Chap 8 Notes: Confidence Interval Estimation

J. Harner A. Billings

Department of Statistics
West Virginia University

Stat 211 Fall 2007

Outline

Chap 8 Notes: Confidence Interval Estimation

Statistical Estimation and Confidence Interval:

- Measurements should be accurate,
  i.e.,close to the true or actual value.

- Measurements have two sources of error:
  1. random measurement error
  2. bias (systematic measurement error)

- So

$$\text{measured value} = \text{true value} + \text{bias} + \text{random error}$$

(From Sec. 5.1)

Law of Large Numbers:

If the sample size mean $\bar{X}$ should be close to the population mean μ.

- This implies that when n is large enough, the random errors in the data tend to "cancel out", so that

$$\begin{aligned}\bar{X} &\simeq \text{true value} + \text{bias} \\ &\simeq \mu + \text{bias}\end{aligned}$$

- Good methods and measuring instruments are necessary to try to eliminate bias.

Example:

Estimate mean height of population of 8^{th} grade male gym students, using sample (random) of 100.

$$\text{Measured height} = \text{True height} + \text{bias} + \text{random error}$$

here,

$$\begin{aligned} \text{bias} &: \text{wearing tennis shoes} \\ \text{random error} &: \text{stance (straight, slouched, etc)} \end{aligned}$$

and

$$\bar{X} \simeq \mu + \text{bias}$$

$$\begin{aligned} \bar{X} &: \text{sample mean height} \\ \mu &: \text{population mean height} \\ \text{bias} &: \text{tennis shoes height} \end{aligned}$$

- Eliminate bias by removing shoes.

Estimating Population Parameters:

- When the sample size n is large, the sample mean $\bar{X}$ and sample standard deviation S should be reasonably good estimators of the population mean μ and population standard deviation σ.

  - $\bar{X}, S$ are (sample) statistics.
  - μ, σ are (population) parameters.

$$\begin{array}{rcl} \text{Capital } \bar{X}, S & - & \text{random variables} \\ \text{Little } \bar{x}, s & - & \text{computed or observed numerical values} \end{array}$$

Random sample and Estimates

- A **repeated trials random experiment**, or just **random experiment**, consists of n **independent** random variables $X_1, X_2, X_3, \ldots, X_n$ each with the same probability distribution (identically distributed); n is the number of **replications**. The collection $X_1, X_2, X_3, \ldots, X_n$ of random variables is called a **random sample**.

Example:

In this example;

Population = all conceivable tosses;

Random sample is 100 tosses;

each toss is independent of all others.

each X has same probability distribution.

X	1	...	6
p(x)	$\frac{1}{6}$	...	$\frac{1}{6}$

n = 100 tosses.

Sampling with replacement from a conceptual population.

Random Sample from Population (Survey sampling)

- Random sampling from a population usually involves sampling without replacement - we usually don't want to have the same subject in the sample twice. Thus the observations are not independent. (Recall our "marble" example from Chap. 4)
- But when the population size N is large, we can "think of" the observation $X_1, X_2, X_3, \ldots, X_n$ as being independent.

Example:

1. Box with 500 Red, 500 yellow.
   Sample without replacement.

- Does first selection "affect" second selection?

$$P(R_2|R_1) = \frac{499}{999} = 0.4995$$

$$P(R_2|R_1) = \frac{500}{999} = 0.5005$$

- **Yes** - but not much!
  Even less "affect" if N is bigger.

Large Population Random Sample Rule:

- When obtaining a random sample from a real population of size N, which is large compared to the sample of size n, the random sample $X_1, X_2, X_3, \ldots, X_n$ can be modeled as being independent and identically distributed.
- Here, "large" means $N \geq 20n$or the size of the sample should be 5% or less the size of the population.

Central Limit Theorem for $\bar{X}$ (CLT):

- Sample Mean $\bar{X}$ is a random variable.

Example:

X = distance to home

Sample Data	$\bar{X}$

- Since $\bar{X}$ is a random variable, $\bar{X}$ has a distribution, which will have a mean and standard deviation. Let $\mu_{\bar{X}}$ be the mean of $\bar{X}$'s distribution, and $\sigma_{\bar{X}}$ be the the Standard deviation of $\bar{X}$'s distribution.
- Recall μ and σ are the mean and standard deviation of the distribution from which we obtain the data (the X's), called the **Parents distribution.**
- $\sigma_{\bar{X}}$, the standard deviation of the distribution of all possible values of $\bar{X}$, is also called **standard Error of $\bar{X}$**, denoted as $\underline{\text{SE}(\bar{X})}$.

- The distribution of all possible values of $\bar{X}$ is called the **sampling Distribution** of $\bar{X}$.

- **But some questions remain:**
  - What values do $\mu_{\bar{X}}$ and SE($\bar{X}$) assume?
  - what is the shape of the sampling distribution of $\bar{X}$?

- We can use the 5-step Method to simulate the probability distribution of possible values of $\bar{X}$ (sampling distribution).

- See the following diagram[1].

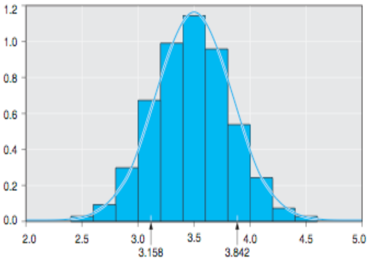


Figure 8.4 Experimental histogram of 10,000 simulated $\bar{x}$-values from a fair die population, $n = 25$.

- Histogram suggests that:
    1. sampling distribution of $\bar{X}$ is approximately normal.
    2. $E(\bar{X}) = \mu_{\bar{X}} \simeq \mu = 3.5$ (from Sec 5.3)
    3. $SE(\bar{X}) = \sigma_{\bar{X}} = 0.34 < \sigma = 1.71$ (from Sec 5.3)

[1] refer to page Page: 438 of the Text Book

Central Limit Theorem for $\bar{X}$ (CLT):

- Will the sampling distribution of $\bar{X}$ always look (approximately) normal?
- what is the value of $\mu_{\bar{X}}$? $SE(\bar{X})$?

- When the sample size n is large (n$\geq$ 20 for this book), we can say something about the sampling distribution of $\bar{X}$:

CLT: Suppose $X_1, X_2, X_3, \ldots, X_n$ is a random sample from a population with mean μ and standard deviation σ. Then the sampling distribution of $\bar{X}$ has

$$\mu_{\bar{X}} = E(\bar{X}) = \mu$$
$$\sigma_{\bar{X}} = SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

- IF **n**$\geq$ **20**, the sampling distribution of $\bar{X}$ will be approximately normal.
  Also
$$P(\bar{X} \leq x) = P\left(Z < \frac{x-\mu}{\sigma/\sqrt{n}}\right)$$

- What if n<20?

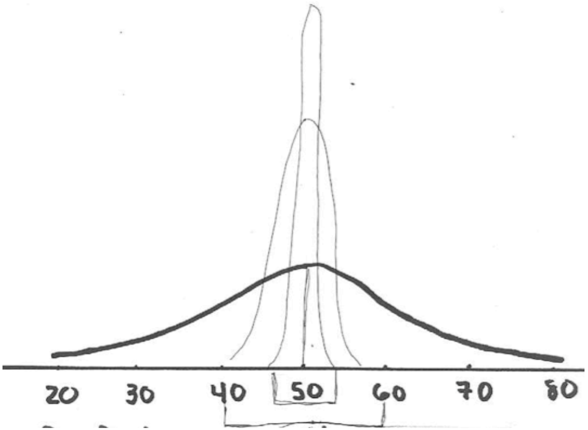
- IF **n< 20**, the sampling distribution of $\bar{X}$ is not necessarily normal, or even approximately normal.

Example:

Suppose random sample of size n = 25 from population with μ = 50 and σ = 10. Describe the Sampling Distribution of $\bar{X}$.

Population Dist			Sampling Dist. of $\bar{x}$		
μ	σ	n	$\mu_{\bar{x}}$	$SE(\bar{x})$	Shape
50	10	16			
50	10	25			
50	10	49			
50	10	100			
75	12	9			
42	24	64			

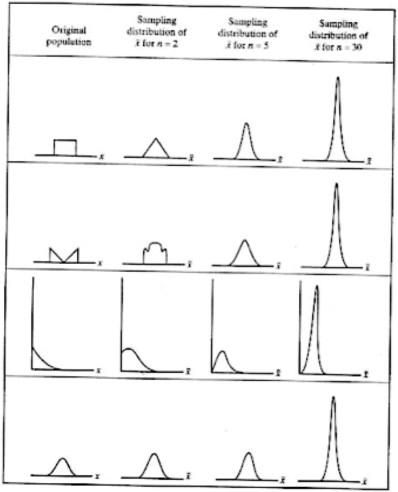
- SE($\bar{X}$) = $\frac{\sigma}{\sqrt{n}}$



Population Distribution – N , μ=50, σ=10

$\bar{X}$'s Distribution (n=25) – N , $\mu_{\bar{X}}$=50, $\sigma_{\bar{x}}$=2

$\bar{X}$'s Distribution (n=100) – N , $\mu_{\bar{X}}$=50, $\sigma_{\bar{x}}$=1



answer depends on the shape of the distribution of the sampled population, as shown by Figure 6.10. Generally speaking, the greater the skewness of the sampled population distribution, the larger the sample size must be before the normal distribution is an adequate approximation for the sampling distribution of $\bar{x}$. For most sampled populations, sample sizes of $n \geq 30$ will suffice for the normal approximation to be reasonable. We will use the normal approximation for the sampling distribution of $\bar{x}$ when the sample size is at least 30.

Example:

Suppose weight of an apple follows a probability distribution with mean 100 grams and standard deviation 16 grams.

a. Assume that the probability distribution is Normal. find the probability that a randomly selected apple weighs more than 105 grams.

$$
\begin{aligned}
Z &= \frac{X-\mu}{\sigma} \\
&= \frac{105-100}{16} \\
&= 0.3125
\end{aligned}
$$

(Contd.)

$$\begin{aligned} P(X > 105) &= P(Z > 0.31) \\ &= 1 - P(Z < 0.31) \\ &= 1 - 0.6217 \\ &= 0.3783 \end{aligned}$$

b. No longer assume that apple weight follows a normal distribution.

   Find the probability that the mean weight of 64 randomly selected apples exceeds 105 grams

   Find P($\bar{X}$>105)

   Since n is large (n = 64$\geq$20), we can use the **CLT**.

(Contd.)

$$
\begin{aligned}
Z &= \frac{X-\mu}{\frac{\sigma}{\sqrt{n}}} \\
&= \frac{105-100}{16/\sqrt{64}} \\
&= 2.50
\end{aligned}
$$

$$
\begin{aligned}
P(\bar{X} > 105) &= P(Z > 2.50) \\
&= 1 - P(Z < 2.50) \\
&= 1 - 0.9938 \\
&= 0.0062
\end{aligned}
$$

Confidence Interval Estimation - Introduction:

- We can use $\bar{X}$ to estimate μ. But it is unlikely that $\bar{X}$ ia a perfect estimate of μ. ($\bar{X}$ is a point estimate of μ)
- Better to say that μ is between two numbers, a and b, i.e,

$$a < \mu < b$$

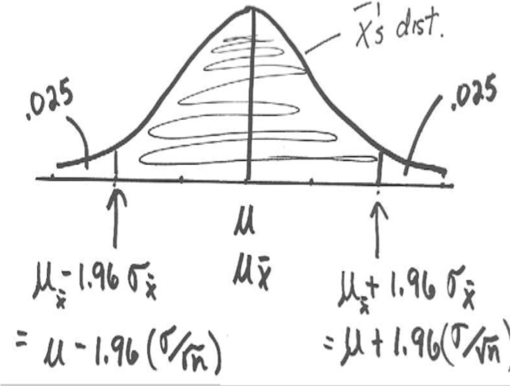
  - This is called an interval estimate of μ
- We can state the interval equivalently as

$$\bar{X} \pm \text{error term}$$

  - What is the **likelihood** that μ really is between a and b? (**Confidence level**)
  - How large is the **error term**?

95% Confidence Interval for μ (Large Sample):

- **Recall:** If n is large (n$\geq$ 20), $\bar{X}$'s distribution is approximately normal, with mean $\mu_{\bar{X}} = \mu$ and standard deviation $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \sigma_{\bar{X}}$



- So 95% of the values of $\bar{X}$ lie between $\mu - 1.96.(\frac{\sigma}{\sqrt{n}})$ and $\mu + 1.96.(\frac{\sigma}{\sqrt{n}})$

$$P\left(\mu - 1.96.(\frac{\sigma}{\sqrt{n}}) < \bar{X} < \mu + 1.96.(\frac{\sigma}{\sqrt{n}})\right) = 0.95$$

which can rearranged

$$P\left(\bar{X} - 1.96.(\frac{\sigma}{\sqrt{n}}) < \mu < \bar{X} + 1.96.(\frac{\sigma}{\sqrt{n}})\right) = 0.95$$

- So the 95% **C.I.** goes from

$$\bar{X} - 1.96.(\frac{\sigma}{\sqrt{n}}) \text{ to } \bar{X} + 1.96.(\frac{\sigma}{\sqrt{n}})$$

- Since we usually don't know σ, we use s.

$$\bar{X} - 1.96.(\frac{s}{\sqrt{n}}) \text{ to } \bar{X} + 1.96.(\frac{s}{\sqrt{n}})$$

Example:

Construct a 95% C.I. for mean weight of full grown dogs,

$$\bar{X} = 44 \text{ lbs}, \quad s = 8 \text{ lbs}, \quad n = 36$$

$$\begin{aligned} \bar{X} \pm 1.96.(\frac{s}{\sqrt{n}}) &= 44 \pm 1.96.(\frac{8}{\sqrt{36}}) \\ &= 44 \pm 2.61 \end{aligned}$$

- So our 95% C.I. is from

$$41.39 \text{ lbs to } 46.61 \text{ lbs}, .$$

- We are 95% confident that the mean weight of all full grown dogs is between

$$41.39 \text{ lbs and } 46.61 \text{ lbs}, .$$

- 95% is called the **Confidence Level.**
- **Interpretation** - See Page: 455 of the Text Book.
- **Note:** Other Confidence Levels are sometimes used by researchers:

$$90\% \qquad \text{Use } Z = 1.645$$

$$99\% \qquad \text{Use } Z = 2.578$$

$$\bar{X} \pm \text{critical value} \times \widehat{SE}(\bar{X})$$

$$\bar{X} \pm Z \times \widehat{SE}(\bar{X})$$

- The error term depends on

  1. The **sample size** n
     (the larger n is, the smaller the error term.)
  2. The **standard deviation** S
     (the smaller S is, the smaller the error term.)
  3. The **confidence level**
     (the higher the confidence level, the larger the critical value is; so the larger the error term is.)

Sample Size for Estimating μ

- The size, n , of the sample depends on

1. **Variability of the data.** (standard deviation S)
2. **Precision of the estimate**, i.e., the size of the error term $Z(\frac{S}{\sqrt{n}})$

Example:

Suppose we wish to estimate μ with a precision of 2 units (i.e., the error term is 2) with 95% confidence. Also suppose we know S = 12.6.
Find the sample size required.

- We must find n such that

$$Z(\frac{S}{\sqrt{n}}) = 2$$

(Contd.)

Assuming that n is "large" for a 95% C.I.

$$Z = 1.96$$

So,

$$\begin{aligned} 1.96 \times \left(\frac{S}{\sqrt{n}}\right) &= 2 \\ 1.96 \times \left(\frac{12.6}{\sqrt{n}}\right) &= 2 \end{aligned}$$

Solving for n yields

$$n = 152.47$$

which we round up to

$$n = 153 \text{ observations}$$

We need at least 153 observations.

100(1-α)% C.I. for Binomial Parameter p (Large Sample):

- Suppose we have data from a binomial experiment, but we don't know the value of p. (**Recall**: p = P(Success))
- We can use a point estimator

$$\hat{p} = \frac{X}{n}$$

  to estimate p.
- We can also use an interval estimate, provided
  1. n.$\hat{p}$ > 5 and
  2. n.(1-$\hat{p}$) > 5
- The 100(1-α)% C.I. will have the form

$$\hat{p} \pm Z \times SE(\hat{p})$$

- We can estimate $SE(\hat{p})$ by

$$SE(\hat{p}) \simeq \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$$

- Thus for a **binomial experiment** with X successes in n trial, a 100(1-α)% C.I. for p is

$$\frac{X}{n} \pm Z \times \sqrt{\frac{\frac{X}{n}.(1-\frac{X}{n})}{n}}$$

or

$$\hat{p} \pm Z \times \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$$

Example:

In a sample of 200 college students, 120 were female and 80 were male. Construct a 95% C.I. for the proportion p of all college students which are female.

- Let Success = female.

  Then X = number of Successes = number of females
  and n = 200.

$$\frac{X}{n} \pm 1.96 \times \sqrt{\frac{\frac{X}{n}.(1-\frac{X}{n})}{n}}$$

$$\frac{120}{200} \pm 1.96 \times \sqrt{\frac{\frac{120}{200}.(1-\frac{X}{120})}{200}}$$

$$0.6 \pm 1.96 \times \sqrt{(0.6).(1-0.6)}200$$

$$0.6 \pm 1.96 \times (0.034641)$$

$$0.6 \pm 0.0679$$

(Contd.)

So a 95% C.I. on p goes from 0.5321 to 0.6679.

- We are 95% confident that the proportion of female college students is between

$$0.5321 \text{ to } 0.6679$$

Example:

Example 8.20 of the Text Book[2].

[2] refer to Page: 458 of the Text Book

Sample Size for Estimating p:

- Given a specified size for the error term, find n.
- **Recall**: For a 95% C.I. on p, the error term is given by

$$1.96 \times \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}}$$

Example: (Contd.)
We wish to estimate p within 0.03 (error term) with 95% confidence.
Our previous example had,

$$n = 200 \text{ and } X = 120$$

So,

$$\hat{p} = \frac{X}{n} = \frac{120}{200} = 0.6$$

(Contd.)

We must solve for n:

$$\begin{aligned} 1.96 \times \sqrt{\frac{\hat{p}.(1-\hat{p})}{n}} &= 0.03 \\ 1.96 \times \sqrt{(0.6).(1-0.6)n} &= 0.03 \end{aligned}$$

Solving for n yields

$$n = 1024.43$$

We need n = 1025 observations.

- **Note:** IF $\hat{p}$ is not available from a "pilot" study, set $\hat{p}$ to 0.5. Then solve for n as in the example above.

Confidence Interval for μ (Small Sample; Data From Normal Population)

I. **Case 1** - $\underline{\sigma \text{ known}}$
Distribution of $\bar{X}$ when population is Normal.

- Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a Normal population with mean μ and standard deviation σ. Then the sampling distribution of $\bar{X}$ is Normal with mean $\mu_{\bar{X}} = \mu$ and $SE(\bar{X}) = \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

- IF $X_1, X_2, \ldots, X_n$ is a random sample from Normal population with parameters μ and σ, then a (1-α)100% C.I. is given by

$$\bar{X} \pm Z.SE(\bar{X})$$

i.e.,

$$\bar{X} \pm Z.\left(\frac{\sigma}{\sqrt{n}}\right)$$

Example:

A random sample of 16 observations from a Normal population with known $\sigma = 18$ yields $\bar{X} = 43.2$

Construct a 99% C.I on μ

$$
\begin{aligned}
\bar{X} \quad &\pm \quad Z.\left(\frac{\sigma}{\sqrt{n}}\right) \\
43.2 \quad &\pm \quad 2.578.\left(\frac{18}{\sqrt{16}}\right) \\
43.2 \quad &\pm \quad 2.578.(4.5) \\
43.2 \quad &\pm \quad 11.6
\end{aligned}
$$

Our 99% C.I. goes from 31.6 to 54.8.

II. **Case 2** - σ unknown

- When σ is unknown, we use a slightly different formula to construct 100(1-α)% C.I.

$$\bar{X} \pm t_{n-1} \times \left(\frac{S}{\sqrt{n}}\right)$$

where the value of t_{n-1} comes from a "t - distribution" with (n-1) degrees of freedom (**Table F**)[3].

- The value of t depends on:
  1. **The Confidence Level**
     for a 95% C.I. use
     $$\frac{\alpha}{2} = \frac{(1 - 0.95)}{2} = 0.025$$
  2. **The Sample size,n,**
     which determines the number of degrees of freedom,
     $$df = n - 1$$

[3] refer to Page: 701 of the Text Book

Example:

Construct a 95% C.I. for the mean weight of apples.

$$\bar{X} = 183 \text{ grams}, S = 14.1 \text{ grams}, n = 16$$

- First, find the value of t.

Here, $\frac{\alpha}{2} = 0.025$ (95% C.I.)
and df = n-1 = 15.

So, t = 2.13

Next, compute

$$\begin{array}{rcl} \bar{X} & \pm & t_{n-1} \times \left(\frac{S}{\sqrt{n}}\right) \\ 183 & \pm & 2.13 \times \left(\frac{14.1}{\sqrt{16}}\right) \\ 183 & \pm & 2.13 \times (3.525) \\ 183 & \pm & 7.5 \end{array}$$

(Contd.)

So, we estimate, with 95% confidence, that the mean weight μ of apple is between 175.5 grams and 190.5 grams.

Example: (Contd.)

Suppose n or α were different. Find t :

1. n =10 α =0.05
2. n = 24 95% C.I.
3. n = 18 90% C.I.

Example:

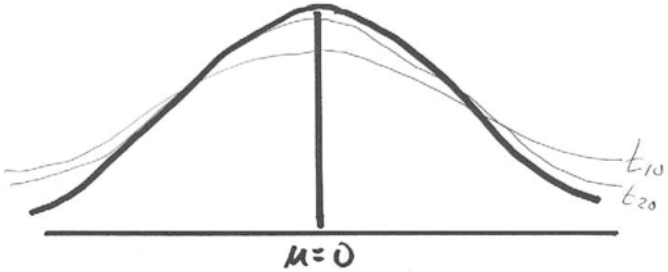
See Example 8.26 of the Text Book[4].

[4] refer to Page: 473 of the Text Book

t - distribution

- The t - distribution is actually a family of distributions, each determined by one parameter, the number of **degrees of freedom (df).**
- All t-distributions posses the following properties:
  1. $\mu = 0$
  2. mound-shaped
  3. symmetric about $\mu = 0$
  4. look like the standard normal distribution (Z), except with a "flatter" middle and "fatter" tails.
  5. $t = \frac{\bar{X}-\mu}{\left(\frac{S}{\sqrt{N}}\right)}$
  6. as n increases,

$$t \longrightarrow Z$$



$N, \mu = 0, \sigma = 1$

t with df = 10

t with df = 20

Degrees Of Freedom (d.f.)

- The number of degrees of freedom equals the number of independent data values used in the calculation, minus the number of restriction placed on the data.

Example:

Suppose we know n = 5 and $\bar{X}$ = 10.

Then df = n - 1 = 4

Since we can "choose" the first 4 data values freely; however the fifth data value must be chosen so that $\bar{X} = 10$.

(Contd.)

For example, the first 4 data values might be

4 6 12 7

Then the fifth value must be

$$21 \qquad \text{(to make } \bar{X} = 10\text{)}$$

or,

The first 4 data values might be

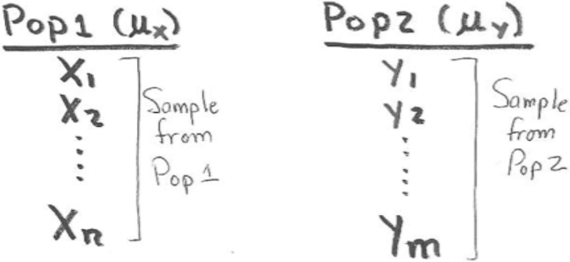
8 11 15 9

and the fifth value must be

7

100(1-α)% C.I. for Difference in Means of Two populations (Unpaired data),$\mu_X - \mu_Y$

- Suppose we obtain random samples from 2 distinct populations. We wish to estimate the difference in population means $\mu_X - \mu_Y$ using a 95% C.I.



  - **Note:** The sizes of the two samples may be different.

- **Case I:** Small Samples from Normal populations

  100(1-α)% C.I. for $\mu_X - \mu_Y$

$$(\bar{X} - \bar{Y}) \pm t.S_P.\sqrt{\frac{1}{n} + \frac{1}{m}}$$

- where

$$\begin{aligned} \bar{X} &= \text{mean of sample 1} \\ n &= \text{size of sample 1} \\ \bar{Y} &= \text{mean of sample 2} \\ m &= \text{size of sample 2} \\ \text{t has df} &= n + m - 2 \end{aligned}$$

$$S_P = \sqrt{\frac{(n-1).S_X^2 + (m-1).S_Y^2}{n+m-2}}$$

- S_P is called the **"pooled standard deviation"**.

$$\begin{aligned} S_X^2 &= \text{variance of sample 1} \\ S_Y^2 &= \text{variance of sample 2} \end{aligned}$$

Example:

Mental task Problem

$n = 15 \text{ males} \quad \bar{X} = 18.3 \text{ min} \quad S_X = 3.4 \text{ min}$

$m = 16 \text{ females} \quad \bar{Y} = 12.6 \text{ min} \quad S_Y = 2.9 \text{ min}$

Find a 95% C.I. for mean difference in time between males and females $\mu_X - \mu_Y$.

- First Compute

$$
\begin{aligned}
S_P &= \sqrt{\frac{(n-1).S_X^2 + (m-1).S_Y^2}{n+m-2}} \\
&= \sqrt{\frac{(15-1).3.4^2 + (16-1).2.9^2}{15+16-2}} \\
&= \sqrt{\frac{287.99}{29}} \\
&= \sqrt{9.931} \\
&= 3.15
\end{aligned}
$$

(Contd.)

For 95%,

$$\frac{\alpha}{2} = \frac{0.05}{2} = 0.025$$
$$df = n + m - 2 = 29$$

So, t = 2.045

$$(\bar{X} - \bar{Y}) \pm t.S_P.\sqrt{\frac{1}{n} + \frac{1}{m}}$$
$$(18.3 - 12.6) \pm 2.045.(3.15).\sqrt{\frac{1}{15} + \frac{1}{16}}$$
$$5.7 \pm 6.442.\sqrt{0.129167}$$
$$5.7 \pm 2.31$$

Our 95% C.I. goes from 3.4 min to 8.0 min.

- **This means:**

  With 95% Confidence, it takes males between 3.4 min and 8.0 min more than females, "on the average" to complete the task.

Assumptions:

1. Measurements in each population are approximately normally distributed.
2. The population standard deviations are equal, i.e.,

$$\sigma_X = \sigma_Y$$

Example:

See Example 8.28 of the Text Book.[5]

[5] refer to the Page: 479 of the Text Book

- **Case II:** Large Samples

  If both sample sizes, n and m are large, we can construct a $100(1-\alpha)\%$ C.I. on $\mu_X - \mu_Y$, as follows

$$(\bar{X} - \bar{Y}) \pm Z.\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

Example:

College statistics professor wants to estimate the difference in performance on EXAM 1 of students who had two or more high school math courses and those students who took fewer than two math courses in high school. Summary of data as follows

Two or more math courses:

$$n = 35 \quad \bar{X} = 84.2 \quad S_X = 10.2$$

Fewer than 2 math courses:

$$m = 45 \quad \bar{Y} = 73.1 \quad S_Y = 14.3$$

(Contd.)

Find 90% C.I. for the difference in mean scores $\mu_X - \mu_Y$.

- A 100(1-α)% C.I. on $\mu_X - \mu_Y$ is given by,

$$(\bar{X} - \bar{Y}) \pm Z.\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

For a 90% C.I., Z = 1.645

Then,

$$\begin{aligned}(84.2 - 73.1) &\pm 1.645.\sqrt{\frac{10.2^2}{35} + \frac{14.3^2}{45}} \\ 11.1 &\pm 1.645.\sqrt{7.51679} \\ 11.1 &\pm 4.51\end{aligned}$$

Our 90% C.I. goes from 6.6 to 15.6

- **This means:**

  With 90% Confidence, students who took two or more high school math courses average 6.6 to 15.6 points higher on statistics EXAM 1 than students who take fewer than two high school math courses.

Confidence Interval for The Difference between Two Population Proportions, $p_1 - p_2$

- We may be interested in estimating (C.I.) the difference in proportion for two independent populations, $p_1 - p_2$.
- Suppose n is the size of a sample from population 1, and m is the size of the sample from population 2.
- If the following conditions hold true:
  1. $n.\hat{p}_1 > 5$
  2. $n.(1 - \hat{p}_1) > 5$
  3. $m.\hat{p}_2 > 5$
  4. $m.(1 - \hat{p}_2) > 5$

(Contd.)

- then we can construct 100(1-α)% C.I. on $p_1 - p_2$

$$(\hat{p}_1 - \hat{p}_2) \pm Z.\sqrt{\frac{\hat{p}_1.(1-\hat{p}_1)}{n} + \frac{\hat{p}_2.(1-\hat{p}_2)}{m}}$$

where $\hat{p}_1 = \frac{X}{n}$ and $\hat{p}_2 = \frac{Y}{m}$

$$X = \text{number of Successes in Sample 1}$$
$$Y = \text{number of Successes in Sample 2}$$

Example: **Flu vaccine**

Estimate the difference in proportion of vaccinated adults who get the flu, and the proportion of unvaccinated adults who get the flu. Independent random samples from each population

Vaccinated:

$$n = 50 \quad X = \text{get flu} = 18 \quad \hat{p}_1 = \frac{X}{n} = 0.36$$

(Contd.)

Unvaccinated:

$$m = 100 \quad Y = \text{get flu} = 48 \quad \hat{p}_2 = \frac{Y}{m} = 0.48$$

- Note that:

1. $n.\hat{p}_1 = 18 > 5$
2. $n.(1 - \hat{p}_1) = 32 > 5$
3. $m.\hat{p}_2 = 48 > 5$
4. $m.(1 - \hat{p}_2) = 52 > 5$

(Contd.)

Lets' construct 90% C.I. on $p_1 - p_2$, the difference in population proportions:

$$
\begin{aligned}
(\hat{p}_1 - \hat{p}_2) &\pm Z.\sqrt{\frac{\hat{p}_1.(1-\hat{p}_1)}{n} + \frac{\hat{p}_2.(1-\hat{p}_2)}{m}} \\
(0.36 - 0.48) &\pm 1.645.\sqrt{\frac{0.36.(1-0.36)}{50} + \frac{0.48.(1-0.48)}{100}} \\
-0.12 &\pm 1.645.\sqrt{0.004608 + 0.002496} \\
-0.12 &\pm 1.645.\sqrt{0.007104} \\
-0.12 &\pm 1.645.(0.084285) \\
-0.12 &\pm 0.13865
\end{aligned}
$$

$$(-0.25865, 0.01865)$$

- With 90% Confidence, we estimate that the proportion of vaccinated adults who get flu is between 25.9% less and 1.9% more than the proportion of unvaccinated adults who get flu.

- **Note:** since 0 is inside the confidence interval, a difference in proportions equal 0 is plausible. Effectively, the result of this experiment indicates that vaccination was not effective in reducing the chance of getting flu.
- **Note:** Flu vaccine is effective for the elderly and ill. Flu vaccine's effectiveness is questionable for generally healthy adults.

100(1-α)% C.I. for Mean of Paired Difference: μ_D

- "Real-world" research is often concerned with comparisons between two populations.
- For some research, it might make sense to "pair" the data values. We can then make estimate and draw inferences about the differences between pairs.
  - Paired Differences:

Population 1	Population 2	Difference (d)
X_1	Y_1	$d_1 = Y_1 - X_1$
X_2	Y_2	$d_2 = Y_2 - X_2$
$\vdots$	$\vdots$	$\vdots$
X_n	Y_n	$d_n = Y_n - X_n$

Example:

See Example 8.30 of the Text Book[6].

[6] refer to Page: 484 of the Text Book

- **Note:** Paired differences also arise in "before" and "after" research.

- 100(1-α)% C.I. for Mean of Paired Difference is given by

$$\bar{d} \pm t.\left(\frac{S_d}{\sqrt{n}}\right)$$

where

$$S_d = \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}}$$

Example:

$$S_d = \sqrt{\frac{\sum (d - \bar{d})^2}{n-1}}$$

Twelve subjects participated in an experiment to study the effectiveness of a certain diet, combined with a program of exercise, in reducing serum cholesterol levels. Table 6.4.1 shows the serum cholesterol levels for the 12 subjects at the beginning of the program (Before) and at the end of the program (After).

Table 6.4.1

Serum Cholesterol Levels for 12 Subjects Before and After Diet-Exercise Program

	Serum Cholesterol		Difference (After-Before)
Subject	**Before (X)**	**After (Y)**	$d_i = (Y_i - X_i)$
1	201	200	−1
2	231	236	+5
3	221	216	−5
4	260	233	−27
5	228	224	−4
6	237	216	−21
7	326	296	−30
8	235	195	−40
9	240	207	−33
10	267	247	−20
11	284	210	−74
12	201	209	+8

(Contd.)

$$\bar{d} = \frac{\sum d}{n} = \frac{-242}{12} = -20.17$$

$$\begin{aligned} S_d &= \sqrt{\frac{\sum d^2 - \frac{(\sum d)^2}{n}}{n-1}} \\ &= \sqrt{\frac{1076 - \frac{(-242)^2}{12}}{11}} \\ &= \sqrt{535.06} \\ &= 23.131 \end{aligned}$$

$$df = n - 1 = 12 - 1 = 11$$

- For 95% C.I.

$$\alpha = \frac{1 - 0.95}{2} = 0.025$$

(Contd.)

$$\bar{d} \pm t.\left(\frac{S_d}{\sqrt{n}}\right) = -20.17 \pm 2.20.\left(\frac{23.131}{\sqrt{12}}\right)$$
$$= -20.17 \pm 2.20.(6.677)$$
$$= -20.17 \pm 14.70$$

So our 95% C.I. on the mean difference is from -34.87 to -5.47.

Interpretation:

You Try!

Assumptions:

The differences are approximately normally distributed.

- **IF** n$\geq$ 30 then the values of t in Table F are almost identical to the corresponding Z values in the standard normal Table E.
  So for n$\geq$ 30, we can construct a C.I. on mean of paired difference using

$$\bar{d} \pm Z.\left(\frac{S}{\sqrt{n}}\right)$$