Confidence Intervals

`List of things to Do:-`

- Compute the point estimate of μ
- Compute the confidence intervals about μ with σ known
- Error term in the confidence interval
- Determine the sample size n for estimating the population means
- Explain the meaning of Confidence interval
- SAS command to calculate C.I (Confidence Interval)

Basics

- **A Population** is any entire collection of people, animals, plants or things from which we may collect data. It is the entire group we are interested in, which we wish to describe or draw conclusions about.

  Example:- The population for a study of infant health might be all children born in the U.K. in the $2000's$.

- A **Sample** is a group of units selected from a larger group (the population). By studying the sample it is hoped to draw valid conclusions about the larger group.

  Example:- The population for a study of infant health might be all children born in the U.K. in the $1980's$. The sample might be all babies born on $7th$ May in any of the years.

Basics

- A **Parameter** is a value, usually `unknown` (and which therefore has to be estimated), used to represent a certain population characteristic.

  For example, the population mean is a parameter that is often used to indicate the average value of a quantity.

  μ represents the population **mean**.
  σ represents the population **standard deviation.**
  p represents the population **proportion.**

- A **Statistic** is a quantity that is calculated from a sample of data. It is used to give information about unknown values in the corresponding population.

  For example, the average of the data in a sample is used to give information about the overall average in the population from which that sample was drawn.

  $\bar{x}$ will estimate μ.
  s will estimate σ.
  $\tilde{p}$ will estimate p.

About Normal Distribution

A normal distribution with mean μ and variance σ^2 is a statistic distribution with probability function

$$N(\mu, \sigma) = P(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $x \in (-\infty, \infty)$, $\mu \in R$ and $\sigma \geq 0$.

While statisticians and mathematicians uniformly use the term **'Normal Distribution'** for this distribution, physicists sometimes call it a **Gaussian distribution** and, because of its curved flaring shape, social scientists refer to it as the **Bell Curve.'**

Basics Continued...

Normal to Standard Normal

Let ,

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

$$N(0,1) = P(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$$

At the back of the book the normal table used is based on the above formula.

Basics Continued...

Finding the Area under the Standard Normal Curve

Basics Continued...

Finding the value of Z_α

Basics Continued...

Binomial to Standard Normal

$$B(n,p) = \binom{n}{x} p^x (1-p)^{n-x}$$

where n is the number of trials, r is the number of

success and p is the probability of success.
If $np(1-p) \geq 10$, the binomial experiment is approximately Normal, with mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$.

Standardize the Binomial random variable X we obtain,

$$Z = \frac{x-\mu}{\sigma} = \frac{x-np}{\sqrt{np(1-p)}}$$

Point Estimates

Def:- A point estimate of a parameter is the value of a statistic that estimates the value of the parameter.

Example:- Suppose $X_1, X_2, , X_{10}$ are random samples which follow normal distribution with mean μ and standard deviation $s = 2.5$. Hence in this problem the parameter μ is unknown and we want to estimate μ from the given 10 samples. In order to understand this fact more clearly we need following *definition:-*

Unbiased Estimator:- An estimate s is said to be an unbiased estimator of the parameter μ if $E(S) = \mu$.

Theorem:- Let $X_1, X_2, ..., X_n$ be the random sample from the population with mean μ. The sample mean $\bar{X}$ is an unbiased estimator for the population mean μ.

Solution:- We have to show that $E(\bar{X}) = \mu$.

$$E(\bar{X}) = E(\frac{\sum X_i}{n}) = \frac{1}{n}E(\sum X_i) = \frac{1}{n}\sum E(X_i)$$

But we know that $E(X_i) = \mu$. Therefore the above equation boils down to

$$E(\bar{X}) = \frac{n\mu}{n} = \mu.$$

Note:- From the above proof we observed that $X_i's$ and $\bar{X}$ are the unbiased estimators of μ. Which one is the best one, the $X_i's$ or the $\bar{X}$. The answer is $\bar{X}$ is the better unbiased estimator, also note that we can find many unbiased estimators. A good question is which is the best. The answer to this question is beyond the reach of this Class.

Hence for your example the

$$\frac{X_1 + X_2 + ... + X_{10}}{10} = \bar{X}$$

estimates μ. Hence the unknown population mean has been estimated by some known fact. Isn't it great!!!!!

Computing the confidence intervals about μ with σ known.

Assumption:- The data should be from a normal population or the sample size is greater than 30. The way to check the normality of the data if sample size is less than 30, is to plot the data and see whether it is approximately a straight line with no outliers. I will talk about outliers in SAS.

`Confidence Interval:-` A confidence interval estimate of a parameter consists of an interval of numbers, along with the measure of the likelihood that the interval contains the unknown parameter.

$$(\bar{x} - z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}})$$

- $\bar{x}$ is the point estimate for μ.
- α our level of confidence.
- the standard deviation of the sample mean. (What represents standard deviation of the sample mean $\bar{x}$?)

C.I can be rewritten as

Point Estimate $\pm$ Margin of Error

Interpretation of Confidence Interval

Let X_1,..., X_k denote K samples, each of size n, from a population with mean μ and standard deviation σ(Known). Let us construct $(1-\alpha)100\%$ C.I for each of these samples of size n. Let the C.I's are denoted by C_1,...,C_k. We will observe that, approximately $(1-\alpha)100\%$ of the C.I's will contain μ.

Suppose $k = 20$ and $\alpha = .10$. Suppose 90% confidence interval for μ is calculated as $(25, 36)$. That means "we are 90% confident" that μ lies between 25 and 36. In other words out of 20 C.I's approximately 18 of $C_k's$ will contain μ.

Error term in the confidence interval

The margin of error term in confidence interval is denoted as **E**.

$$\mathbf{E} = z_{\frac{\alpha}{2}} * \frac{\sigma}{\sqrt{n}}$$

The error **E** suggests that if we increase the sample size then we will expect less error, that means the confidence interval will become more shallower, hence will be a good estimator of the unknown population parameter.

Determining the sample size n

The sample size required to estimate the population mean μ, with a given level of confidence and with a specified margin of error(or minimum error tolerance) **E** is given by

$$n = (\frac{z_{\frac{\alpha}{2}} * \sigma}{\mathbf{E}})^2$$

Miles on Cavalier

A researcher is interested in approximating the mean number of miles on three-year old Chevy Cavaliers. She finds a random sample in Orlando, Florida area and obtained the following result

37,815	20,000	57,103	46,585	24,822
49,678	30,983	52,969	8,000	39,862
6,000	65,192	34,285	30,906	41,841
39,851	43,000	74,362	52,664	33,587
52,896	45,280	30,000	41,713	76,315
22,442	43,301	52,899	41,526	28,381
55,163	51,812	36,500	31,947	16,529

- `Obtain a point estimate of the population mean number of miles on a three-year old Cavalier.`

  Ans:- $\bar{x} = 40,463.11$ miles is the point estimate for population mean. That means we can roughly claim that a three-year old Cavalier will have 40,463.11 miles on it.

- `Construct a` 99% `confidence interval for the population mean number of miles on a three-year old Cavalier assuming that` $\sigma = 16,100$.

  Ans:- The 99% confidence interval is given by

  $$(40463.11 - z_{\frac{0.01}{2}} * \frac{16100}{\sqrt{35}} \ , \ 40463.11 + z_{\frac{0.01}{2}} * \frac{16100}{\sqrt{35}})$$

  or

  $$(33456.14 \ , \ 47469.97).$$

  Hence we are 99% confident that the population mean number of miles will lie between 33456.14 and 47469.97.

- From the above result can we say something regarding the population mean number of miles in U.S.A. If not then how nicely the data describes the mean population miles for a three-year old car in an around Florida.

  Ans:- No. As this is a local Florida data.
  The confidence interval is pretty big. Hence we can say that this interval is a moderate estimate of the population mean at 99% level of significance. In order to estimate the population mean more accurately we need to look at other factors, like sex, driving conditions, age group and so on.. .

Computing the confidence intervals about μ with σ uknown.

Assumption:- The data should be from a normal population or the sample size is greater than 30.

Confidence Interval in this case is given by

$$(\bar{x} - t_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}} \quad , \quad \bar{x} + t_{\frac{\alpha}{2}} * \frac{s}{\sqrt{n}})$$

where t denotes the Student's-t distribution with n-1 degrees of freedom. Everything remains the same as above.

Using the same data set from the previous problem Construct a 99% confidence interval for the population mean number of miles on a three-year old Cavalier assuming that $\sigma = unknown$.

Ans:- The 99% confidence interval is given by

$$(40463.11 - t_{\frac{0.01}{2}} * \frac{s}{\sqrt{35}} \ , \ 40463.11 + t_{\frac{0.01}{2}} * \frac{s}{\sqrt{35}})$$

$$= (40463.11 - t_{\frac{0.01}{2}} * \frac{16149.65}{\sqrt{35}} \ , \ 40463.11 + t_{\frac{0.01}{2}} * \frac{16149.65}{\sqrt{35}})$$

$$(33015.17 \ , \ 47911.06).$$

Hence we are 99% confident that the population mean number of miles will lie between 33015 and 47911, when σ is unknown.

SAS Program

```
options linesize=80 nodate; run;
title 'Miles in Three-Year old Chevy Cavalier';
data Chevy;
input miles;
cards;
37815
20000
57103
46585
24822
49678
30983
52969
8000
39862
6000
65192
34285
30906
41841
39851
43000
74362
52664
33587
52896
45280
30000
41713
76315
22442
43301
52899
41526
28381
55163
51812
36500
31947
16529 ;
proc print data=Chevy;
run;
proc means data=chevy;
proc means data=Chevy lclm uclm alpha=0.01 ;
VAR miles;
run;
```

SAS Output

Miles in Three-Year old Chevy Cavalier in Orlando, Florida

Obs	miles
1	37815
2	20000
3	57103
4	46585
5	24822
6	49678
7	30983
8	52969
9	8000
10	39862
11	6000
12	65192
13	34285
14	30906
15	41841
16	39851
17	43000
18	74362
19	52664
20	33587
21	52896
22	45280
23	30000
24	41713
25	76315
26	22442
27	43301
28	52899
29	41526
30	28381
31	55163
32	51812
33	36500
34	31947
35	16529

SAS Output Continued...

The MEANS Procedure

N	Mean	Std Dev	Maximum	Minimum
35	40463.11	16149.65	6000	76315.00

Lower 99% CL for Mean	Upper 99% CL for Mean
33015.17	47911.06

Problem on Sample Size Determination

The dean of a college wants to use the mean of a random sample to estimate the average amount of time students take to get from one class to the next, and she wants to be able to assert with probability 0.95 that her error will be at most 0.30 minute. If she knows from studies of a similar kind that it is reasonable to let $\sigma = 1.50$ minute, how large a sample will she need?

Ans:- The things that are given to us

- $E = 0.30$
- $\sigma = 1.50$
- $z\frac{.05}{2} = 1.96$

We need to find the sample size n.

$$n = (\frac{z_{\frac{\alpha}{2}} * \sigma}{\mathbf{E}})^2 = (\frac{z_{\frac{0.05}{2}} * 1.50}{0.30})^2 = 96.04$$

Computing the confidence intervals about a Population proportion.

Let p denote the population proportion. The best point estimate for p is denoted by $\tilde{p}$.

$$\frac{x}{n} \quad = \quad \tilde{p} \quad \rightarrow \quad p$$

The **Confidence Interval** in this case is denoted by

$$(\tilde{p} - z_{\frac{\alpha}{2}} * \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \quad , \quad \tilde{p} + z_{\frac{\alpha}{2}} * \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}})$$

Determining the Sample size n The sample size required to obtain a $(1-\alpha)\%100$ confidence interval for p with a given margin of error E, is given by

$$n = \tilde{p}(1-\tilde{p})(\frac{z_{\frac{\alpha}{2}}}{E})^2$$

The Death Penalty(8.3 Problem 21)

In a Harris Poll conducted in July, 2000, 64% of the people polled answered yes to the following question: "Do you believe in capital punishment, that is the death penalty, or are you opposed to it?" The margin of Error was ±3%, and the estimate was made with 95% confidence. How many people were surveyed?

Ans:- The things that are given to us

- $E = 0.03$
- $\tilde{p} = 0.64$
- $z_{\frac{.05}{2}} = 1.96$

We need to find the sample size n.

$$n = \tilde{p}(1-\tilde{p})(\frac{z_{\frac{\alpha}{2}}}{E})^2 = 0.64(1-0.64)(\frac{1.96}{0.03})^2 = 983.4$$

Hence 984 people were surveyed.

Computing the confidence intervals about a Population Variance.

If a random sample of size n is obtained from a normally distributed population with mean μ and standard deviation σ, then

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

has a chi-square distribution with $n-1$ degrees of freedom.
The **Confidence Interval** about σ^2 is given by

$$\left(\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}} \quad , \quad \frac{(n-1)s^2}{\chi^2_{\frac{1-\alpha}{2}}}\right)$$

Tests of Hypothesis

`List of things to do:-`

- Determine the null and alternative hypothesis from a claim
- Type I and Type II error
- Calculation of Power

Basic Definitions:- The `Null Hypothesis is`, denoted by H_0, is a statement to be tested. The null hypothesis is assumed to be true until we have evidence against it.
The `Alternative Hypothesis` denoted by H_1, is a claim to be tested.

Ways to set up H_0 and H_1:-

Figuring out `Null` and `Alternative` Hypothesis from a given data.

The following data is a survey of 15 Blood pressure patients before and after taking B.P medication for 6 months.

No. of Observations	Before Medication	After Medication
1	140	134
2	123	126
3	150	130
4	174	140
5	134	120
6	147	142
7	132	125
8	150	120
9	190	150
10	162	133
11	167	135
12	168	142
13	145	129
14	145	120
15	159	130

Let μ_0 denote the mean Blood Pressure **before** taking Medication.
Let μ_1 denote the mean Blood Pressure **after** taking Medication.

$$H_0 : \mu_0 = \mu_1$$

$$H_1 : \mu_0 > \mu_1$$

Hence we claim that after taking medication the Blood Pressure will go down.

Different Types of Hypothesis Testing

- `Simple vs Simple`
  Example:-

$$H_0 : \mu = 0$$
$$H_1 : \mu = 3$$

- `Simple vs Composite`
  Example:-

$$H_0 : \mu = 0$$
$$H_1 : \mu \geq 0$$

- `Composite vs Composite`
  Example:-

$$H_0 : \mu < 0$$
$$H_1 : \mu \geq 0$$

Type I and Type II Error

	H_0 `Is true`	H_1 `Is True`
`Do not Reject` H_0	$\surd$	Type II Error
`Reject` H_0	Type I Error	$\surd$

- Hence a Type I Error would occur if the Null Hypothesis is Rejected when, in Reality the Null Hypothesis is True.
- Power of a Hypothesis Test is the Probability of Rejecting Null Hypothesis when, in reality H_0 is False. Hence its a correct decision.

We are now ready to define the Level of Significance and Power of a Hypothesis Test.

Level of Significance:- The level of Significance, α, is the probability of making a Type I Error.

$$\alpha = P(RejectingH_0|H_0isTrue)$$

Probability of type II Error:- Probability of Type II Error is denoted by β and is defined as

$$\beta = P(NotRejectingH_0 \quad | \quad H_1istrue).$$

Probability of type II Error:-

Remarks:- In problems we fix α, as .05(95%), .01(99%), .1(90%) . That means we fix the margin of error and the small it is , the good for the experiment. In general it is known that Type I error is more severe than Type II error, hence always want a small α. On the other hand we will like to get a bigger Power to decide how powerful the test is.

Problem 27. `Potato Consumption:-` According to the Statistical Abstract of the United States, the mean per capita consumption of potatoes in 1999 was 48.3 pounds. A Researcher believes that the potatoes consumption has risen since then.

1. Determine the Null and the Alternative Hypothesis.
   Ans:-
   $$H_0 : \mu = 48.3$$
   $$H_1 : \mu > 48.3$$

2. Suppose, in reality the mean capita consumption of potatoes is 48.3 pounds. Was a Type I or Type II error committed. If we test this hypothesis at the $\alpha = 0.05$ level of significance, what is the probability of committing a Type I error.
   Ans:- Hence H_0 is true and its rejected, we are committing a Type I Error. And the Probability of Type I Error is $\alpha = 0.05$.

Important Numbers

Confidence Level	α	z_α
90%(One Sided Test)	0.100	1.285
95%(One Sided Test)	0.050	1.645
95%(Two Sided Test)	0.025	1.960
98%(One Sided Test)	0.02	2.055
99%(One sided Test)	0.010	2.325
99%(Two Sided Test)	0.005	2.575

Note:-

- z_α is a Number ranging from $-\infty$ to $+\infty$.
- If the significance level is not stated then assume that we are talking about 95% confidence level and hence we will use $\alpha = 0.05$.

The Critical Region or Rejection Region for Two Tailed, Left Tailed and Right Tailed Hypothesis

Two Tailed:-

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Rejection Region:-

$$Z_{Cal} > Z_{\frac{\alpha}{2}}.$$

or

$$Z_{Cal} < -Z_{\frac{\alpha}{2}}.$$

Where $Z_{Cal} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Left Tailed:-

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu < \mu_0$$

Rejection Region:-

$$Z_{Cal} < -Z\alpha.$$

Where $Z_{Cal} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Right Tailed:-

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

Rejection Region:-

$$Z_{Cal} > Z\alpha.$$

Where $Z_{Cal} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Problems of Power

For a given population with $\sigma = \$12$, we want to test the null hypothesis $\mu = \$75$ on the basis of random sample of size $n = 100$. If the Null hypothesis is rejected when $\bar{x}$ is greater than or equal to \$76.50. and otherwise it is accepted, find

- the probability of Type I error.

  Ans:-
  α = Probability of Type I Error = P(Reject $H_0 | H_0$ is True)
  The rejection region is given as $\bar{x} \geq 76.50$.
  Mathematically,

$$\alpha = P(\bar{x} \geq 76.50 \mid H_0 is True)$$

$$\alpha = P(\bar{x} \geq 76.50 | \ \mu_0 = \$75)$$

$$\alpha = P(Z = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{76.50-\mu_0}{\frac{\sigma}{\sqrt{n}}}) = P(Z \geq \frac{76.50-75}{\frac{12}{\sqrt{100}}})$$

$$\alpha = P(Z \geq \frac{76.50-75}{\frac{12}{\sqrt{100}}}) = P(Z \geq 1.25) = \mathbf{0.1056}$$

  Hence level of confidence is approx. 90%.

- The power of the hypothesis when $\mu_0 = 75.30$. Is it a good power?
  Ans:-
  Power = P(Reject $H_0|H_1$ is True)
  Mathematically, Power = $P(\bar{x} \geq 76.50|\ \mu_0 = 75.30)$

$$
\begin{aligned}
Power = P(Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{76.50 - \mu_0}{\frac{\sigma}{\sqrt{n}}}) \\
= P(Z \geq \frac{76.50 - 75.30}{\frac{12}{\sqrt{100}}}) \\
= P(Z \geq \frac{76.50 - 75.30}{\frac{12}{\sqrt{100}}}) \\
= P(Z \geq 1) = \mathbf{0.1587}
\end{aligned}
$$

  Conclusion:- Its a Awful Power. A good Power should be close to 1. And .1587 is not.

- The probability of Type II Error when $\mu = 77.22$. Is it good?
  Ans:-
  Power = P(Reject $H_0|H_1$ is True)
  Mathematically, Power = $P(\bar{x} \geq 76.50|\ \mu_0 = 77.22)$

$$
\begin{aligned}
Power = P(Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \geq \frac{76.50 - \mu_0}{\frac{\sigma}{\sqrt{n}}}) \\
= P(Z \geq \frac{76.50 - 77.22}{\frac{12}{\sqrt{100}}}) \\
= P(Z \geq \frac{76.50 - 77.22}{\frac{12}{\sqrt{100}}}) \\
= P(Z \geq -0.6) = \mathbf{0.7257}
\end{aligned}
$$

  Conclusion:- Its a Good Power. A good Power should be close to 1. And .7257 is close to 1.

Problem 5, Sec 9.6:- In order to test $H_0 : \mu = 20$ v.s $H_1 : \mu < 20$, a simple random sample of size $n = 18$ is obtained from a population that is known to be normally distributed with $\sigma = 3$.

- What would it mean to make a Type II Error?
  Ans:- A Type II Error would occur if the sample data led to a conclusion of not rejecting $\mu = 20$ when in fact $\mu < 20$.

- If the researcher decides to test this hypothesis at $\alpha = 0.05$ level of significance, compute the Power of the test if the true population mean is 17.4. What is P(Type II Error)?
  Ans:-

$$\text{Power} = \text{P(Reject } H_0 | H_1 \text{ is True)}$$

`Computing The Rejection Region:-` Observe that this is a **Left Tailed Test** , hence the rejection region is given in terms of $\bar{x}$ as follows

$$\bar{x} < \mu - z_\alpha \frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} < 20 - 1.645 \frac{3}{\sqrt{18}} = 18.84$$

$$\text{Power} = \text{P(Reject } H_0 | H_1 \text{ is True)}$$

$$= \text{P}(\bar{x} < 18.84 \ \mid \ \mu = 17.40)$$

Standardizing we get

$$= P(Z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \leq \frac{18.84 - \mu_0}{\frac{\sigma}{\sqrt{n}}}) = P(Z \leq \frac{18.84 - 17.40}{\frac{3}{\sqrt{18}}})$$

$$= P(Z \leq 2.04)$$

$$= 0.9793$$

Geometrically:-

- Redo part b if the true population mean is 19.2.

Problem 12, Sec 9.6:- A school administrator states that students whose first language learned is not english do not score differently on the Math portion of the SAT Exam from students whose first language is English. The mean SAT math score of students whose first language is english is 516., according to the data obtained from the College Board. Suppose the researcher obtains a simple random sample of 20 students whose first language learned was not english, SAT math scores are normally distributed with a population standard standard deviation of 109.

- What would it mean to make a Type II Error?
  Ans:- A Type II Error would occur if the sample data led to a conclusion of not rejecting $\mu = 516$ when in fact $\mu \neq 516$.

- If the researcher decides to test this hypothesis at $\alpha = 0.05$ level of significance, compute the Probability of Type II Error if the true population mean is 505. What is Power of the test?

Ans:-

P(Type II Error) = P(**Do Not Reject** $H_0|H_1$ is True)

Computing The Rejection Region:- Observe that this is a **Two Tailed Test** , hence the rejection region is given in terms of $\bar{x}$ as follows

$$\bar{x} > \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \text{ or } \bar{x} < \mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

Hence do not reject H_0 if , $\mu - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \bar{x} < \mu + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

or

$$516 - z_{\frac{0.05}{2}}\frac{109}{\sqrt{20}} < \bar{x} < 516 + z_{\frac{0.05}{2}}\frac{109}{\sqrt{20}}$$

$$516 - 1.96\frac{109}{\sqrt{20}} < \bar{x} < 516 + 1.96\frac{109}{\sqrt{20}}$$

Hence D not reject Reject H_0 if $\bar{x}$ lies $(468.23, 563.77)$.

P(Type II Error) = P(**Do Not Reject** $H_0|H_1$ is True)

$$= \text{P}(468.23 < \bar{x} < 563.77|\mu = 505)$$

Standardizing we get..

$$= P(\frac{468.23-505}{\frac{109}{\sqrt{20}}} < Z < \frac{563.77-505}{\frac{109}{\sqrt{20}}})$$

$$= P(-1.51 < Z < 2.41)$$
$$= .9920 - 0.0655 = 0.9265 = \beta$$

Hence the **Power** of this Two Tailed Test is
$1 - .9265 = 0.0735$. Which is a `Bad Power.`

The Classical method of Testing Hypothesis about μ, σ known

Steps used in Classical method of Testing Hypothesis about μ, σ known...

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Write down α value given in the problem and find the Critical Value Z_{tab}, z_α or $z_{\frac{\alpha}{2}}$ depending on the nature of the test. Also Draw the rejection region or the critical region.

Step 3 Compute the test Statistic

$$Z = Z_{Cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Step 4 Compare the critical value from the test Statistic. In other words compare Z_{Cal} with Z_{tab}.

Step 5 State the conclusion.

Note:- The procedure presented requires that the data comes from a population that is Normally distributed or the sample size is more that 30.

Problem 12, Sec 9.2:- A school administrator states that students whose first language learned is not english do not score differently on the Math portion of the SAT Exam from students whose first language is English. The mean SAT math score of students whose first language is english is 516., according to the data obtained from the College Board. Suppose a simple random sample of 20 students whose first language learned was not english, results in a sample mean SAT math score of 518. SAT math scores are normally distributed with a population standard standard deviation of 109.

- Why it is necessary for SAT math scores to be normally distributed in order to test the claim using the Steps?

  Ans:- Since the sample size is less than 20, we need the SAT math scores to be Normally distributed.

- Test the Researcher's claim using the Classical Approach at the $\alpha = 0.10$ level of significance.
  Ans:-

  The Classical Approach

  Step 1

  $$H_0 : \mu = 516$$
  $$H_1 : \mu \neq 516$$

  Step 2 $\alpha = 0.10$. As this is a two tailed test the rejection region is given as $Z \geq Z_{\frac{\alpha}{2}} = 1.645$ or $Z \leq -Z_{\frac{\alpha}{2}} = -1.645$. Geometrically

  Step 3 The Test Statistic

  $$Z_{Cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{518 - 516}{\frac{109}{\sqrt{20}}} = 0.08$$

  Step 4 As the test statistic Z_{Cal} doesn't lie in the critical region, we do not reject the Null Hypothesis.

  Step 5 Hence there is `not enough evidence` to support the claim that the mean SAT math score of Students whose first language is not english is different from 516.

Testing a Hypothesis about μ, σ known, p- Value Approach

Steps used in Testing Hypothesis about μ, σ known, p-value approach...

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Compute the test Statistic

$$Z = Z_{Cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Step 3 Determine the p- value.

For `Two tailed` test the P-Value is
$2P(Z > |Z_{Cal}|)$.
For `Left tailed` test the P-Value is $P(Z < Z_{Cal})$.
For `Right tailed` test the P-Value is
$P(Z > Z_{Cal})$.

Step 4 Reject the Null Hypothesis if the P-Value is **less** than the level of Significance α.

Step 5 State the conclusion.

Same Problem 12, Sec 9.2:- A school administrator states that students whose first language learned is not English do not score differently on the Math portion of the SAT Exam from students whose first language is English. The mean SAT math score of students whose first language is English is 516., according to the data obtained from the College Board. Suppose a simple random sample of 20 students whose first language learned was not English, results in a sample mean SAT math score of 518. SAT math scores are normally distributed with a population standard standard deviation of 109.

Test the Researcher's claim using the Classical Approach at the $\alpha = 0.10$ level of significance. **Ans:-**

The p-value Approach

Step 1

$$H_0 : \mu = 516$$

$$H_1 : \mu \neq 516$$

Step 2 The Test Statistic

$$Z_{Cal} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{518 - 516}{\frac{109}{\sqrt{20}}} = 0.08$$

Step 3 p- Value is $2P(Z > |0.08|) = 2P(Z > 0.08) = 2(1 - 0.5319) = 2(0.4681) = 0.9362$

Step 4 As the test statistic p-value is not less than $\alpha = 0.05$, we do not reject the Null Hypothesis.

Step 5 Hence there is `not enough evidence` to support the claim that the mean SAT math score of Students whose first language is not English is different from 516.

T Test

The Classical method of Testing Hypothesis about μ, σ unknown

Steps used in Classical method of Testing Hypothesis about μ, σ unknown...

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Write down α value given in the problem and find the Critical Value t_{tab}, t_{α} or $t_{\frac{\alpha}{2}}$ and also mention the degrees of freedom. Also Draw the rejection region or the critical region.

Step 3 Compute the test Statistic

$$t = t_{Cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Step 4 Compare the critical value from the test Statistic. In other words compare t_{Cal} with t_{tab}.

Step 5 State the conclusion.

Note:- The procedure presented requires that the data comes from a population that is Normally distributed or the sample size is more that 30.

Testing a Hypothesis about μ, σ unknown, p- Value Approach

Steps used in Testing Hypothesis about μ, σ unknown, p-value approach...

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Compute the test Statistic

$$t = t_{Cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Step 3 Determine the p- value.

Step 4 Reject the Null Hypothesis if the P-Value is **less** than the level of Significance α.

Step 5 State the conclusion.

Problem 15, Sec 9.3:- In 1989, the average age of an inmate on a death row was 36.2 years of age, according to the data obtained from U.S Department of justice. A sociologist wants to test the **Claim that the average age of a death row inmate has changed since then**. She randomly selects 32 death row inmates and finds that there mean age is **38.9**, with a standard deviation of 9.6.

- Using the Classical approach test the sociologist's claim at the $\alpha = 0.05$ level of significance.
  Ans:-

  Step 1

  $$H_0 : \mu = 36.2$$
  $$H_1 : \mu \neq 36.2$$

  Step 2 This is a two sided test, with 31 df and so the critical values are $\pm t_{0.025} = \pm 2.040$.

Step 3

$$t = t_{Cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{38.9 - 36.2}{\frac{9.6}{\sqrt{32}}} = 1.591$$

.

Step 4 1.591 doesn't lie in the critical region, hence we will not reject the NULL HYPOTHESIS.

Step 5 Hence there is not enough evidence to support the claim that the mean age of death -row inmates is different from 36.2 years.

- Determine and Interpret the P-value.
  Ans:-

  How to find the P value from given t_{Cal} and df?

Solution to Problem 20, Sec 9.3:-

Step 1

$$H_0 : \mu = 1.62$$
$$H_1 : \mu < 1.62$$

Step 2 This is a one sided test, with 11 df and so the critical values are $-t_{0.10} = -1.363$.

Step 3

$$t = t_{Cal} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{1.6141 - 1.62}{\frac{0.0053}{\sqrt{12}}} = -3.856$$

.

Step 4 -1.363 lie in the critical region, hence we reject the NULL HYPOTHESIS.

Step 5 Hence there is enough evidence to support the claim that the mean wt of Maxfli XS Golf balls is less than 1.62 ounces.

Determine and Interpret the P-value.

Ans:-

`How to find the P value from given` t_{Cal} `and df?`

The Classical method of Testing Hypothesis about population proportion

Steps used in Classical method of Testing Hypothesis about population proportion

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Write down α value given in the problem and find the Critical Value Z_{tab}, z_α or $z_{\frac{\alpha}{2}}$ depending on the nature of the test. Also Draw the rejection region or the critical region.

Step 3 Compute the test Statistic

$$Z = Z_{Cal} = \frac{\tilde{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Where $\tilde{p}$ is the sample Proportion.

Step 4 Compare the critical value from the test Statistic. In other words compare Z_{Cal} with Z_{tab}.

Step 5 State the conclusion.

Note:- For Normality you should check $np_0(1-p_0) \geq 10$

Solution to Problem 12, Sec 9.4:-

Step 1

$$H_0 : p_0 = 0.249$$

$$H_1 : p_0 < 0.249$$

Note:-
$np_0(1 - p_0) = 150 * 0.249(1 - 0.249) = 28 \geq 10$,
hence the normality requirement is fulfilled.

Step 2 $\alpha = 0.05$. As this is a Left tailed test the rejection region is given as $Z \leq -Z_\alpha = -1.645$.
From the Survey, $\tilde{p} = \frac{28}{150} = 0.187$

Step 3 The Test Statistic

$$Z_{Cal} = \frac{\tilde{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.187 - 0.249}{\sqrt{\frac{0.249(1-0.249)}{150}}} = -1.76$$

Step 4 As the test statistic Z_{Cal} lies in the critical region, we reject the Null Hypothesis.

Step 5 Hence there is `enough evidence` to support the claim that the percentage of 30-40 year old male who do not exercise has decreased from its 1998 level.

Inference on Two Samples

- Paired Tests(Dependent Samples).
- Welch's t test(Independent Samples).
- Pooled Two sample t test.
- F Test

Paired Tests(Dependent Samples).

Definition:- The sampling method is **dependent** when the individuals selected in one sample are used to determine the individuals to be in the second sample. As an example we use Paired T test, when, Before and After kinds of comparison, the ages of Husbands and wives, first half and second half of the exam, cars stocked and sold by used car dealers, and numerous other kinds of situations in which data are naturally Paired. Steps..

Assumptions:-
1.The sample is obtained using simple random sampling.
2.the sample data are paired.
3. the differences are normally distributed or , the sample size is greater than 30.

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Write down α value given in the problem and find the Critical Value t_{tab}, t_α or $t_{\frac{\alpha}{2}}$ and also mention the degrees of freedom. Also Draw the rejection region or the critical region.

Step 3 Compute the test Statistic

$$t = t_{Cal} = \frac{\bar{d} - 0}{\frac{s_d}{\sqrt{n}}} = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

Step 4 Compare the critical value with the test Statistic. In other words compare t_{Cal} with t_{tab}.

Step 5 State the conclusion.

Confidence Intervals

A $(1-\alpha).100\%$ confidence interval for μ_d is given by

$$(\bar{d} - t_{\frac{\alpha}{2}}.\frac{s_d}{\sqrt{n}}, \bar{d} + t_{\frac{\alpha}{2}}.\frac{s_d}{\sqrt{n}})$$

Problem 8, Sec 10.1:-

Observations	1	2	3	4	5	6	7	8
X_1	19.4	18.3	22.1	20.7	19.2	11.8	20.1	18.6
X_2	19.8	16.8	21.1	22.0	21.5	18.7	15.0	23.9

1. Determine $d_i = X_1 - X_2$.

   Ans:-

X_1	X_2	$X_1 - X_2$
19.4	19.9	-0.4
18.3	16.8	1.5
22.1	21.1	1.0
20.7	22.0	-1.3
19.2	21.5	-2.3
11.8	18.7	-6.9
20.1	15.0	5.1
18.6	23.9	-5.3

2. Compute $\bar{d}$ and s_d.

   Ans:-

   options linesize=80 nodate; run;
   data pair;
   infile 'path'
   input chaos;
   proc means data=pair ;
   var chaos;
   run;

`Output`

3. Test the claim that $\mu_d \neq 0$ at the $\alpha = 0.01$ level of significance.

4. Compute the 99% confidence Interval about the population mean difference $\mu_d = \mu_1 - \mu_2$.

Welch's t test(Independent Samples).

Definition:- The sampling method is **independent** when the individuals selected in one sample do not dictate which individuals are to be in the second sample.
Assumptions:-
1.The sample is obtained using simple random sampling.
2.the samples are independent.
3. the population from which the samples are drawn are normally distributed or the sample sizes are greater than 30.

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Write down α value given in the problem and find the Critical Value t_{tab}, t_α or $t_{\frac{\alpha}{2}}$ and the degrees of freedom will be smaller of $n_1 - 1$ or $n_2 - 1$. Also Draw the rejection region or the critical region.

Step 3 Compute the test Statistic

$$t = t_{Cal} = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

Step 4 Compare the critical value with the test Statistic. In other words compare t_{Cal} with t_{tab}.

Step 5 State the conclusion.

Improving the Conservative df of Welch's t test

The Welch's t test is conservative, in order to get accurate answer we should use

$$df = \frac{(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2})^2}{\frac{(\frac{{s_1}^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{{s_2}^2}{n_2})^2}{n_2 - 2}}$$

Software will use this particular formula for more accurate answer.

Confidence Intervals about the difference of two means

A $(1-\alpha).100\%$ confidence interval for μ_d is given by

$$(\bar{x_1} - \bar{x_2}) \pm t_{\frac{\alpha}{2}} * \sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}.$$

where $t_{\frac{\alpha}{2}}$ is obtained using smaller of $n_1 - 1$ or $n_2 - 1$ degrees of freedom or the improved formula.

Problem 9, Sec 10.2:- An engineer wanted to know whether the strength of two different concentrate mix designs differed significantly. He randomly selected 9 cylinders, measuring 6 inches in diameter and 12 inches in height, into which mixture 67-0-301 was poured. After 28 days, he measured the strength(in pounds per square inch) of the cylinder. He also randomly selected 10 cylinders of mixture 67-0-400 and performed the same test. the results are as follows:-

Mixture 67-0-301	**Mixture 67-0-400**
3960	4070
4090	4890
3100	5020
3830	4330
3200	4640
3780	5220
4080	4190
4040	3730
2940	4120
—	4620

1. The data are obtained from simple random sample that are independent.

2. test the claim that mixture 67-0-400 is stronger than 67-0-301 at $\alpha = 0.05$ level of significance.

Ans:-

Step 1

$$H_0 : \mu_{400} = \mu_{301}$$

$$H_1 : \mu_{400} > \mu_{301}$$

μ_{400} = Mean strength of mixture 67-0-400 .
μ_{400} = Mean strength of mixture 67-0-301 .

Step 2 This is a right tailed test. df = $\min\{9, 8\}$.
$t_{cri} = 1.860$. Hence Reject H_0 if $t_{tab} > 1.860$.

Step 3 $\bar{x}_{400} = 4483, n_{400} = 10$

$\bar{x}_{301} = 3669, n_{301} = 9$

$$t = t_{Cal} = \frac{(\bar{x}_{400} - \bar{x}_{301}) - (\mu_{400} - \mu_{301})}{\sqrt{\frac{s_{400}^2}{n_{400}} + \frac{s_{301}^2}{n_{301}}}} = 3.8$$

Step 4 As $3.8 > 1.860$, we reject H_0.

Step 5 Hence there is enough evidence to conclude that the mixture 67-0-400 is stronger than 67-0-301.

3. Construct a 90% confidence interval about $\mu_{400} - \mu_{301}$, and interpret the result.

Ans:- The confidence interval is given by

$$(\bar{x}_{400} - \bar{x}_{301}) \pm t_{\frac{\alpha}{2}} * \sqrt{\frac{s_{400}^2}{n_{400}} + \frac{s_{301}^2}{n_{301}}}.$$

$$= 814 \pm 398$$

$$= (416, 1212)$$

We are 90% confident that the the true mean

difference is between 416 and 1212.

`Pooled Two sample t test.`

Pooled t-test is used for independent sample case and it assumes equal population variance. While Welch's t-test doesn't assume any thing about the population variance. Comparison between Welch's t-test and Pooled t-test,

Pooled t-test
Assumes that population variance is same
$df = n_1 + n_2 - 2$
$t = t_{Cal} = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ where $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
Not conservative as compared to Welch's t-test
P(Type II Error) is normally not so high as Welch's t-test
Disadvantage is how to find out the fact that the population variance are the same? The answer is F-Test which we will learn later in this section.

Problems of Pooled t-test

The following random samples are measurements of the heat-producing capacity (in millions per calories of tons) of coal from two mines:-

Mine 1:- 8380 8180 8500 7840 7990
Mine 2:- 7660 7510 7910 8070 7790

Use the 0.05 level of significance to test wether the difference between the means of these two samples is significant, assuming that $\sigma_1 = \sigma_2$.

Ans:-

Step 1

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Step 2 $\alpha = 0.05$, $df = n_1 + n_2 - 2 = 8$, Reject H_0 if $|t_{cal}| > t_{cri} = 2.306$.

Step 3 $\bar{x}_1 = 8178$, $s_1 = 271.1$,$\bar{x}_2 = 7788$, $s_2 = 216.8$.

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} = 245.5$$

$$t_{Cal} = \frac{(\bar{x_1} - \bar{x_2})}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = 2.51$$

Step 4 As $2.51 > 2.306$, we Reject the Null Hypothesis.

Step 5 Hence there is enough evidence to conclude that the difference between the two sample mean is significant.

Inference about Two Sample Population Standard Deviation

`When to decide that the population Variances are Equal?`

In order to answer this question we need the following assumption:-

- The samples are independent simple random samples.
- The population from which the samples are drawn must be normally distributed.

Introduction of Fisher's F-distribution

If $\sigma_1^2 = \sigma_2^2$ and s_1^2 and s_2^2 are sample variances from independent simple random samples of size n_1 and n_2 respectively, drawn from normal populations, then

$$F = \frac{s_1^2}{s_2^2}$$

follows F-Distribution with $n_1 - 1$ numerator degrees of freedom and $n_2 - 1$ denominator degrees of freedom. About critical values of F-Distribution:-

Steps for Hypothesis Tests on Two Population Standard Deviation

Step 1 Figuring out the **Null** and **Alternative** Hypothesis.

Step 2 Write down α value given in the problem and find the Critical Value F_{cri} as follows:-

- **Two Tailed**
  $F_{cri} = F_{1-\frac{\alpha}{2},n_1-1,n_2-1}$ and $F_{cri} = F_{\frac{\alpha}{2},n_1-1,n_2-1}$.

- **Left Tailed**
  $F_{cri} = F_{1-\alpha,n_1-1,n_2-1}$.

- **Right Tailed**
  $F_{cri} = F_{\alpha,n_1-1,n_2-1}$.

Step 3

$$F_{cal} = \frac{s_1^2}{s_2^2}$$

Step 4 F_{cal} lies in the critical region or not?

Step 5 State the conclusion.

Calculate $F_{0.975,12,14}$

Problem 11, Sec 10.4:-

	Sample for population 1	Sample for Population 2
n	26	19
s	9.9	6.4

Test the claim that $\sigma_1 > \sigma_2$ at $\alpha = 0.01$ level of significance for the given sample data.

Ans:-

Step 1

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 > \sigma_2$$

Step 2 $\alpha = 0.01$. Right tailed test. Therefore $F_{cri} = F_{0.01,26-1,19-1} = F_{0.01,25,18} = 2.84$. Reject H_0 if $F_{cal} > 2.84$.

Step 3

$$F_{cal} = \frac{s_1^2}{s_2^2} = \frac{9.9^2}{6.4^2} = 2.39$$

Step 4 As F_{cal} is not in the critical region , we do not reject H_0.

Step 5 There is not enough evidence to support the claim that $\sigma_1 > \sigma_2$.

χ^2 Goodness of fit Test

`Definition:-` A **goodness-of-fit test** is an inferential procedure used to determine whether a frequency distribution follows a claimed distribution.

In other words A statistical test in which the validity of one hypothesis is tested without specification of an alternative hypothesis is called a **goodness-of-fit test.**The idea behind the chi-square goodness-of-fit test is to see if the sample comes from the population with the claimed distribution. Another way of looking at that is to ask if the frequency distribution fits a `specific pattern`.
`Two values` are involved, an observed value, which is the frequency of a category from a sample, and the other is expected frequency, which is calculated based upon the claimed distribution.(Sometimes know as expected counts $E_i = np_i$)
The idea is that if the `observed frequency is really close to the claimed (expected) frequency`, then the square of the deviations will be small. The square of the deviation is divided by the expected frequency to weight frequencies.

The test statistics

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad i = 1, 2, ..., k$$

approximately follows a Chi-Square Distribution with $n - 1$ degrees of freedom if the following assumptions are met.

- The data are obtained from a random sample.
- All expected frequency must be at least 1.
- At most 20% of the expected frequencies are less than 5.

`The following are properties of the goodness-of-fit test`

1. The data are the observed frequencies. This means that there is only one data value for each category.
2. The degrees of freedom is one less than the number of categories, not one less than the sample size.
3. It is always a **right tail test**.
4. It has a chi-square distribution. The value of the test statistic doesn't change if the order of the categories is switched.

Steps for Testing A Claim Using a Goodness-of-Fit Test:-

Step 1 A claim is made regarding a distribution. The claim is used to determine Null and Alternative Hypothesis.
H_0: The random variable follows the claimed distribution.

Step 2 Calculate the Expected frequency for each of the K categories.

Step 3 Verify the assumptions for goodness -of- fit Test.

Step 4 Compute the test Statistic

$$\chi^2{}_{Cal} = \sum \frac{(O_i - E_i)^2}{E_i}.$$

Step 5 Find $\chi^2{}_{cri} = \chi^2{}_{\alpha}$ with $k - 1$ degrees of freedom.

Step 6 Reject H_0 if $\chi^2{}_{cal} > \chi^2{}_{cri}$.

Step 7 State the conclusion.

Bicycle Deaths:- A researcher wanted to determine whether bicycle deaths were uniformly distributed over the days of the week. She `randomly` selected 200 deaths that involved a bicycle, recorded the day of the week on which the day occurred, and obtained the following results:

Day of week	Frequency
Sunday	16
Monday	35
Tuesday	16
Wednesday	28
Thursday	34
Friday	41
Saturday	30

Is there reason to believe that the day of the week on which a fatality occurs on a bicycle occurs with equal frequency at the $\alpha = 0.05$ level of significance?

Ans:-

Step 1

$$H_0 : p_1 = p_2 = p_3 = p_4 = p_5 = p_6 = p_7 = \frac{1}{7}$$

H_0: at least one of the p_i will is different from $\frac{1}{7}$.

Step 2,3,4 will follow from the following diagram.

Observed Count(O_i)	Expected Count(E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
16	$200(\frac{1}{7}) = 28.57$		
35	28.57		
16	28.57		
28	28.57		
34	28.57		
41	28.57		
30	28.57		

Step 5 $df = 7 - 1 = 6, {\chi^2}_{0.05} = 12.592$

Problems of χ^2 Goodness of fit Test

1. H_0 : The random variable X is **binomial with** $n = 4$**,** $p = 0.3$.
H_1 : The random variable X is not binomial with $n = 4$, $p = 0.3$.

X	0	1	2	4	5
Observed	260	400	280	50	10

- Find the Expected Frequency.
- Determine the χ^2 test Statistic.
- Determine the degrees of freedom.
- Test the hypothesis at the $\alpha = 0.05$ level of significance.

Problems of χ^2 Goodness of fit Test

2.According to the manufacturer of $M\&Ms$, 30% of the plain $M\&Ms$ in a bag should be brown, 20% yellow, 20% red, 10% orange, 10% blue, 10% green. A student wanted to determine whether a randomly selected bag of plain $M\&Ms$ had contents that followed this distribution. He counted the number of $M\&Ms$ that were colored and obtained the following results.

Color	Frequency
Brown	125
Yellow	77
Red	90
Blue	31
Orange	42
Green	35

Test the claim that plain $M\&Ms$ follow the distribution stated by manufacturer of $M\&Ms$ at 0.05 level of significance.

Problems of χ^2 Goodness of fit Test

3.

Number of Radio Messages	Observed Frequency
0	70
1	57
2	46
3	20
4	5
5	2

Number of Radio Messages	Expected Frequency
0	44.6
1	66.9
2	50.2
3	25.1
4 or more	13.1

Note:- `The expected frequency table was constructed using Poisson Distribution with` $\lambda = 1.5$.

H_0 : The population sampled has the Poisson distribution with $\lambda = 1.5$.

Test the claim at 0.01 level of significance.

Curve Fitting

Model : $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
Where
1.X_i is the independent variable (predictor variable)
2.Y_i is the dependent variable or (response variable)
3. β_1 is the slope and β_0 is the Y-intercept.
4. ϵ_i is the Error term which follows $N(0, \sigma^2)$.

This model is a population (notice the Greek $\beta's$ and ϵ known as estimators) simple (meaning one X variable) linear (meaning straight line) regression model.
If we take the `expected value` (average) of the Regression model it becomes $E(Y_i) = \beta_0 + \beta_1 X_i$.

(**1**) The disappears because $E(\epsilon_i) = 0$, that is we expect our error to be zero.
(**2**) This model indicates that given a level of X_i , you can generate the mean of the probability distribution of the Y_i's at that X_i level.

We must `estimate the population Regression line using sample data.`
The equation of the Least-squares Regression Line is given by

$$\hat{Y}_i = b_1 x + b_0$$

b_1 and b_0 are the estimates for β_1 and β_0 respectively. We can calculate b_1 and b_0 from Normal equations. ("y-hat" is the Predicted Value for any given x.)

Normal Equations

$$\sum y = nb_0 + b_1 \sum x$$
$$\sum xy = b_0 \sum x + b_1 \sum x^2$$

where Σ means summing over all the data points. Solving the above Normal equations we obtained b_1 and b_0 as follows.

$$b_1 = \frac{\sum xy - \frac{1}{n}(\sum x)(\sum y)}{\sum x^2 - \frac{1}{n}(\sum x)^2} = \frac{S_{xy}}{S_{xx}}$$

$$b_0 = \frac{\sum y - b_1(\sum x)}{n}$$

Where

$$S_{xx} = \sum x^2 - \frac{1}{n}(\sum x)^2$$

is the variance of observed $x_i's$ multiplied by the factor $n-1$.

We will use similar notations later on like

$$S_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2$$

is the variance of observed $y_i's$ multiplied by the factor $n-1$.

`A few more Notations...`

Standard Deviation of $x_i's$ is denoted by $s_x = \sqrt{\frac{S_{xx}}{n-1}}$.

Standard Deviation of $y_i's$ is denoted by $s_y = \sqrt{\frac{S_{yy}}{n-1}}$.

What are the steps we require to fit a Least Square Line from the given Data

1. Plot the data. Equivalently draw the **Scatter Diagram.**

2. Write down the Population Simple Linear Regression Model.
   $(y = \beta_0 + \beta_1 x + \epsilon)$.

3. Two **Normal equations.**

4. Solve Normal Equations to find b_1 and then b_0.
   Plug in back b_1 and b_0 in the Model.

5. Fit the Least Square Line in the Scatter Diagram.

Problems on Curve Fitting

Number of weeks x	Hearing Range y
47	15.1
56	14.1
116	13.2
178	12.7
19	14.6
75	13.8
160	11.9
31	14.8
12	15.3
164	12.6
43	14.7
74	14.0

Here x is the length of time that a person has been living near a major airport directly in the flight path of departing jets, and y is his or her hearing range(in thousands of cycles per second)

Solution:- Solved in the class.

Multiple regression

Model:- $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

The least square multiple regression line is

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$$

The normal equations from which b_0, b_1, b_2 are obtained are,

$$\sum y = nb_0 + b_1 \sum x_1 + b_2 \sum x_2$$
$$\sum yx_1 = b_0 \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2$$
$$\sum yx_2 = b_0 \sum x_2 + b_1 \sum x_1 x_2 + b_2 \sum x_2^2$$

Important Formulas

$$S_{xx} = \sum x^2 - \frac{1}{n}(\sum x)^2$$

$$S_{yy} = \sum y^2 - \frac{1}{n}(\sum y)^2$$

$$S_{xy} = \sum xy - \frac{1}{n}(\sum x)(\sum y)$$

$$s_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{S_{yy} - b_1 S_{xy}}{n-2}}$$

$$s_{b_1} = \frac{s_e}{\sqrt{S_{xx}}}$$

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$SST = S_{yy}$$

$$SSE = S_{yy} - b_1 S_{xy}$$

$$SSR = b_1 S_{xy}$$

ANALYSIS OF VARIANCE

Motivation:- Suppose we are asked to compare the cleansing action on the basis of following whiteness reading of three Detergents on 15 dinner plates. The data is listed below.

Detergent X	77	81	71	76	80
Detergent Y	72	58	74	66	70
Detergent Z	76	85	82	80	77

`Other factors at the time of washing the dinner plates :-`

- how dirty the plates were?
- washing time?
- water temperature and hardness while cleaning.
- what kind of instrument used for washing?

If we calculate the means of whiteness reading we will have 77 for detergent X, 68 for detergent Y, 80 for detergent Z.
`Remember that a significant test may show that the differences between sample means are too large to be attributed to chance,`

`but cannot say why the differences occurred?`
Hence we should introduce a `Controlled Experiment procedure` , where every above factors remain the same during the washing process. That means same washing time, same dirty ness, use water of exactly the same temperature and hardness , inspecting the instrument after each use and so many other factors... .
Isn't the above Controlled Experiment tedious?

Ans:-
So what we do to get rid of this Controlled Factors? We can conduct an experiment in which none of these factors is controlled , but in which we protect ourselves against their effect by **RANDOMIZATION**. That is we design, or plan, the experiment in such a way that the variation caused by these factors can be combined under the general heading of "Chance".

Mathematically what we do?

Def:- An **analysis of variance** expresses a measure of the total variation in a set of data as a sum of terms, each attributed to a specific source, or cause, of variation.

The Assumptions of Analysis of Variance

1. Treatment effects are additive.
2. Experimental errors - are random.
3. Experimental errors are independently distributed.
4. Experimental errors follow a normal distribution.
5. Experimental errors have mean zero and constant variance.

Source					
Treatment 1	x_{11}	x_{12}	x_{13}	...	x_{1n}
Treatment 2	x_{21}	x_{22}	x_{23}	...	x_{2n}
...	...	...	...	...	...
Treatment K	x_{K1}	x_{K2}	x_{K3}	...	x_{Kn}

notations:-

$$Sample\ in\ i^{th}\ row\ and\ j^{th}\ coloumn = x_{ij}$$

$$Treatment\ K\ variance = s_i^2$$

$$Grand\ Mean = \bar{x} = \frac{x_{11} + x_{12} + ... + x_{Kn}}{Kn}$$

$$Mean\ Treatment\ 1 = \bar{x}_1 = \frac{x_{11} + x_{12} + ... + x_{1n}}{n}$$

$$Mean\ Treatment\ 2 = \bar{x}_2 = \frac{x_{21} + x_{22} + ... + x_{2n}}{n}$$

.

.

$$Mean\ Treatment\ K = \bar{x}_k = \frac{x_{k1} + x_{k2} + ... + x_{kn}}{n}$$

$$\text{SST} = \text{SS(Tr)} + \text{SSE}$$

Equivalently we can re-write the above equations;

$$\text{SST} = \text{SSB} + \text{SSW}$$

Mathematically

$$SST = \sum\sum(x_{ij} - \bar{x})^2$$
$$SSB = n\sum(\bar{x}_i - \bar{x})^2$$
$$SSW = \sum\sum(x_{ij} - \bar{x}_i)^2 = (n-1)(s_1^2 + s_2^2 + ... + s_K^2)$$

Test $H_0 : \mu_1 = \mu_2 = \mu_3 = ... = \mu_K$

AnOVA TABLE

Source	DF	Sum of Squares	Mean Square	F
Treatments	K-1	SSB	$MSB = \frac{SSB}{K-1}$	$F = \frac{MSB}{MSW}$
Error	K(n-1)	SSW	$MSW = \frac{SSW}{K(n-1)}$	
Total	nK-1	SST		

note:- This AnOVA works when each Treatment have same number of observations. The Formulas for SSB, and SSW will slightly change when we have different observations in each treatments.

COMPUTInG FORMULAS FOR SUM OF SQUARES (Equal Sample Size)

$$T = x_{11} + x_{12} + ... + x_{Kn}$$

$$T_i = x_{i1} + x_{i2} + ... + x_{in}$$

$$\boxed{\begin{array}{c} SST = \sum\sum x_{ij}^2 - \frac{1}{Kn}T^2 \\ SSB = \frac{1}{n}\sum T_i^2 - \frac{1}{Kn}T^2 \\ SSW = SST - SSB \end{array}}$$

COMPUTING FORMULAS FOR SUM OF SQUARES (UnEqual Sample Size)

Source					
Treatment 1	x_{11}	x_{12}	x_{13}	...	x_{1n_1}
Treatment 2	x_{21}	x_{22}	x_{23}	...	x_{2n_2}
...	...	...	...	...	...
Treatment K	x_{K1}	x_{K2}	x_{K3}	...	x_{Kn_K}

$$T = x_{11} + x_{12} + ... + x_{Kn_K}$$

$$T_i = x_{i1} + x_{i2} + ... + x_{in_1}$$

$$N = n_1 + n_2 + ... + n_K$$

$$\boxed{\begin{array}{c} SST = \Sigma\,\Sigma\, x_{ij}^2 - \frac{1}{N}T^2 \\ SSB = \Sigma\, \frac{T_i^2}{n_i} - \frac{1}{N}T^2 \\ SSW = SST - SSB \end{array}}$$

WILCOXON SIGNED RANK TEST FOR DEPENDENT SAMPLES (Small ($n \leq 30$))

Null Hypothesis Alternate Hypothesis	$H_0 : M_D = 0$ $H_0 : M_D \neq 0$	$H_0 : M_D = 0$ $H_0 : M_D > 0$	$H_0 : M_D = 0$ $H_0 : M_D < 0$
Test Statistic	Smaller(T_-, T_+)	T_-	T_+
Rejection Region	$T \leq T_0$	$T_- \leq T_0$	$T_+ \leq T_0$

1. T_- = Sum of Ranks of all negative differences.
2. T_+ = Sum of Ranks of all positive differences.
3. T_0 = Critical Value.

WILCOXON SIGNED RANK TEST FOR DEPENDENT SAMPLES (Large($n > 30$))

Null Hypothesis Alternate Hypothesis	$H_0 : M_D = 0$ $H_0 : M_D \neq 0$	$H_0 : M_D = 0$ $H_0 : M_D > 0$	$H_0 : M_D = 0$ $H_0 : M_D < 0$
Test Statistic	z_{cal}	z_{cal}	z_{cal}
Rejection Region	$\lvert z \rvert > z_{\frac{\alpha}{2}}$	$z < -z_\alpha$	$z > z_\alpha$

Where T_+ is sum of Ranks of all positive differences.

$$z = \frac{T_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Problem WILCOXON SIGNED RANK TEST FOR DEPENDENT SAMPLES (Small ($n \leq 30$))

A	B	A-B	$\lvert A-B \rvert$	Dummy Ranking	Ranking
7	9				
4	5				
8	8				
9	8				
3	6				
6	10				
8	9				
10	8				
9	4				
5	9				

COMPARING TWO POPULATION INDEPENDENT SAMPLES
or
WILCOXON RANK SUM TEST
(SMALL($n_1 \leq 20$, $n_2 \leq 20$))

`Location 1:-`	0.37	0.70	0.75	0.30	0.45	0.16	0.62	0.73	0.33
`Location 2:-`	0.86	0.55	0.80	0.42	0.97	0.84	0.24	0.51	x

Sorting the numbers in ascending order and Ranking

Data	Location	Rank
0.16	1	**1**
0.24	2	2
0.30	1	**3**
0.33	1	**4**
0.37	1	**5**
0.42	2	6
0.45	1	**7**
0.51	2	8
0.55	2	9
0.62	1	**10**
0.69	2	11
0.70	1	**12**
0.73	1	**13**
0.75	1	**14**
0.80	2	15
0.84	2	16
0.86	2	17
0.92	2	18
0.97	2	19

$$T_1 = 1 + 3 + 4 + 5 + 7 + 10 + 12 + 13 + 14 = 69$$

$$T_2 = 2{+}6{+}8{+}9{+}11{+}15{+}16{+}17{+}18{+}19 = 121 = \frac{19 * 20}{2} - 69$$

Null Hypothesis	$H_0 : M_1 = M_2$	$H_0 : M_1 = M_2$	$H_0 : M_1 = M_2$
Alternate Hypothesis	$H_0 : M_1 \neq M_2$	$H_0 : M_1 > M_2$	$H_0 : M_1 < M_2$
N0.	Test	.	Statistic
$n_1 < n_2$	T_1	T_1	T_1
$n_1 > n_2$	T_2	T_2	T_2
$n_1 = n_2$	Either one	Either one	Either one

Rejection Region

`Problem:`

The following are the burning times of random samples of two kinds of emergency flares:

Brand 1	Brand 2
17.2	13.6
18.1	19.1
19.3	11.8
21.1	14.6
14.4	14.3
13.7	22.5
18.8	12.3
15.2	13.5
20.3	10.9
17.5	14.8

Use the Wilcoxon Rank Sum Test at 0.05 l.o.s to test whether it is reasonable to say that on the average brand 1 flares are better than Brand 2 flares.

Summary of Non Parametric Tests for Final Exam

- **ONE SAMPLE SIGN TEST FOR A POPULATION MEDIAN (Small ($n \leq 30$))**
  We are testing , Null Hypothesis $H_0 : M = 0$, M is the population Median.

- **WILCOXON SIGNED RANK TEST FOR DEPENDENT SAMPLES (Small ($n \leq 30$))**
  We are testing , Null Hypothesis $H_0 : M_D = 0$, M_D is the population Median difference.

- **WILCOXON SIGNED RANK TEST FOR DEPENDENT SAMPLES (Large ($n > 30$))**
  We are testing , Null Hypothesis $H_0 : M_D = 0$

- **COMPARING TWO POPULATION INDEPENDENT SAMPLES or WILCOXON RANK SUM TEST (SMALL($n_1 \leq 20$, $n_2 \leq 20$))**
  We are testing , Null Hypothesis $H_0 : M_1 = M_2$
  M_1 and M_2 are Population median for sample 1 and sample 2 respectively.

- **KRUSKAL-WALLIS H-TEST FOR COMPLETELY RANDOMIZED DESIGN (ONE WAY ANOVA)**
  We are testing , Null Hypothesis $H_0 : M_1 = M_2 = M_3 = ... = M_k$

- **FRIEDMAN F_r-TEST FOR COMPLETELY RANDOMIZED BLOCK DESIGN (TWO WAY ANOVA)**
  We are testing , Null Hypothesis H_0 : Effects of Each Treatment = 0, Effect of Each Block = 0

- **SPEARMAN'S RANK CORRELATION COEFFICIENT**
  We are testing , Null Hypothesis $H_0 : \rho = 0$, where ρ is the population correlation between x and y.