac{S}{\sqrt{n}} \qquad (4)$$

where $t_{1-\frac{\alpha}{2}}$ is the value with **n-1** degrees of freedom leaving an area of $\frac{\alpha}{2}$ to the right.

EX (3):

The contents of **7** similar containers of sulphuric acid are **9.8, 10.2, 10.4, 9.8, 10, 10.2, 9.6** litters. Find a 95% confidence interval for the mean of all such containers assuming an approximate normal distribution.

Solution:

For 95% confidence interval for the mean μ:

$n = 7$, $\bar{X} = 10$, $S = 0.283$,

$$at: 1-\alpha = 0.95 \rightarrow \alpha = 0.05 \rightarrow \frac{\alpha}{2} = 0.025 \rightarrow t_{1-\frac{\alpha}{2},\, n-1} = t_{0.025,6} = t_{0.975,\, 6} = 2.447$$

$$\bar{X} \pm t_{1-\frac{\alpha}{2},\, n-1}\left(\frac{S}{\sqrt{n}}\right) \Rightarrow 10 \pm (2.447)\left(\frac{0.283}{\sqrt{7}}\right) \Rightarrow 10 \pm 0.262$$

$$9.738 < \mu < 10.262 \rightarrow P(9.738 < \mu < 10.262) = 0.95$$

3.4 Estimating Confidence Interval for proportion P (with Large Sample):

If r is the proportion of successes in a random sample of size n then the $(1-\alpha)100\%$ confidence interval for the population proportion is given by:

$$r - d < P < r + d$$

Where $$d = z_{1-\frac{\alpha}{2}} \sqrt{\frac{r(1-r)}{n}}$$

where $z_{1-\frac{\alpha}{2}}$ is the z - value leaving an area of $\frac{\alpha}{2}$ to the right and d is the maximum value of the error.

EX (4):

A new rocket – launching system is being considered for deployment of small, short – rang rockets. The existing system has $P = 0.8$ as the probability of a successful launch. A sample of **40** experimental launches is made with the new system and **34** are successful. Construct a 95% confidence interval for P.

Solution:

95% **confidence interval for the proportion** P **:**

$$n = 40 \ , \ r = \frac{34}{40} = 0.85 \ \ , \ 1 - r = 0.15$$

$$1-\alpha=0.95\rightarrow\alpha=0.05\rightarrow1-\frac{\alpha}{2}=0.975\rightarrow z_{1-\frac{\alpha}{2}}=1.96$$

$$r\mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{r(1-r)}{n}}=0.85\mp(1.96)\sqrt{\frac{(0.85)(0.15)}{40}}=0.85\mp0.111$$

$$0.739<P<0.961\rightarrow p(0.739<P<0.961)=0.95$$

Theorem (3):

If r is used as an estimate of P, we can be $(1-\alpha)100\%$ confident that the error will not exceed $d=z_{1-\frac{\alpha}{2}}\sqrt{\frac{r(1-r)}{n}}$.

EX (5):

In Ex. 4, find the error of P.

Solution:

The error will not exceed the following value:

$$d=z_{1-\frac{\alpha}{2}}\sqrt{\frac{r(1-r)}{n}}=(1.96)\sqrt{\frac{(0.85)(0.15)}{40}}=0.111$$

Theorem (4):

If r is used as an estimate of P we can be $(1-\alpha)100\%$ confident that the error will be less than a specified amount d when the sample size is approximately:

$$n=\frac{z^2_{1-\frac{\alpha}{2}}\quad r(1-r)}{d^2}$$

Then the fraction of n is rounded up.

EX (6):

How large a sample is required in Ex. 4 if we want to be 95% confident that our estimate of P is within **0.02?**

Solution:

$$d = 0.02 \ , z_{1-\frac{\alpha}{2}} = 1.96 \ , r = 0.85 \ , 1 - r = 0.15$$

$$n = \frac{(1.96)^2(0.85)(0.15)}{(0.02)^2} = 1224.51 \cong 1225$$

3.5 Estimating the Difference between Two populations Means

3.5.1 Independent samples :

3.5.1.1 Confidence Interval for $\mu_1 - \mu_2$ when $\sigma_1^2 \, and \, \sigma_2^2$ Known:

If $\bar{X}_1 \, and \, \bar{X}_2$ are the means of independent random samples of size $\mathbf{n_1}$ and $\mathbf{n_2}$ from populations with known variances $\sigma_1^2 \, and \, \sigma_2^2$ respectively, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad (5)$$

where $Z_{1-\frac{\alpha}{2}}$ is the Z-value leaving an area of $\frac{\alpha}{2}$ to the right .

EX (7):

A standardized chemistry test was given to **50** girls and **75** boys. The girls made an average grade of **76**, while the boys made an average grade of **82**. Find a 96% confidence interval for the difference $\mu_1 - \mu_2$ where μ_1 is the mean score of all boys and μ_2 is the mean score of all girls who might take this test. Assume that the population standard deviations are **6** and **8** for girls and boys respectively.

Solution:

Girls	boys
$n_1 = 50$	$n_2 = 75$
$\bar{X}_1 = 76$	$\bar{X}_2 = 82$
$\sigma_1 = 6$	$\sigma_2 = 8$

96% **confidence interval for the mean** $\mu_1 - \mu_2$**:**

$$1-\alpha = 0.94 \rightarrow \alpha = 0.04 \rightarrow \frac{\alpha}{2} = 0.02 \rightarrow Z_{1-\frac{\alpha}{2}} = 2.05$$

$$(\bar{X}_1 - \bar{X}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(82-76) \pm (2.05)\sqrt{\frac{36}{50} + \frac{64}{75}} \Rightarrow 6 \pm 2.571$$

$$3.429 < \mu_1 - \mu_2 < 8.571 \rightarrow P(3.429 < \mu_1 - \mu_2 < 8.571) = 0.96$$

3.5.1.2 Confidence Interval for $\mu_1 - \mu_2$ when σ_1^2 *and* σ_2^2 Unknown

---

but equal variances:

$\bar{X}_1$ *and* $\bar{X}_2$ are the means of independent random samples of size $\mathbf{n_1}$ and $\mathbf{n_2}$ respectively from approximate normal populations with unknown but equal variances, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) \pm t_{1-\frac{\alpha}{2},\, v}\ S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (6),$$

Where $$S_P = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 + n_2 - 2}} \qquad (7)$$

is the pooled estimate of the population standard deviation and $t_{1-\frac{\alpha}{2},\, v}$ is the $\mathbf{t}$ – value with $v = n_1 + n_2 - 2$ degrees of freedom leaving an area of $\frac{\alpha}{2}$ to the right.

EX (8):

The independent sampling stations were chosen for this study, one located downstream from the acid mine discharge point and the other located upstream. For **12** monthly samples collected at the downstream station the species diversity index had a mean value $\bar{X}_1 = 3.11$ and a standard deviation $S_1 = 0.771$ while **10** monthly samples had a mean index value $\bar{X}_2 = 2.04$ and a standard deviation $S_2 = 0.448$. Find a 90% confidence interval for the difference between the population means for the two locations, assuming that the populations are approximately normally distributed with equal variances.

Solution:

Station 1	Station 2
$n_1 = 12$	$n_2 = 10$
$\bar{X}_1 = 3.11$	$\bar{X}_2 = 2.04$
$S_1 = 0.771$	$S_2 = 0.448$

90% **confidence interval for the mean** $\mu_1 - \mu_2$ **:**

$$S_P = \sqrt{\frac{11(0.771)^2 + 9(0.448)^2}{12+10-2}} = 0.646$$

$$at\ \ 1-\alpha = 0.90 \to \alpha = 0.1 \to \frac{\alpha}{2} = 0.05 \to t_{1-\frac{\alpha}{2}, n_1+n_2-2} \to t_{0.95\ ,\ 20} = 1.725$$

$$(3.11 - 2.04) \pm (1.725)(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} \Rightarrow 1.07 \pm 0.477$$

$$0.593 < \mu_1 - \mu_2 < 1.547 \to P(0.593 < \mu_1 - \mu_2 < 1.547) = 0.90$$

3.5.2 dependent (paired) samples :

Sometimes samples from two populations are not independent but paired by some important characteristics. For example, studying the height of the father and his first son, studying the weights of twins. So we can say that samples from two populations will be paired or dependent samples whenever the unit taken from one population is related to a unit taken from the second population. in the case of paired samples, it is more efficient to look at the differences in the values for the paired populations.

Also, we can say that we have paired samples when we have one population and get a random sample then we measure two characteristics on the same experimental unit, for example, suppose we take a sample of 20 people and measure the weights before exercising and after exercising. Then, we have one population and two variables:

1- The weights before exercising

2- The weights after exercising

But the both values are for the same person and thus, are paired.

Suppose we have a sample of size n from the first population $x_1, x_2, \dots x_n$ and a sample of size n from the second population $y_1, y_2, \dots y_n$, the study depends on the difference between the values of the two samples that we denoted by $D_i = x_i - y_i$ which has normal distribution with mean $\mu_D = \mu_{x-}\mu_y$ and unknown varianceσ_D^2.Thus, to begin analyze, we need the following information from samples:

$\overline{D} = the\ sample\ mean\ of\ the\ differences = \frac{\sum D_i}{n}$

$s_D^2 = the\ sample\ variance\ of\ the\ differences = \frac{\sum D_i^2 - n\overline{D}^2}{n-1}$

Therefore, the confidence interval for the difference is given by:

$$\overline{D} - d < \mu_D < \overline{D} + d$$

Where $d = t_{n-1,1-\frac{\alpha}{2}} \frac{s_D}{\sqrt{n}}$

Ex(9) page 111

location	Time 1	Time 2	D_i
1	25	105	-80
2	63	79	-16
3	79	107	-28
4	82	74	8
5	29	74	-45
6	38	68	-30
Total			-191

$\overline{D} = \frac{\sum D_i}{n} = \frac{-191}{6} = -31.83$

$s_D = \sqrt{\frac{\sum D_i^2 - n\overline{D}^2}{n-1}} = 29.492$ and $\alpha = 0.05$

$d = t_{5,0.975} \frac{s_D}{\sqrt{n}} = (2.5706).\frac{29.492}{\sqrt{6}} = 30.9499$

$$-62.7833 < \mu_D < -0.88336$$

3.6 Estimating Confidence Interval for Population Variance

We must assume that the populations are normal for all tests on population variances. We will also need two new distributions: one for variance and the other for the case of comparing two population variances.

3.6.1 Estimating Confidence Interval for Population Variance σ^2

If we have a random sample of size n from a normal population with unknown variance σ^2; thus the confidence interval is given by:

$$\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}$$

3.6.2 Estimating Confidence Interval for standard deviation σ

By taking the square root of the previous confidence interval for σ^2, we obtain the confidence interval of the population standard deviation as:

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}}$$

Note that:

To find the table value of chi-square at a degree of freedom doesn't exist in the table, we use the following rule

$$\chi^2_{df,1-\alpha} = \chi^2_{V1,1-\alpha} + \frac{df - v_1}{v_2 - v_1}\left(\chi^2_{V2,1-\alpha} - \chi^2_{V1,1-\alpha}\right)$$

Where

df: is the needed degree of freedom

V1:the closest degree of freedom lower than the needed df.

V2: the closest degree of freedom greater than the needed df.

Ex(10)

Find the table value $\chi^2_{34,0.05}$

Solu.

`Since df=34 is not in the table, we find V1 = 30 and V2 = 35. From the rule we have:

$$\chi^2_{34,0.05} = \chi^2_{30,0.05} + \frac{34-30}{35-30}\left(\chi^2_{35,0.05} - \chi^2_{30,0.05}\right)$$

$$= 18.493 + \frac{8}{10}(22.465 - 18.493)$$

$$= 18.493 + \frac{8}{10}(3.972) = 18.493 + 3.1776 = 21.6706$$

Ex(11)

A random sample of size 48 unit of a certain type of banana, the average of the length is15.7 cm and variance 1.2. assume that the length of banana have a normal distribution, then

1- Estimate the 90% confidence interval for the variance of length.
2- Estimate the 90% confidence interval for the standard deviation of length.

Solu.

n = 48 S^2 = 1.2 , 1-α = 0.90

$$\frac{(n-1)S^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1,\frac{\alpha}{2}}}$$

$$\frac{47\ (1.2)}{\chi^2_{47,0.95}} < \sigma^2 < \frac{47\ (1.2)}{\chi^2_{47,0.05}}$$

$$\chi^2_{47,0.05} = 30.612 + \frac{47-45}{50-45}(34.764 - 30.612) = 32.2728$$

$$\chi^2_{47,0.95} = 61.656 + \frac{47-45}{50-45}(67.505 - 61.656) = 63.9956$$

$$\frac{47(1.2)}{63.9956} < \sigma^2 < \frac{47(1.2)}{32.2728}$$

Thus the confidence interval for the variance is:

$$0.88131059 < \sigma 2 < 1.747601696$$

Also, the confidence interval for the standard deviation is:

$$(0.938781059 < \sigma < 1.321968871)$$

3.7 Estimating Confidence Interval for two Population Variances

If we have two methods to estimate something and we find that the two means are the same. Which method should we choose? We should choose the method which is less variable.

Therefore, for tests involving two variances, ***it is necessary to have independent samples from two normal populations*** . Then, the confidence interval for their ratio is :

$$\frac{s_1^2/s_2^2}{F_{1-\frac{\alpha}{2},\, n_1-1\,,n_2-1}} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2/s_2^2}{F_{\frac{\alpha}{2},\, n_1-1\,,n_2-1}}$$

Note that

$$F_{\frac{\alpha}{2},n_1-1,n_2-1} = \frac{1}{F_{1-\frac{\alpha}{2},n_2-1,n_1-1}}$$

Ex(12)

$F_{0.995,12,15} = 4.25$ but we cannot find the value of $F_{0.005,12,15}$ directly from the table, so

$$F_{0.005,12,15} = \frac{1}{F_{0.995,15,12}} = \frac{1}{4.72} = 0.21186$$

Ex(13)

In a study of milk in Riyadh markets in 1979, independent samples of raw milk and pasteurized milk were collected. The total bacteria count per ml(divided by 10^3) was measured:

	Sample size	**Mean**	**Standard deviation**
Raw Milk	31	31291.547	70521.846
Pasteurized Milk	23	784.838	2358.172

Assuming normal distributions with unequal variances, find a 99% confidence interval for the ratio of the variances of raw and pasteurized milk.

Solu.

Interval for their ratio is :

$$\frac{s_1^2/s_2^2}{F_{1-\frac{\alpha}{2},\,n_1-1\,,n_2-1}} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{s_1^2/s_2^2}{F_{\frac{\alpha}{2},\,n_1-1\,,n_2-1}}$$

Since,

$$F_{1-\frac{\alpha}{2},\, n_1-1\,,n_2-1} = F_{0.995,30\,,22} = 2.98$$

$$F_{\frac{\alpha}{2},\, n_1-1\,,n_2-1} = F_{0.005,30\,,22} = \frac{1}{F_{0.995,22\,,30}} = \frac{1}{2.282} = 0.3546$$

Therefore, the confidence interval will be as follows:

$$\frac{(70521.846)^2/(2358.172)^2}{2.98} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{(70521.846)^2/(2358.172)^2}{0.3546}$$

$$\frac{1589.6226}{2.98} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1589.6226}{0.3546}$$

$$533.4304 \leq \frac{\sigma_1^2}{\sigma_2^2} \leq 4482.7357$$

3.8 Estimating Confidence Interval for two Populations Proportions

If we have two independent random samples, then we can obtain the confidence interval of the difference for the two populations proportions as follows:

$$(r_1 - r_2) - d < P_1 - P_2 < (r_1 - r_2) + d$$

$$d = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{r_1(1-r_1)}{n_1} + \frac{r_2(1-r_2)}{n_2}}$$

Ex(14)

Two machines A,B. If we get a random sample of size 40 from machine A with defective proportion 0.18 and a random sample of size 60 from machine B with defective proportion 0.14. Estimate the difference of defective proportion with 95% confidence interval.

Solu.

n1 = 40 r1 = 0.18 1 - α = 0.95

n2 = 60 r2 = 0.14

thus, $Z_{0.975} = 1.96$

$$(r_1 - r_2) - d < P_1 - P_2 < (r_1 - r_2) + d$$

$$d = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{r_1(1 - r_1)}{n_1} + \frac{r_2(1 - r_2)}{n_2}}$$

$$d = (1.96)\sqrt{\frac{0.18(1 - 0.18)}{40} + \frac{0.14(1 - 0.14)}{60}} = 0.0878$$

$$(0.18 - 0.14) - 0.0878 < P_1 - P_2 < (0.18 - 0.14) + 0.0878$$

$$-0.0478 < P_1 - P_2 < 0.1278$$

Ex(15)

If take two samples A,B of children with the following answers about drinking milk during meals in the two populations:

samples	Yes	No	Total
A	77	143	220
B	88	92	180

Find with 99% confident the difference of the proportion for the children who always drink milk during meals in the two populations.

Solu.

$n_1 = 220$ $a_1 = 77$ $r_1 = \frac{77}{220} = 0.35$

$n_2 = 180$ $a_2 = 88$ $r_2 = \frac{88}{180} = 0.489$

$1 - \alpha = 0.99$ $\frac{\alpha}{2} = 0.005$ $Z_{1-\frac{\alpha}{2}} = Z_{0.995} = 2.575$

$$(r_1 - r_2) - d < P_1 - P_2 < (r_1 - r_2) + d$$

$$d = Z_{1-\frac{\alpha}{2}} \sqrt{\frac{r_1(1 - r_1)}{n_1} + \frac{r_2(1 - r_2)}{n_2}}$$

$$d = (2.575)\sqrt{\frac{0.35(1 - 0.35)}{220} + \frac{0.489(1 - 0.489)}{180}} = 0.12673$$

$(0.35 - 0.489) - 0.12673 < P_1 - P_2 < (0.35 - 0.489) + 0.12673$

$-0.26573 < P_1 - P_2 < 0.01227$