Quantitative Methods 1 - Summary - QM - NOTES

Quantitative Methods 1 (University of Melbourne)

Quantitative Methods 1

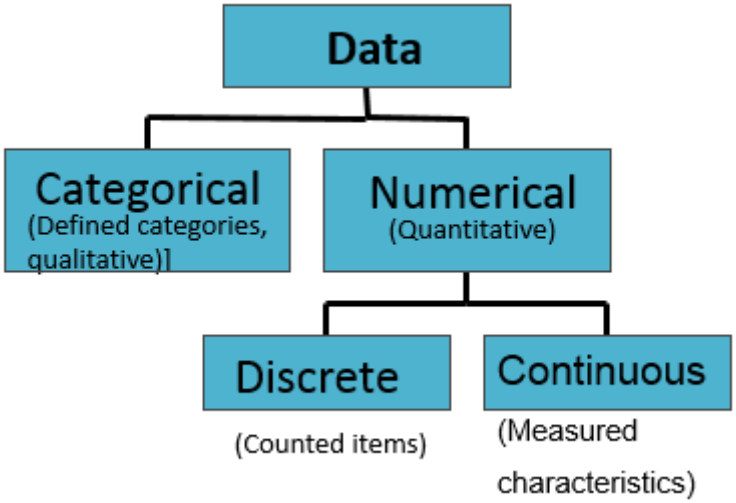
Key definitions

A **population** consists of all the members of a group about which you want to draw a conclusion.
A **sample** is the portion of the population selected for analysis.
A **parameter** is a numerical measure that describes a characteristic of a population.
A **statistic** is a numerical measure that describes a characteristic of a sample.

Descriptive statistics is collecting, summarising, and presenting data.
Inferential statistics is drawing conclusions about a population based on sample data, i.e. estimating a parameter based on a statistic.

Inferential Statistics
Estimation: Estimate the population mean income (parameter) using the sample mean income (statistic).
Hypothesis testing: Test the claim that the population mean income is $80,000.

Defining Data



Categorical (Qualitative)
- Simply classifies data into categories e.g. marital status, hair colour, gender.

Numerical (Discrete)
- Counted items (finite number of items) e.g. number of children, number of people who have type O blood.

Numerical (Continuous)
- Measured characteristics (infinite number of items) e.g. weight, height.

Graphical Techniques

What is a frequency distribution?
A frequency distribution is a summary table in which data are arranged in numerically ordered classes or intervals.
The number of observations in each ordered class or interval becomes the corresponding frequency of that class or interval.

Why use a frequency distribution?
It is a way to summarise numerical data.

It condenses the raw data (i.e. large datasets) into a more useful form.
It allows for a quick visual interpretation of the data and first inspection of the shape of the data.

Class Intervals and Class Boundaries
Each data value belongs to one, and only one, class.
Each class grouping has the same width.
Determine the width of each interval by:

$$\text{Width of Interval} \cong \frac{\text{Range}}{\text{Number of desired class groupings}}$$

General guidelines:

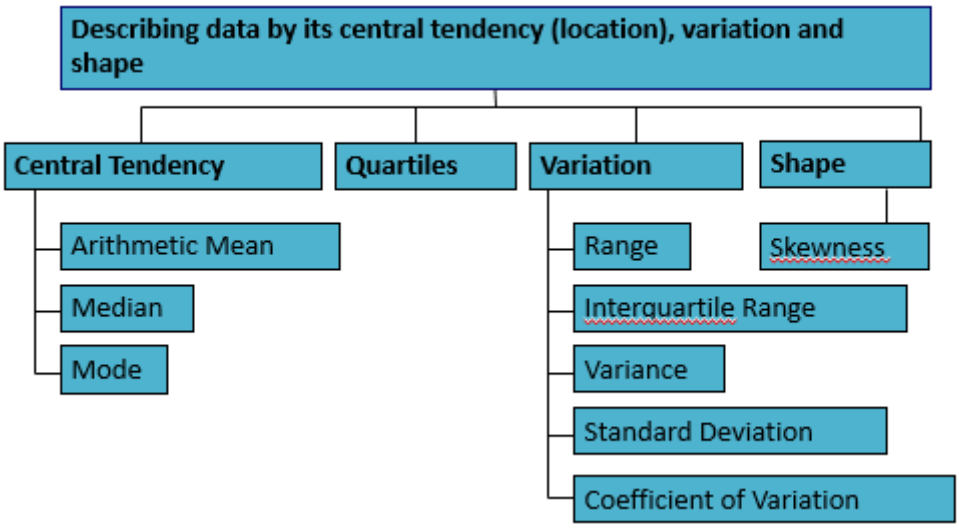
- Usually at least 5 but no more than 15 groupings
- Class boundaries must be mutually exclusive
- Classes must be collectively exhaustive
- Round up the interval width to get desirable endpoints

Graphing Numerical Data: The Histogram
A graph of the data in a frequency distribution is called a **histogram.**
The **class boundaries** (or **class midpoints**) are shown on the horizontal axis.
The vertical axis is either **frequency, relative frequency** or **percentage.**
Bars of the appropriate heights are used to represent the frequencies (number of observations) within each class or the relative frequencies (percentage) of that class.

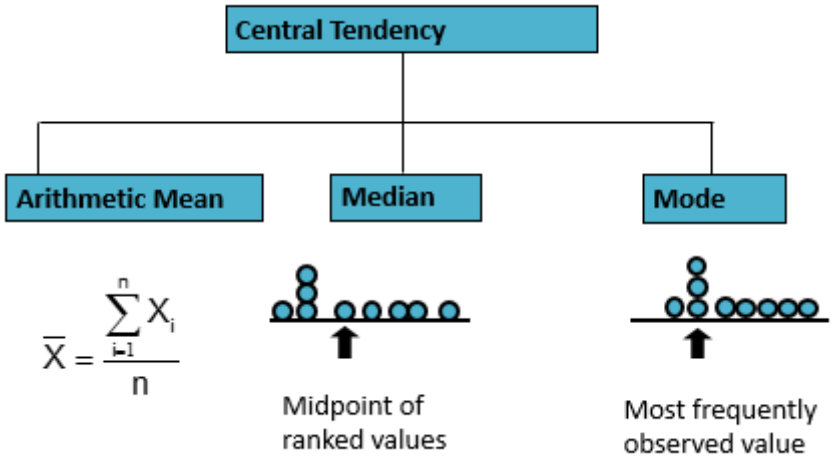
Scatter Diagrams
Scatter diagrams are used to examine possible relationships between two numerical variables.
In a scatter diagram one variable is measured on the vertical axis (Y) and the other variable is measured on the horizontal axis (X).

Numerical Descriptive Measures

Describing Data
Graphing data is useful to understand how data can be used to organize, present and summarized.
It is also useful to use numerical methods to summarize data in ways that cannot be easily visualized, but also in ways that make the data easier to compare.



Measures of Central Tendency



Arithmetic Mean in which a sample of size n, the same mean denoted, $\bar{X}$, is calculated:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Where Σ means to sum or add up.
This formula is affected by extreme values.

The median is an ordered array, the median is the 'middle' number in which 50% of the data is above and 50% of the data is below.
Its main advantage over the arithmetic mean is that it is not affected by extreme values.

To find the location of the median, it is found by: $L = \frac{n+1}{2}$

This formula does not give the **value** of the median but the **position** of the median.
Rule 1: if the number of values in the data set is odd, the median is the middle ranked value.
Rule 2: if the number of values in the data set is even, the median is the mean (average) of the two middle ranked values.

The **mode** is a measure of central tendency in which is the value that occurs most often (most frequent) in the data set. It is not affected by extreme values and unlike the mean and median, there may be no unique (single) mode for a given data set.
It is used for either numerical or categorical (nominal) data.

Quartiles
Quartiles split the ranked data into four segments with an equal number of values per segment.

The first quartile, Q_1, is the value for which 25% of the observations are smaller and 75% are larger.
The second quartile, Q_2, is the same as the median for which 50% of the observations are smaller and 50% are larger.
Only 25% of the observations are greater than the third quartile Q_3.

Similar to the median, we find a quartile by determining the value in the appropriate **position** in the **ranked** data: First quartile position: $L_{Q_1}=\frac{n+1}{4}$

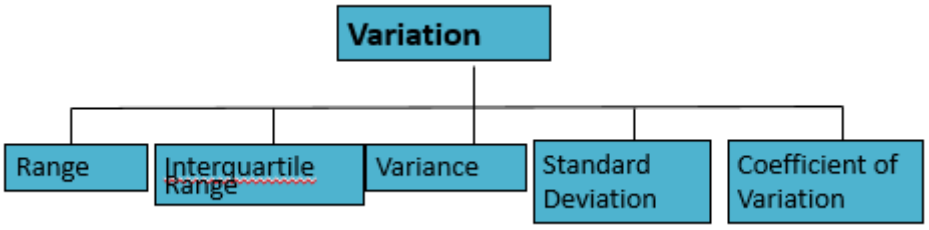
Second quartile position: $L_{Q_2}=\frac{n+1}{2}$ (same as the median)

Third quartile position: $L_{Q_3}=\frac{3(n+1)}{4}$

Where *n* is the number of observed values (sample size).

Rule 1: If the value is an integer, Q is that ranked value.
Rule 2: If the value is a fractional half i.e. 2.5, 3.5, etc then the quartile is the mean of the corresponding ranked values.
Rule 3: If the value is neither and integer nor a fractional half i.e. 2.75, 3.25, etc, round the number to the nearest ranked value and Q is that ranked value.

Measures of Variation



Measures of variation give information on the **spread** or **variability** of the data values.

The range is the simplest measure of variation. It is the difference between the largest and the smallest values in a set of data.

The **interquartile range (IQR)** is like the median and Q_1 and Q_3 in which the IQR is a resistant summary measure (resistant to the presence of extreme values).
We can eliminate outlier problems by using the **interquartile range** as high- and low-valued observations are removed by calculations.

$$IQR = Q_3 - Q_1$$

The **sample variance s^2** measures the average scatter around the mean.

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

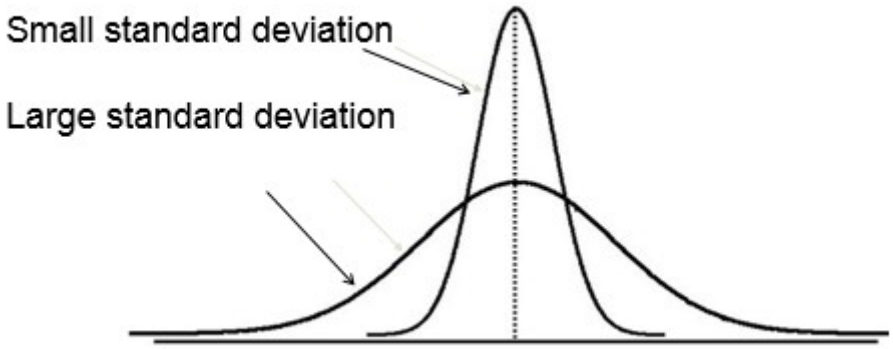
Where

$\bar{X}$ = mean
n = sample size
X_i = i^{th} value of the variable X

The **sample standard deviation**, **s**, is the most commonly used measure of variation and has the same units as the original data. This shows the variation about the mean.

$$s = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

The graph below shows the types of variation.



Variance and Standard Deviation

Advantages:

- Each value in the data set is used in the calculation
- Values far from the mean are given extra weight as deviations from the mean are squared

Disadvantages:

- Sensitive to extreme values (outliers)
- Measures of absolute variation, not relative variation i.e. we cannot compare between data sets with different units or widely different means.

The Z Score

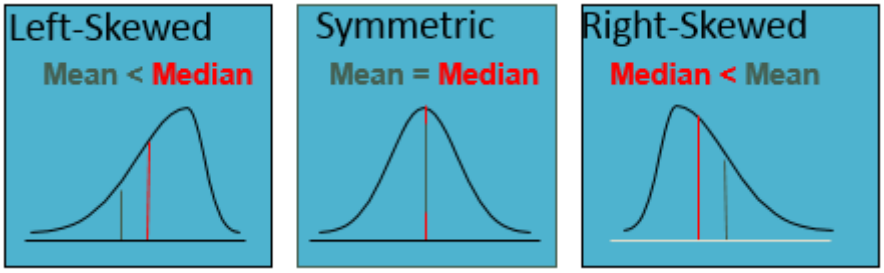
A z-score is a measure of relative standing that takes into consideration both mean & standard deviation. For each observation in the dataset, we can estimate a z-score on the basis of which we can identify whether an observation is an outlier. The difference between a given observation and the mean, divided by the standard deviation.

$$Z = \frac{X - \bar{X}}{S}$$

Shape of a Distribution

This describes how data are distributed, the measures of shape.

- Symmetric or skewed



Numerical Measures for a Population

The population summary measures are called **parameters.** The population mean is the sum of the values in the population divided by the population size, N.

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = \frac{X_1 + X_2 + \cdots + X_N}{N}$$

Population variance is the average of the squared deviations of values from the mean.

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Population Standard Deviation shows the variation about the mean and is the square root of the population variance. IT has the same units as the original data.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}}$$

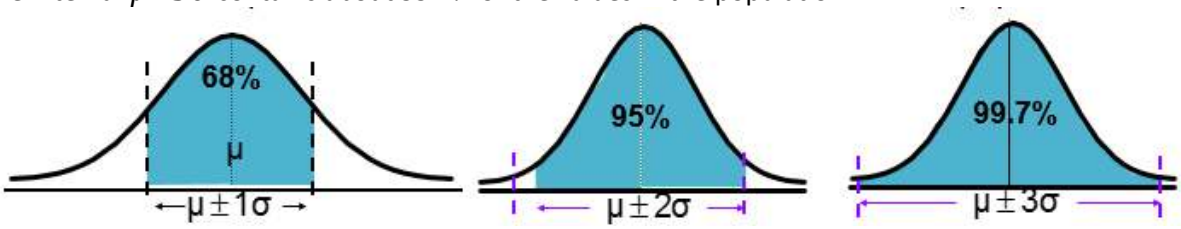
The Empirical Rule

If the data distribution is approximately bell-shaped, then

The interval $\mu \pm 1\sigma$ contains about 68% of the values in the population.

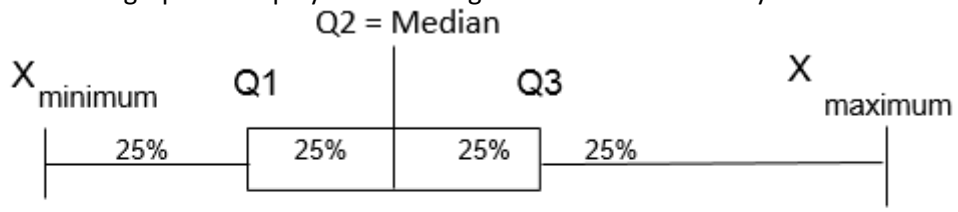
The interval $\mu \pm 2\sigma$ contains about 95% of the values in the population

The interval $\mu \pm 3\sigma$ contains about 99.7% of the values in the population



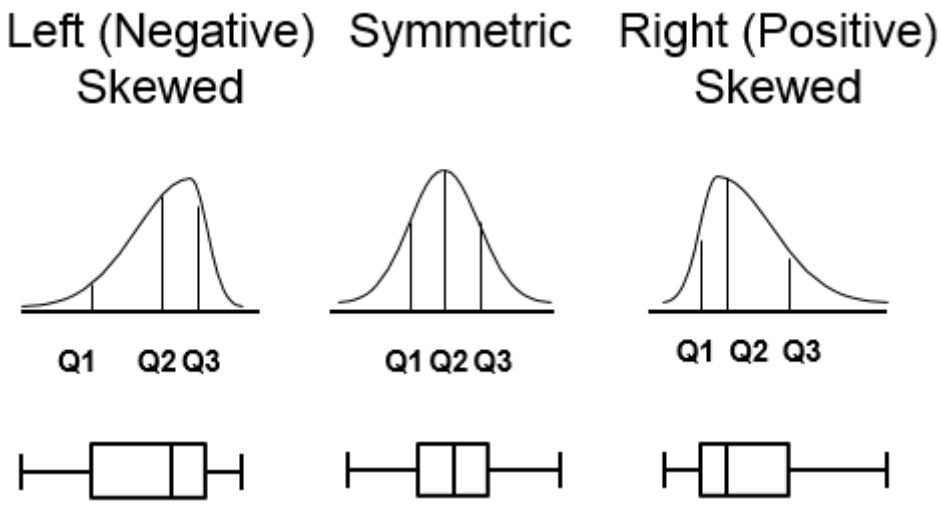
Exploratory Data Analysis

Box-and-Whisker Plot: A graphical display of data using the 5 number summary:



Minimum($X_{smallest}$) -- Q1 -- Median -- Q3 -- Maximum ($X_{largest}$)

The box-and-whisker plots correspond to the distributions as follows:



The **sample covariance** measures the strength of the linear relationship between **two numerical variables.**

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

A positive covariance means that there is a positive linear relationship and a negative covariance means there is a negative linear relationship.
By using this formula, it is only concerned with the direction of the relationship and no causal effect is implied. It is not a measure of relative strength and is affected by units of measurement.

Correlation measures the **relative strength** of the linear relationship between two variables.

$$r=\frac{\text{cov}\left(X,\text{Y}\right)}{S_X S_Y}$$, where $$S_X=\sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$$ and $$S_Y=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-\overline{Y})^2}{n-1}}$$.

A feature of correlation coefficient r is that it ranges between -1 and 1 where:

- The closer to -1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker the linear relationship

Basic Probability Theory

Probability is a numerical value that represents the chance, likelihood, possibility that an event will occur (always between 0 and 1).
An **event** is each possible outcome of a variable.

Events

Simple event (denoted A)

- An outcome from a sample space with one characteristic

Complement of an event A (denoted A')

- All outcomes that are not part of event A.

Joint event (denoted A∩B, pronounced A intersect B)

- Involves two or more characteristics simultaneously

Mutually exclusive events

- Events that cannot occur together

Collectively exhaustive events

- One of the events must occur. The set of events covers the entire sample space.

Probability

The probability of any event must be between 0 and 1, inclusively.
$0 \le P(A) \le 1$, for any event A
The sum of the probabilities of all mutually exclusive and collectively exhaustive events is 1.
P(A) + P(B) = 1, if A and B are mutually exclusive and collectively exhaustive.

Computing Joint and Marginal Probabilities

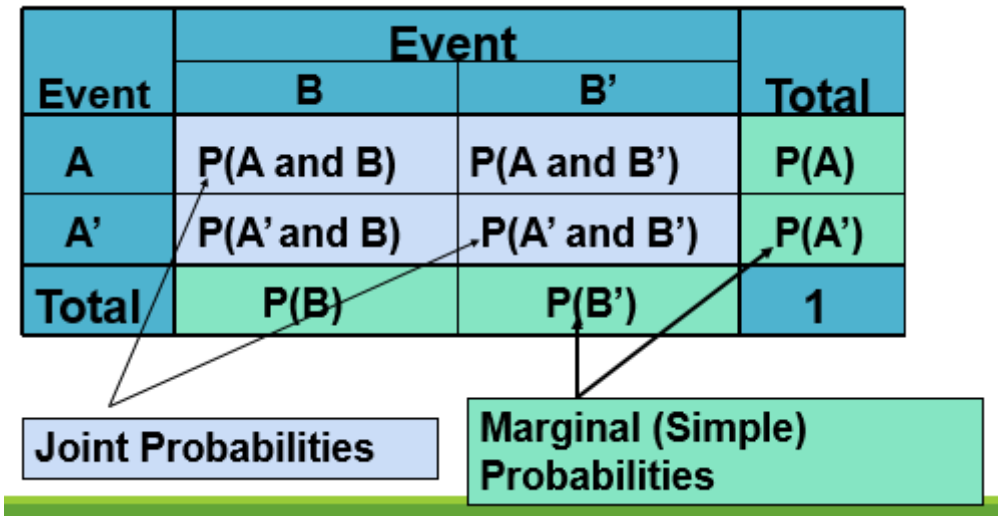
The probability of a joint event, A and B:

$$P(A \text{ and } B)=\frac{\textit{number of outcomes satisfying A and B}}{\textit{total number of elementary outcomes}}$$

Computing a marginal (or simple) probability:

$$P(A)=P(A \text{ and } B_1)+P(A \text{ and } B_2)$$

Where B_1 and B_2 are mutually exclusive and collectively exhaustive events.



General addition rule

P(A or B) = P(A) + P(B) – P(A and B)

If A and B are mutually exclusive, then

P(A and B) = 0, so the rule can be simplified

P(A or B) = P(A) + P(B) **WHERE A AND B ARE MUTUALLY EXCLUSIVE.**

Computing Conditional Probabilities

A **conditional probability** is the probability of one event, given that another event has occurred:

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Where P(A and B) = joint probability of A and B
P(A) = marginal probability of A
P(B) = marginal probability of B

Statistical Independence

Two events are **independent** if, and only if:
P(A|B) = P(A)
Events A and B are independent when the probability of one event is not affected by the other event.

Multiplication Rules

Multiplication rule for two events A and B:
P(A and B) = P(A|B) x P(B)

Note: If A and B are **independent** then
P(A|B) = P(A) and the multiplication rule simplifies to P(A and B) = P(A) x P(B)

Marginal Probability

Marginal probability for event A:

$P(A) = P(A \text{ and } B_1) + P(A + B_2) = P(A|B_1) + P(A|B_2) \times P(B_2)$
Where B_1 and B_2 are mutually exclusive and collectively exhaustive events.

Discrete Probability Theory

A **random variable** represents a possible numerical value from an uncertain event. **Discrete** random variables can only assume a countable number of values.

What is a Probability Distribution?

A **probability distribution** provides the possible values of the random variable and their corresponding probabilities. A probability distribution can be in the form of a table, graph, or mathematical formula.

Requirements of a discrete probability distribution:

$$\sum P(X=x)=1$$
$$0 \le P(X=x) \le 1$$

Discrete Random Variable Summary Measures

Expected value (or mean) of a discrete random variable (weighted average).

$$\mu = \mathrm{E}(X) = \sum_{i=1}^{N} X_i P(X_i)$$

Variance of a discrete random variable – definition formula

$$\sigma^2 = \sum_{i=1}^{N} [X_i - E(X)]^2 P(X_i)$$

Variance of a discrete random variable – alternative calculation formula

$$\sigma^2 = \sum_{i=1}^{N} X_{i^2} P(X_i) - E(X)^2$$

Where E[x] = expected value of the discrete random variable x,
X_i = the i^{th} outcome of the discrete random variable x,
$P(X_i)$ = probability of the i^{th} occurrence of x

The Covariance

The **covariance** measures the direction of a linear relationship between two variables.

Definition formula for covariance

$$\sigma_{XY} = \sum_{i=1}^{N} [X_i - E(X)][(Y_i - E(Y)]\ P(X_i Y_i)$$

Calculation formula for covariance

$$\sigma_{XY} = \sum_{i=1}^{N} X_i Y_i P(X_i Y_i) - E(X)E(Y)$$

Where X_iY_i is the i^{th} outcome of the discrete random variables X and Y respectively,
$P(X_iY_i)$ = probability of the i^{th} occurrence of X and Y

The Sum of Two Random Variables

Expected value of the sum of two random variables

$$E(X+Y) = E(X) + E(Y)$$

Variance of the sum of two random variables

$$\mathrm{Var}(X+Y) = \sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\sigma_{XY}$$

Standard deviation of the sum of two random variables

$$\sigma_{X+Y}=\sqrt{\sigma^2_{X+Y}}$$

Combinations

The number of **combinations** of selecting X objects out of n objects is

$$\left(\frac{n}{X}\right)={}_nC_x=\frac{\mathrm{n!}}{\mathrm{X!}(n-X)!}$$

Where:

n! = n(n-1)(n-2)…(2)(1)

x! = x(x-1)(x-2)…(2)(1)

0! = 1 (by definition)

The Binomial Distribution Formula

$$P(X)=\frac{n!}{X!(n-X)!}p^X(1-p)^{n-X}$$

Where:

P(X) = probability of **X** successes in **n** trials, with probability of success **p** on each trial

X = number of 'successes' in sample

n = sample size (number of trials or observations)

p = probability of 'success'

1 – p = probability of failure

Characteristics of the Binomial Distribution

Mean

$$\mu=E(X)=\mathrm{np}$$

Variance and standard deviation

$$\sigma^2=\mathrm{np}(1\text{-}\mathrm{p})$$

$$\sigma=\sqrt{\mathrm{np}(1\text{-}\mathrm{p})}$$

Where:

n = sample size

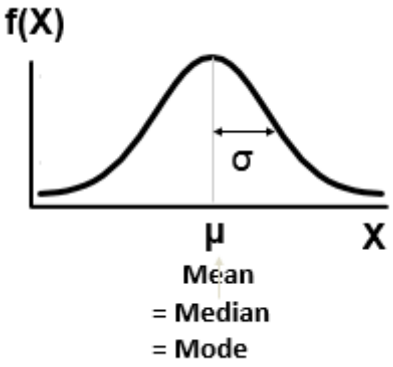
p = probability of success

(1 – p) = probability of failure

Continuous Probability Distributions

A **continuous random variable** is a variable that can assume any value on a continuum (can assume an infinite number of values).

The Normal Distribution

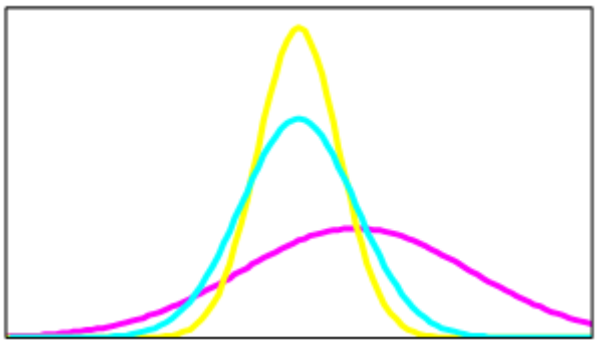


Features of a normal distribution:

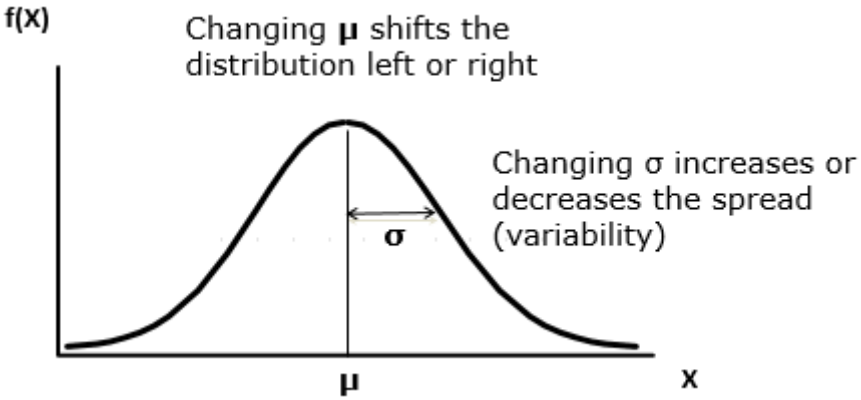
- Bell-shaped
- Symmetrical
- Mean, median and mode are equal
- Central location is determined by the mean, μ
- Spread is determined by the standard deviation, σ
- The random variable X has an infinite theoretical range: $+\infty$ to $-\infty$

Many Normal Distributions

There are many distributions that we can create by varying the parameters **μ** and **σ**, we obtain different normal distributions.



This is because:



<u>The Normal Probability Density Function</u>

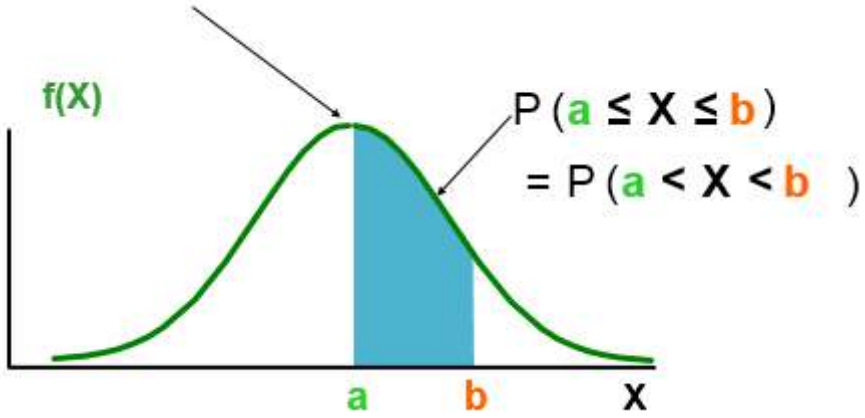
The formula for the normal probability density function is:

$$f(X)=\frac{1}{\sqrt{2\pi}\sigma}e^{-(1/2)[(X-\mu)/\sigma]^2}$$

Where:
e = the mathematical constant approximated by 2.71828
Π = the mathematical constant approximated by 3.14159
μ = the population mean
σ = the population standard deviation
X = any value of the continuous variable

Finding Normal Probabilities

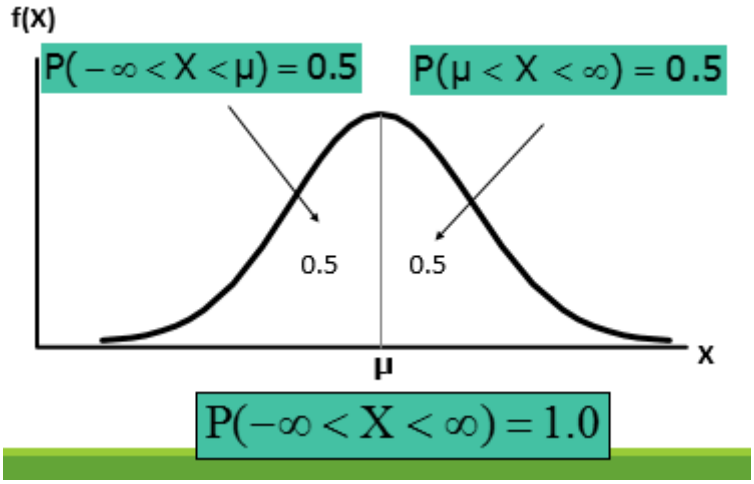
Probability is measured by the area under the curve.



Note that the probability of any individual value is zero since the X axis has an infinite theoretical range: + ∞ to - ∞

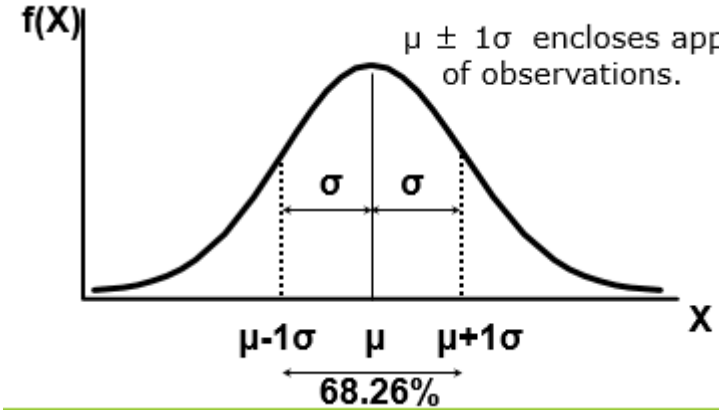
Probability as Area Under the Curve

The total area under the curve is 1, and the curve is symmetric so half is above the mean and half is below.

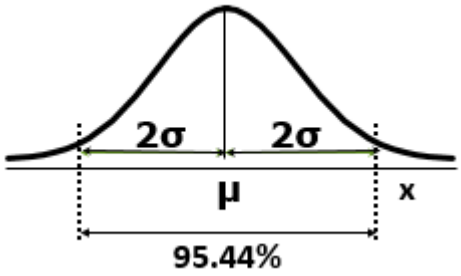


Empirical Rules

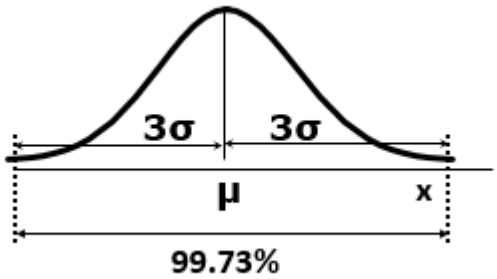
What can we say about the distribution of values around the mean?
There are some general rules:



μ ± 2σ covers approx 95.44% of observations.



μ ± 3σ covers approx 99.73% of observations.



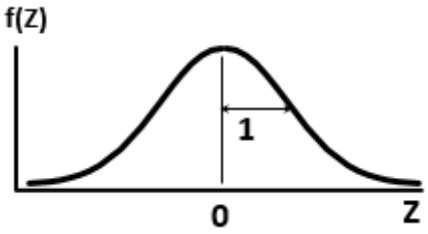
Translation to the Standardised Normal Distribution

Any normal distribution (with any mean and standard deviation combination) can be transformed into the standardised normal distribution (z).

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

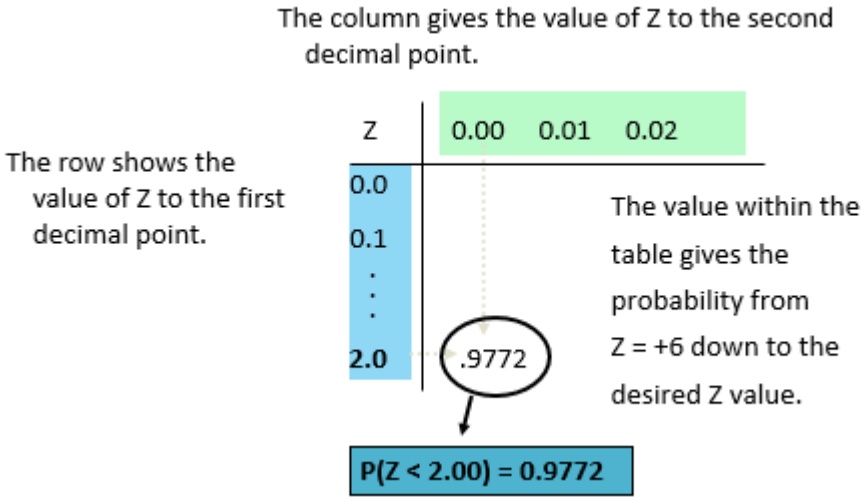
The Standardised Normal Distribution

Is also known as the Z distribution which has a mean of 0 and the standard deviation of 1.



Values above the mean have positive z-values and values blow the mean have negative z-values.

The Standardised Normal Table



General Procedure for Finding Probabilities

To find P(a < x < b) when X is distributed normally:

1. Draw the normal curve for the problem in terms of X.
2. Translate X-values to Z-values and put Z values on your diagram.
3. Use the Standardised Normal Table.

Finding the X Value for a Known Probability

1. Draw a normal curve placing all known values on it, such as mean of X and Z.
2. Shade in area of interest and find cumulative probability.
3. Find the Z value for the known probability.
4. Convert to X units using the formula.

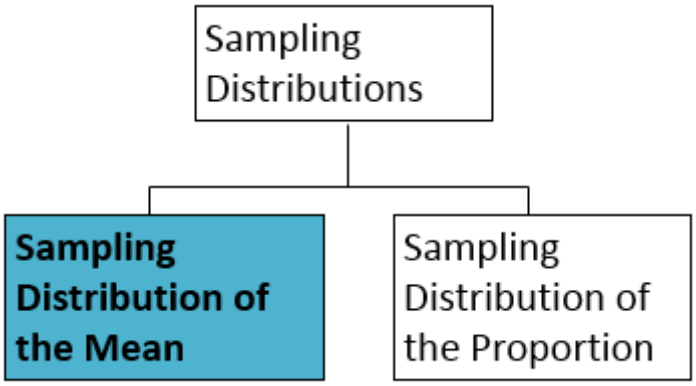
$$X = \mu + Z\sigma$$

This is the Z formula rearranged in terms of X.

Sampling Distributions and Data Sampling

Recall that a sample is a set of objects or people selected out of a population. Presumably, it is taken in a process that is representative of the population that you care about. Yet, not all samples are equal.

A **sampling distribution** is a distribution of all of the possible values of a statistic for a given sample size selected from a population.

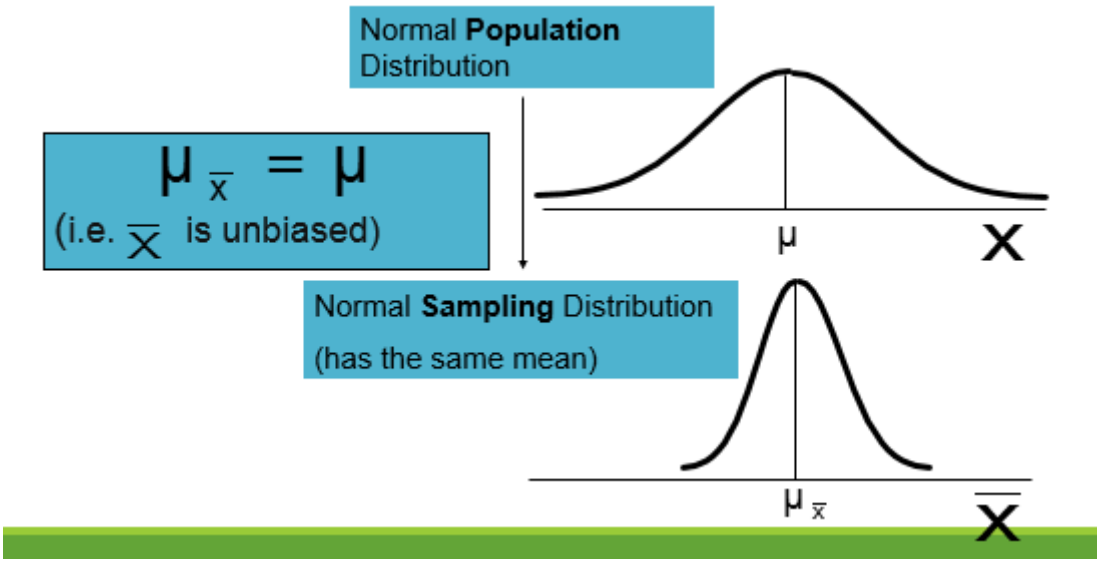


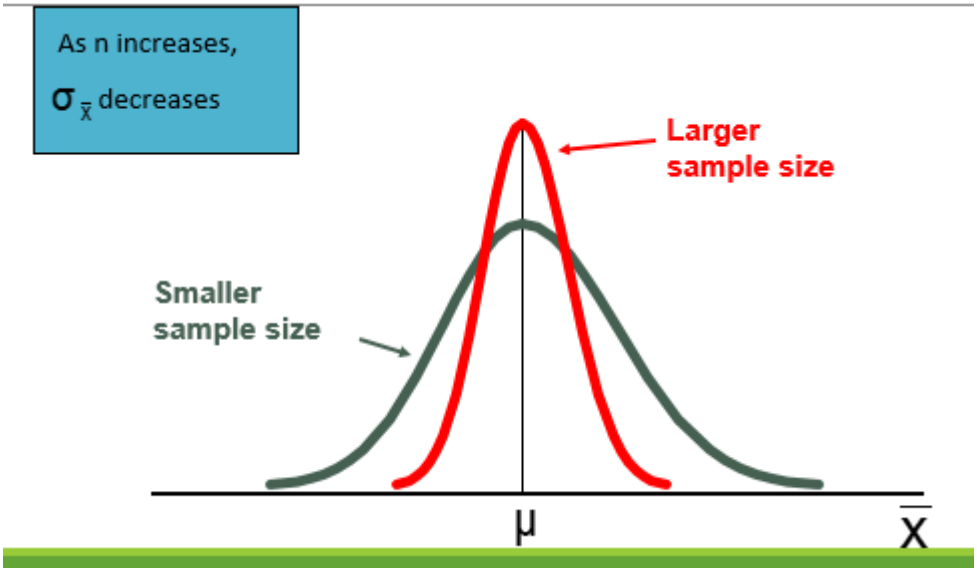
If the Population is Normal

If a population is **normal** with mean μ and standard deviation σ, the sampling distribution of x̄ is also **normally distributed** with:

$\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

Sampling Distribution Properties



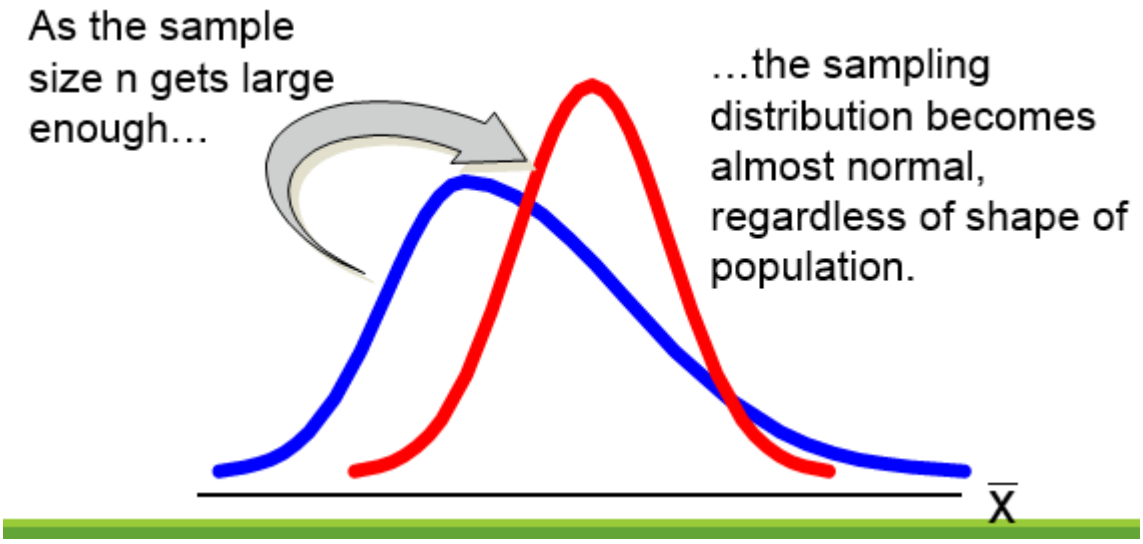


If the Population is NOT Normal

We can apply the **Central Limit Theorem**, which states that regardless of the shape of individual values in the population distribution, **as long as the sample size is large enough** (generally n ≥ 30) the sampling distribution of $\bar{x}$ will be approximately normally distributed with:

$\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

The Central Limit Theorem



Z Formula for Sampling Distribution

If the population is normal OR the Central Limit Theorem is applicable, we can use the normal distribution and the Z table to find probabilities for the sample mean. The relevant formula is:

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Where:

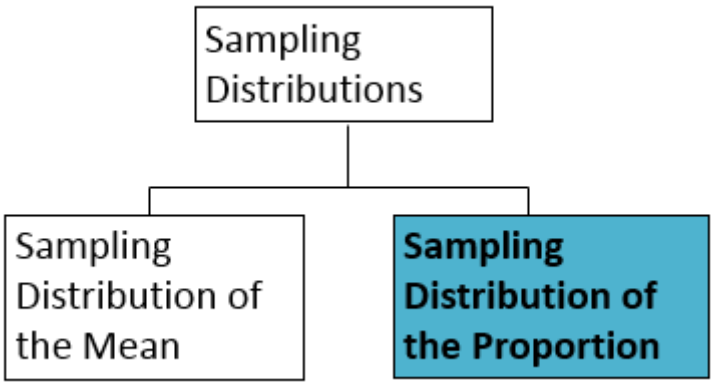
$\acute{X}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

Sampling Distribution of the Proportion



π is the proportion of items in the **population** with a characteristic of interest.

p is the **sample proportion** and provides an *estimate* of π

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

$$= \frac{\text{number of items in the sample having the characteristic of interest}}{\text{sample size}}$$

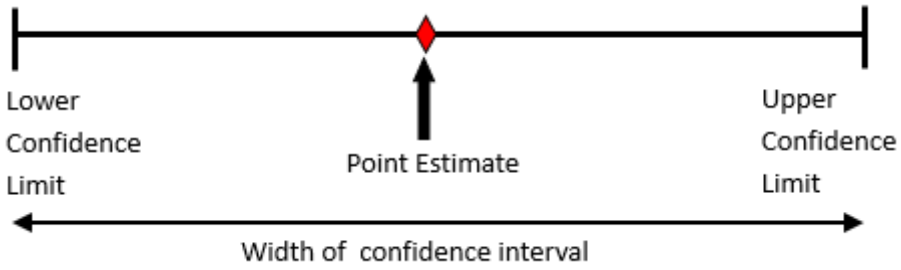
Selecting all possible samples of a certain size, the distribution of all possible sample proportions is the sampling distribution of the proportion.

Confidence Intervals

Point and Interval Estimates

A **point estimate** is the value of a single sample statistic.

A **confidence interval** provides a range of values constructed around the point estimate.



Point Estimates

We can estimate a Population Parameter…		with a Sample Statistic (Point Estimate).
Mean	μ	$\bar{X}$
Proportion	π	p

Confidence Interval Estimate

An interval gives a **range** of values:

- Takes into consideration variation in sample statistics from sample to sample
- Based on observations from 1 sample
- Gives information about closeness to unknown population parameters
- Stated in terms of level of confidence
- Can never be 100% confident

Confidence Interval

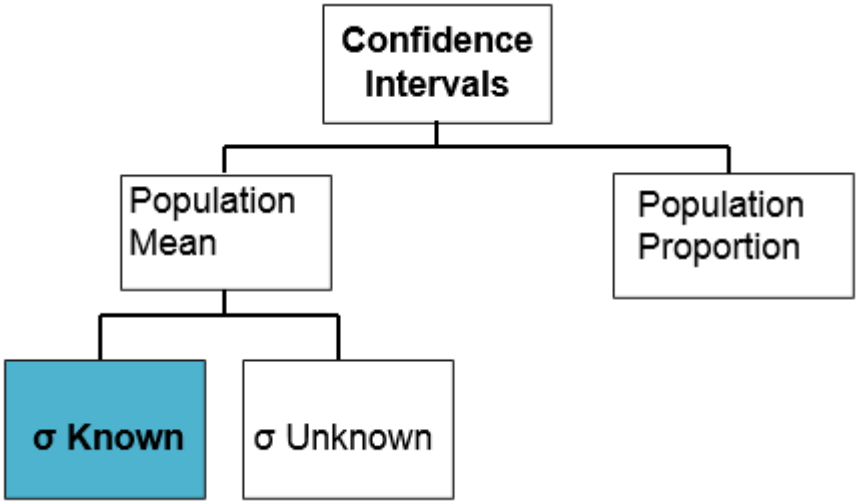
The general formula for all confidence intervals is:

Point Estimate ± (Critical Value)*(Standard Error)

Represents confidence for which the interval will contain the unknown population parameter.

Common confidence levels = 90%, 95% or 99%:

- Also written $(1 - \alpha) = 0.90, 0.95$ or 0.99
  A relative frequency interpretation:
- In the long run, 90%, 95% or 99% of all the confidence intervals that can be constructed (in repeated samples) will contain the unknown true parameter.
  A specific interval will either contain or will not contain the true parameter.
- No probability involved in a specific interval.



Confidence Interval for μ (σ Known)

Assumptions:

- Population standard deviation σ is known
- Population is normally distributed
- If population is not normal, use Central Limit Theorem

Confidence interval estimate

$$\overline{X} \pm Z\frac{\sigma}{\sqrt{n}}$$

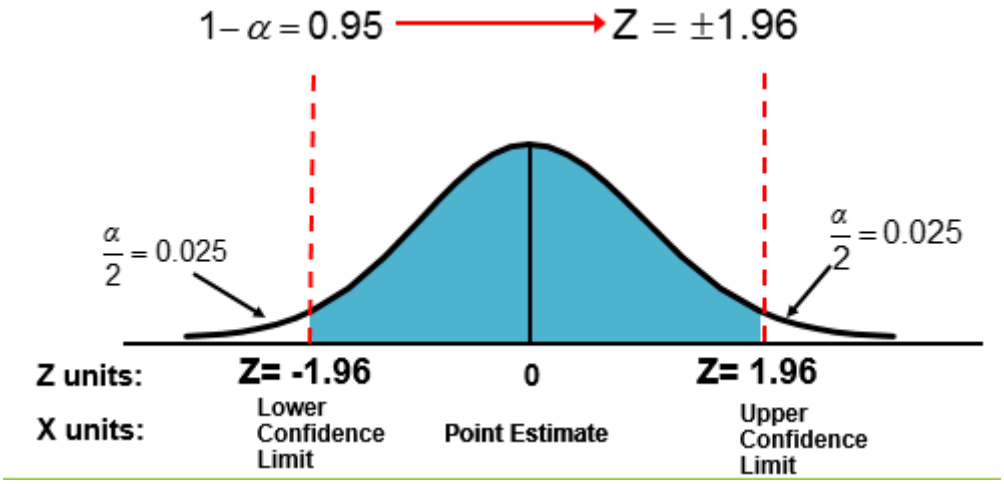
Where $\acute{X}$ is the point estimate

Z is the normal distribution critical value for a probability of $\frac{\alpha}{2}$ in each tail

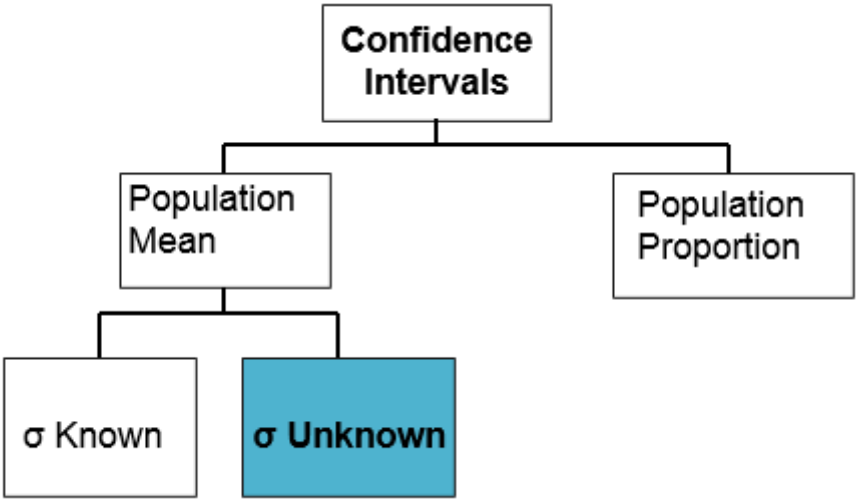
$\frac{\sigma}{\sqrt{n}}$ is the standard error

Finding the Critical Z Value

Consider a 95% confidence interval



Confidence Intervals



Confidence Interval for μ (σ Unknown)

If the population standard deviation σ is unknown, we can **substitute the sample standard deviation, s.**

This introduces extra uncertainty, since s is variable from sample to sample. So we **use the Student t distribution** instead of the normal distribution:

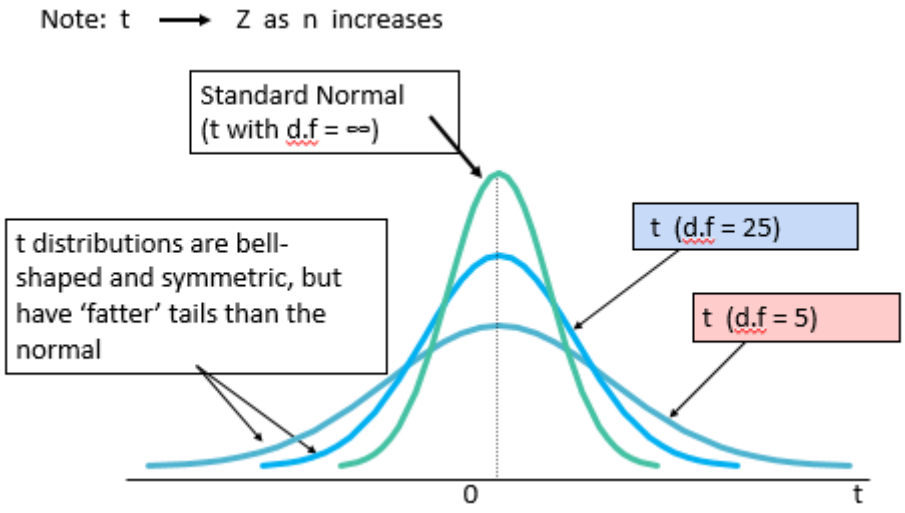
- The t value depends on degrees of freedom denoted by sample size minus 1
- D.f are number of observations that are free to vary after sample mean has been calculated

The confidence interval estimate when σ is unknown is:

$$\overline{X} \pm t_{n-1} \frac{S}{\sqrt{n}}$$

Where t is the critical value of the t distribution with n – 1 degrees of freedom and an area of $\frac{\alpha}{2}$ in each tail.

Student's t Distribution



Hypothesis Testing

A hypothesis is a statement (assumption) about a population parameter.

The Null Hypothesis, H_0

States the belief or assumption in the current situation (status quo).

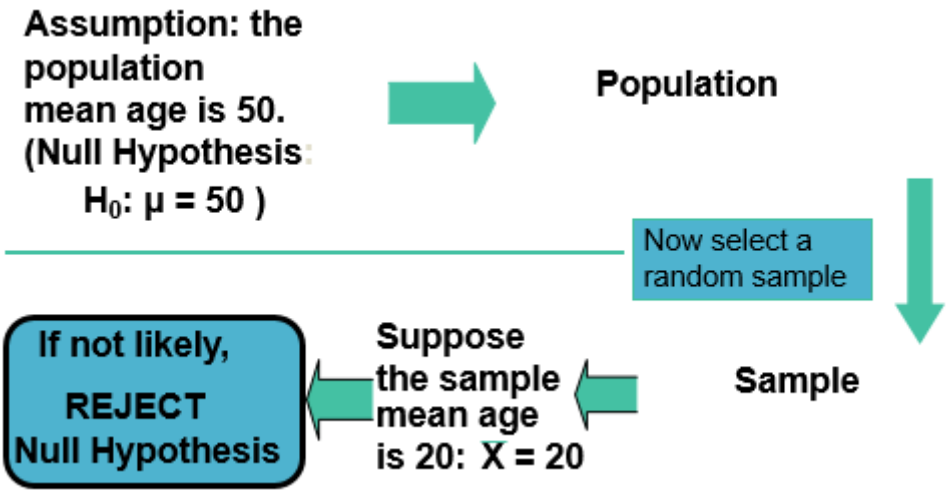
Begin with the assumption that the null hypothesis is true. Similar to the notion of innocent until proven guilty.

Always contains '=', '≤' or '≥' sign.

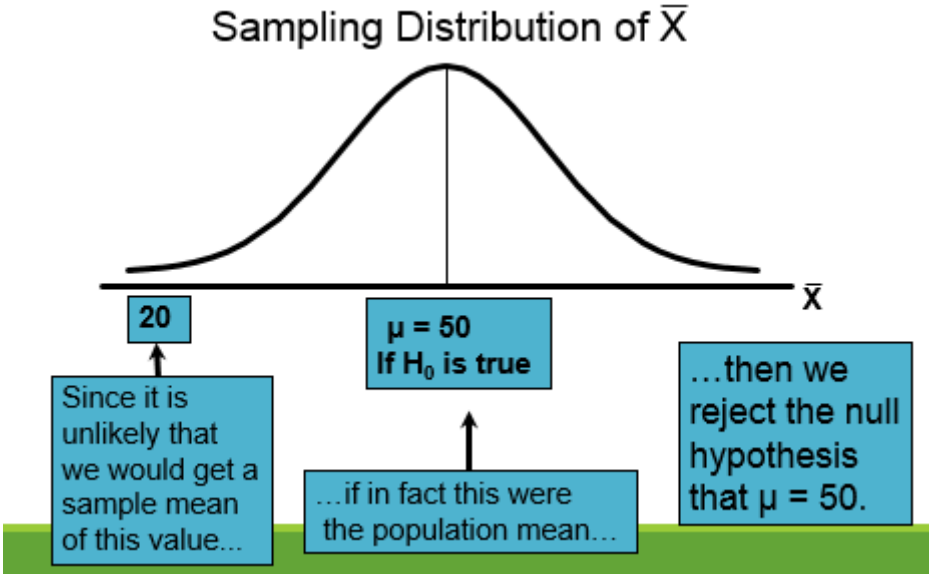
It's always about a population parameter

The Alternative Hypothesis, H_1

The alternative hypothesis is the opposite of the null hypothesis in which it challenges the status quo. Can only contain either the '<, '>' or '≠' sign.



Reason for Rejecting the Null Hypothesis



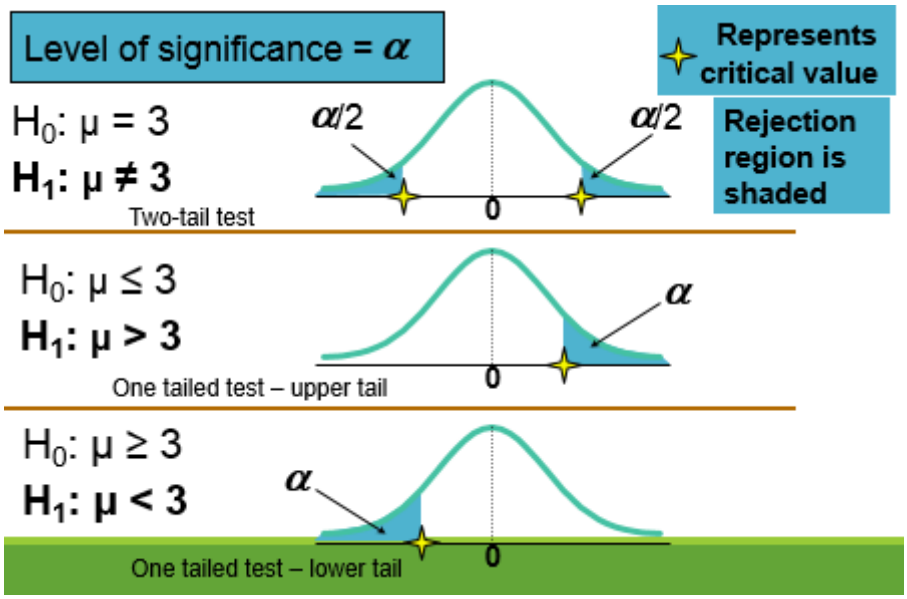
The Level of Significance, α

This defines the unlikely values of the sample statistic if the null hypothesis is true.

- Defines **rejection region** of the sampling distribution
- Designated by α (level of significance)

It is selected by the researcher at the beginning and provides the critical value(s) of the test.

Level of Significance and the Rejection Region



Errors in Making Decisions

Type I error

- Rejected a true null hypothesis
- Considered a serious type of error

Type II error

- Failed to reject a false null hypothesis (or accept a null hypothesis when it is false)

The probability of Type I error is α

- Called **level of significance** of the test
- Set by the researcher in advance

The probability of Type II error is β

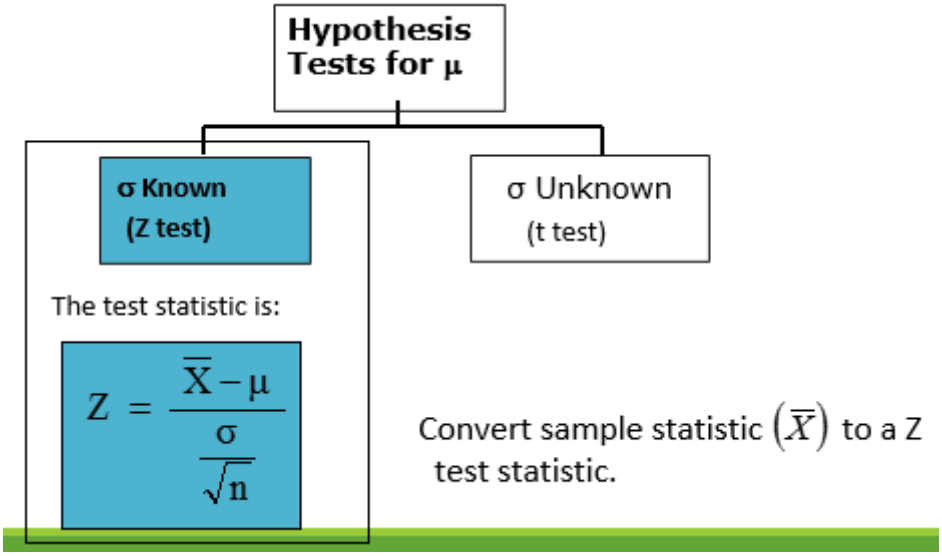
Outcome and Probabilities

Possible Hypothesis Test Outcomes

	Actual Situation	
Decision	H_0 True	H_0 False
Do Not Reject H_0	No error $(1 - \alpha)$	Type II Error (β)
Reject H_0	Type I Error (α)	No Error $(1 - \beta)$

Key: Outcome (Probability)

Z Test of Hypothesis for the Mean (σ Known)



Critical Value Approach to Testing

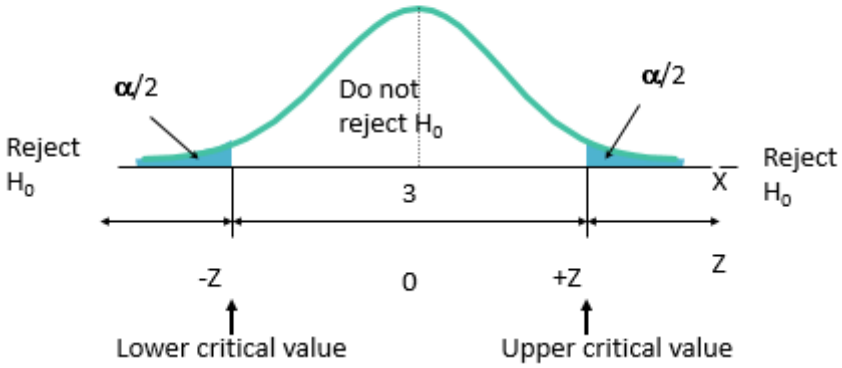
For a two-tailed test for the mean, σ known:

1. Convert sample statistic ($\overline{X}$), to the test statistic (Z statistic).
2. Determine the critical Z values for a specified level of significance α from a table.
3. Decision rule: if the test statistic falls in the rejection region, reject H_0; otherwise do not reject H_0.

Two-tail Tests

There are two cut-off values **(critical values)**, defining the regions of rejection.

Eg. $H_0: \mu = 3$ vs $H_1: \mu \neq 3$



Six Steps in Hypothesis Testing

- State the null hypothesis, H_0, and the alternative hypothesis, H_1
- Choose the level of significance, α, and the sample size, n
- Determine the appropriate test statistic and sampling distribution
- Determine the critical values that divide the rejection and non-rejection regions
- Collect data and compute the value of the test statistic
- Make the statistical decision and state the managerial conclusion
    - If the test statistic falls into the non-rejection region, do not reject the null hypothesis H_0. If the test statistic falls into the rejection region, reject the null hypothesis.
    - Express the managerial conclusion in the context of the real-world problem.

The P-value approach to testing

P-value: Probability of obtaining a test statistic more extreme.
(≤ or ≥) than the observed sample value, given H_0 is true.

- Also called the observed level of significance
- Smallest value of α for which H_0 can be rejected
- Obtain the P-value from table

If p-value < α, reject H_0
If p-value ≥ α, do not reject H_0

One-tail Tests

In many cases, the alternative hypothesis focuses on a particular direction

$H_0: \mu \geq 3$
$H_1: \mu < 3$

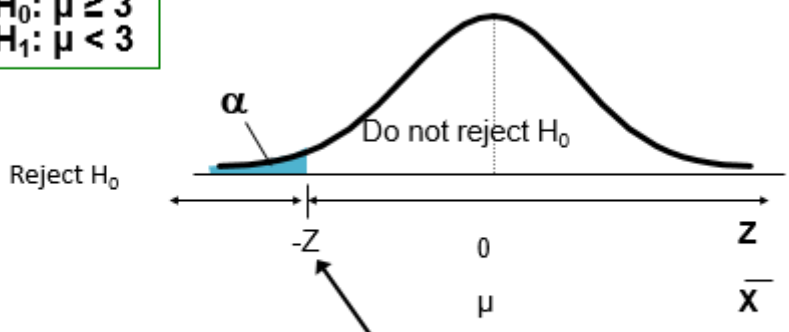
This is a **lower**-tail test since the alternative hypothesis is focused on the lower tail below the mean of 3.

$H_0: \mu \leq 3$
$H_1: \mu > 3$

This is an **upper**-tail test since the alternative hypothesis is focused on the upper tail above the mean of 3.

Lower-tail Tests

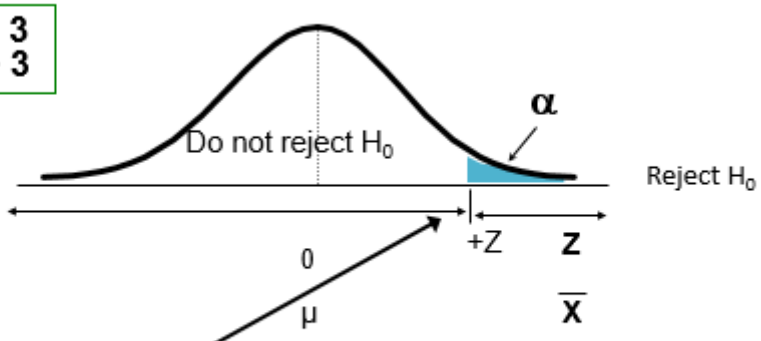
$H_0: \mu \geq 3$
$H_1: \mu < 3$



There is only one critical value (in this case -Z), since the rejection region is in only one tail (the lower tail).

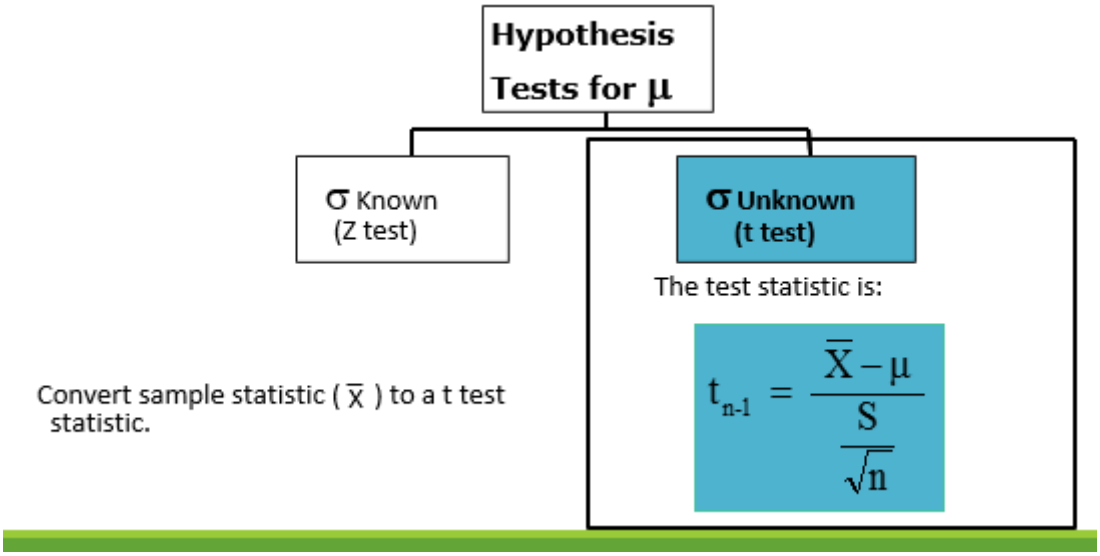
Upper-tail Tests

$H_0: \mu \leq 3$
$H_1: \mu > 3$



There is only one critical value (in this case +Z), since the rejection region is in only one tail (the upper tail)

t Test of Hypothesis for the mean (σ Unknown)



Simple Linear Regression

Introduction to Regression Analysis

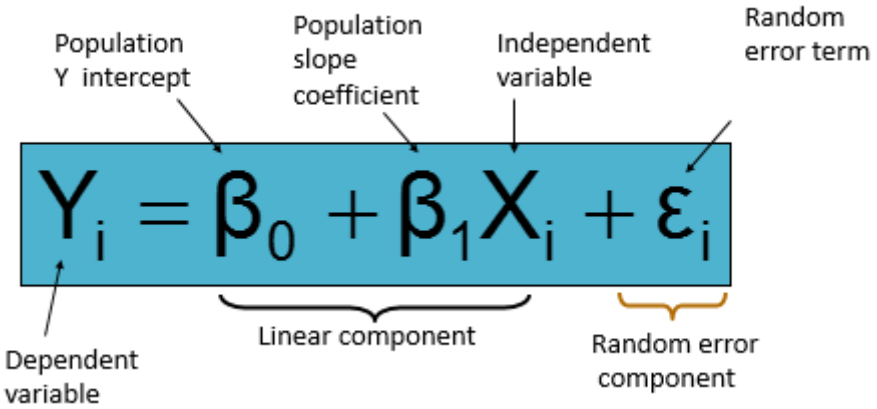
Regression analysis is used to:

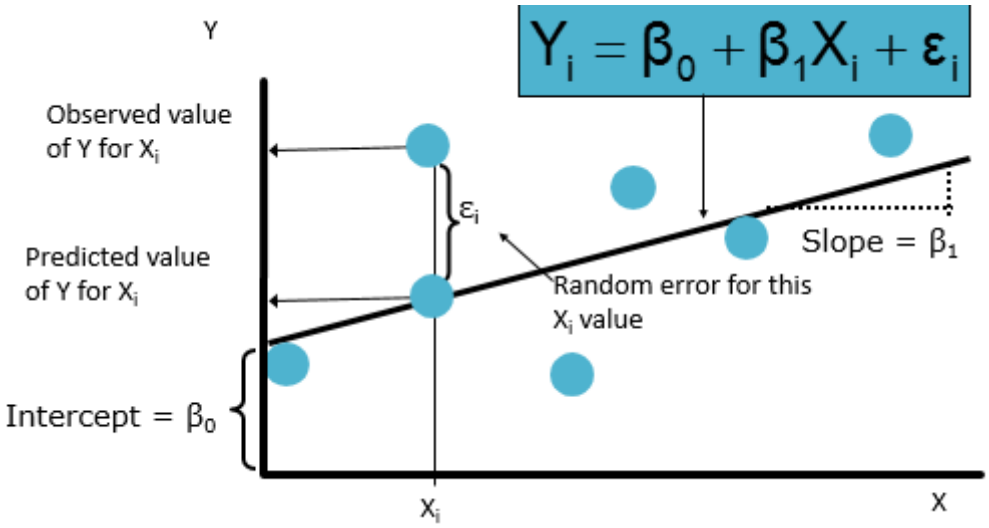
- Predict the value of a dependent variable (Y) based on the value of at least one independent variable (X)
- Explain the impact of changes in an independent variable on the dependent variable

Dependent variable (Y): the variable we wish to predict or explain (response variable)
Independent variable (X): the variable used to explain the dependent variable (explanatory variable)
Simple linear regression:

- Only **one independent variable,** x
- Relationship between X and Y described by a linear function
- Changes in Y are assumed to be caused by changes in X

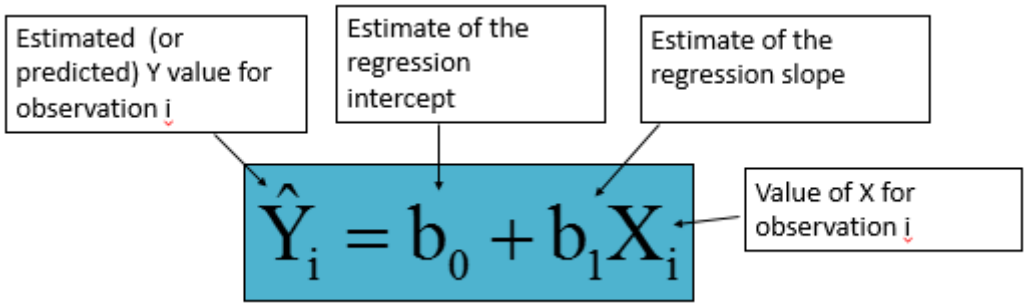
Simple Linear Regression Model





Simple Linear Regression Equation (Prediction Line)

The simple linear regression equation provides an estimate of the population regression line.



The individual random error terms e_i have a mean of zero.

Least Squares Method

b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared differences between actual values (Y) and predicted values ($\hat{Y}$).

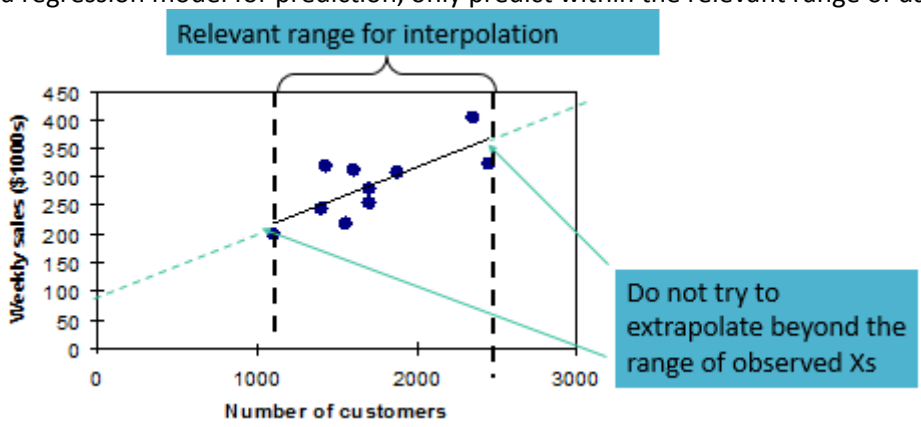
$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (b_0 + b_1 X_i))^2$$

b_0 is the estimated average value of Y when the value of X is zero.

b_1 is the estimated change in the average value of Y as a result of a one-unit change in X.

Interpolation vs. Extrapolation

When using a regression model for prediction, only predict within the relevant range of data.



Measures of Variation

Total variation is made up of two parts.

$$SST = SSR + SSE$$

Total Sum of Squares	Regression Sum of Squares	Error Sum of Squares
$SST = \sum(Y_i - \bar{Y})^2$	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	$SSE = \sum(Y_i - \hat{Y}_i)^2$
Measures the variation of the Y_i values around their mean Y.	Explained variation attributable to the relationship between X and Y.	Variation attributable to factors other than the relationship between X and Y.

Coefficient of Determination, r^2

The **coefficient of determination** is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

The coefficient of determination is also called **r-squared** and is denoted **r^2.**

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Note: $0 \le r^2 \le 1$

Standard Error of the Estimate

The standard deviation of the variation of observations around the regression line is estimated by:

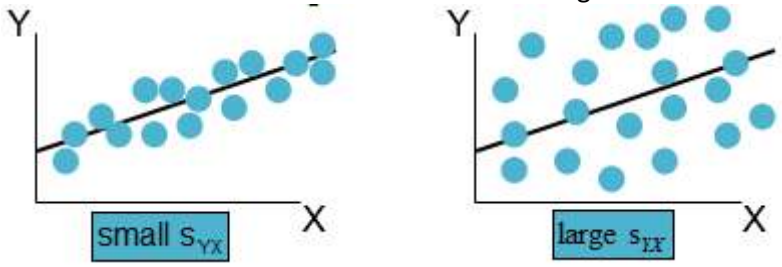
$$S_{YX} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}}$$

Where SSE = error sum of squares

n = sample size

Comparing Standard Errors

S_{YX} is a measure of the variation of observed Y values from the regression line.



The magnitude of S_{YX} should always be judged relative to the size of the Y values in the sample data.

Assumptions of Regression

Use the acronym LINE

Linearity – The underlying relationship between X and Y is linear
Independence of errors – Error values are statistically independent
Normality of error – Error values (ε) are normally distributed for any given value of X
Equal variance – The probability distribution of the errors has constant variance

Residual Analysis

The residual for observation i, e_i, is the difference between its observed and predicted value.

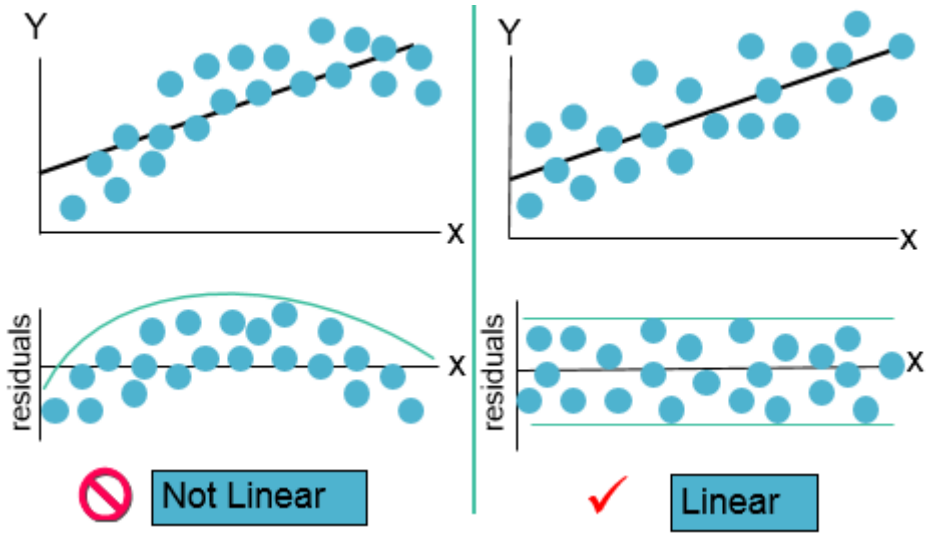
$$e_i = Y_i - \hat{Y}_i$$

Check the assumptions of regression by examining the residuals:

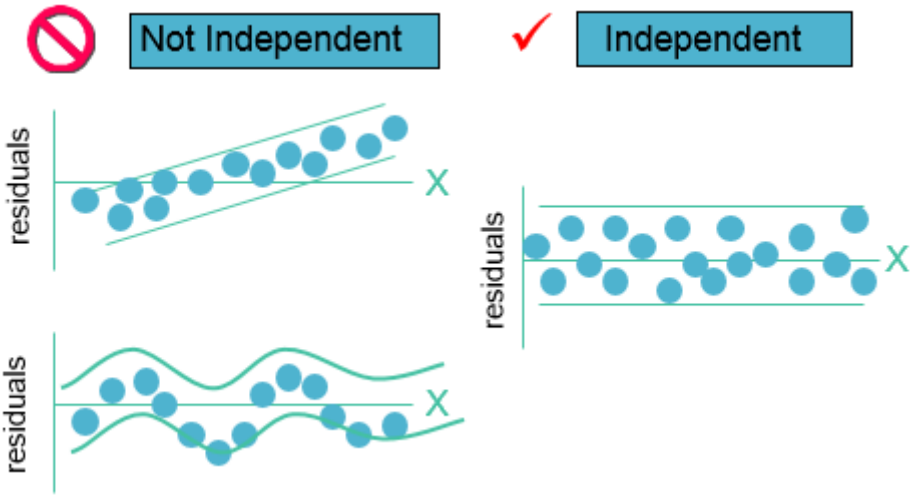
- Examine the linearity assumption
- Evaluate independence assumption
- Evaluate normal distribution assumption
- Examine for constant variance for all levels of X

Graphical Analysis of Residuals:

- Can plot residuals vs X.

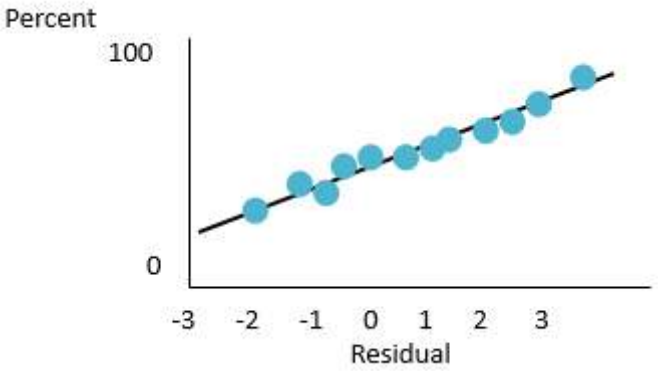


Residual Analysis for Independence

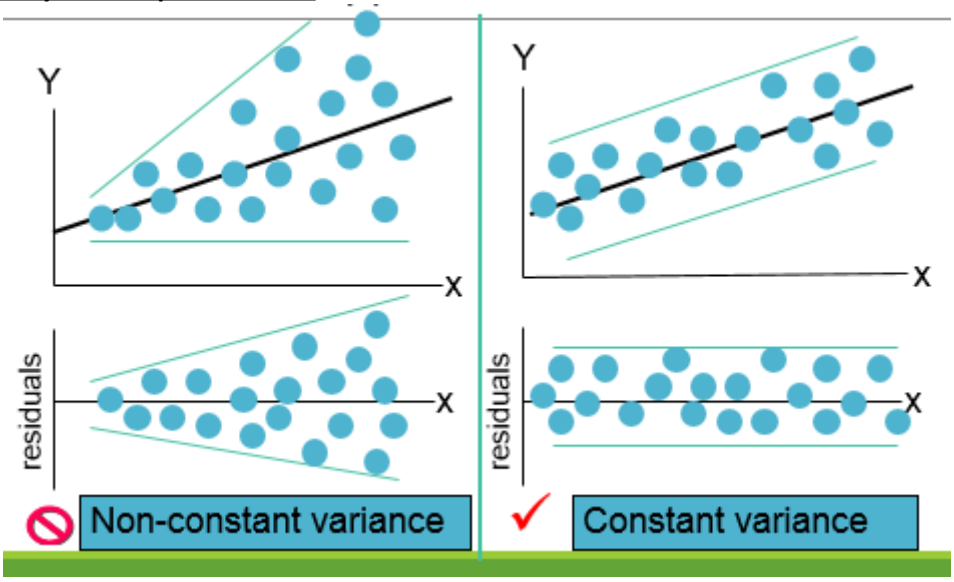


Residual Analysis for Normality

A normal probability plot of the residuals can be used to check for normality.



Residual Analysis for Equal Variance



Inferences about the Slope

The standard error of the regression slope of coefficient (b_1) is estimated by:

$$S_{b_1} = \frac{S_{YX}}{\sqrt{SSX}} = \frac{S_{YX}}{\sqrt{\sum (X_i - \overline{X})^2}}$$

Where S_{b1} = estimate of the standard error of the least squares slope

$$S_{YX} = \sqrt{\frac{SSE}{n-2}}$$ = standard error of the estimate

Inference about the Slope: t Test

t test for a population shape: Is there a linear relationship between X and Y?

Null and alternative hypotheses

H_0: $\beta_1 = 0$ (no linear relationship)

H_1: $\beta_1 \neq 0$ (linear relationship does exist)

$$t = \frac{b_1 - \beta_1}{S_{b_1}}$$

Test statistic with d.f. = n – 1

Where b_1 = regression slope coefficient

β_1 = hypothesized slope
S_b = standard error of the slope

F Test for Significance

F Test statistic:

$$F = \frac{MSR}{MSE} \quad \text{,where} \quad MSR = \frac{SSR}{k} \qquad MSE = \frac{SSE}{n-k-1}$$

F follows an F distribution with k numerator and (n – k – 1) denominator degrees of freedom.
k = the number of independent (explanatory) variables in the regression model.

Multiple Linear Regression

The Multiple Regression Model

Multiple Regression Model with k Independent Variables:

Y-intercept | Population slopes | Random Error

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

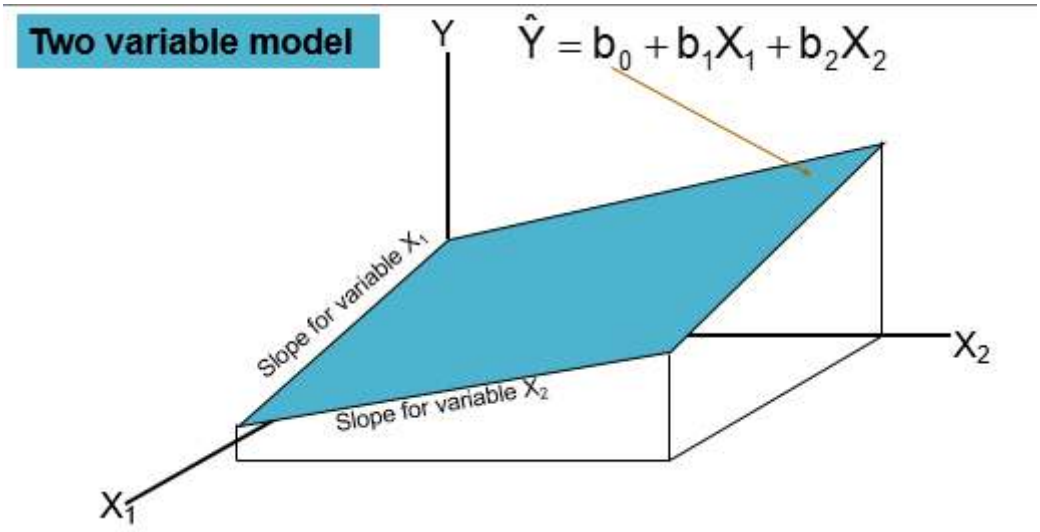
Multiple Regression Equation

The coefficients of the multiple regression model are estimated using sample data.

Multiple regression equation with k independent variables:

Estimated (or predicted) value of Y | Estimated intercept | Estimated slope coefficients

$$\hat{Y}_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki}$$



Are Individual Variables Significant?

Shows if there is a linear relationship between the variable X_i and Y.

Hypotheses

H_0: $\beta_j = 0$ (no linear relationship)

H_1: $\beta_j \neq 0$ (linear relationship does exist)

Use t tests of individual variable slopes (between X_i and Y)

Test statistic:

$$t = \frac{b_j - 0}{S_{b_j}}$$, (df = n- k – 1)

Coefficient of Multiple Determination

Reports the proportion of total variation in Y explained by all X variables taken together.

$$r^2 = \frac{SSR}{SST} = \frac{\text{regression sum of squares}}{\text{total sum of squares}}$$

Adjusted r^2

r^2 never decreases when a new X variable is added to the model.

- This can be a disadvantage when comparing models.

What is the net effect of adding a new variable?

- We lose a degree of freedom when a new X variable is added
- Did the new X variable add enough explanatory power to offset the loss of one degree of freedom?

Shows the proportion of variation in Y explained by all X variables adjusted for the number of X variables used.

$$r^2_{adj} = 1 - \left[(1 - r^2) \left(\frac{n-1}{n-k-1} \right) \right]$$

Where n = sample size, k = number of independent variables

- Penalises excessive use of unimportant independent variables
- Smaller than r^2
- Useful in comparing among models

Is the Model Significant?

F Test for Overall Significance of the Model:

- Shows if there is a linear relationship between all of the X variables considered together and Y.

Hypotheses:

$H_0: \beta_1 = \beta_2 = \ldots = \beta_k = 0$ (no linear relationship)

$H_1:$ at least one $\beta_i \neq 0$ (at least one independent variable affects Y)

F Test for Overall Significance

Test statistic:

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}}$$

Where F has:

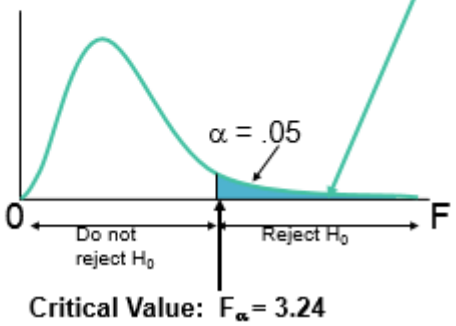
k = numerator

n – (k + 1) = (n – k – 1) degrees of freedom

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_1:$ at least one $\beta_i \neq 0$

$\alpha = .05$; $df_1 = 3$ $df_2 = 16$



Critical Value: $F_\alpha = 3.24$

Test Statistic:

$$F = \frac{MSR}{MSE} = 15.88$$

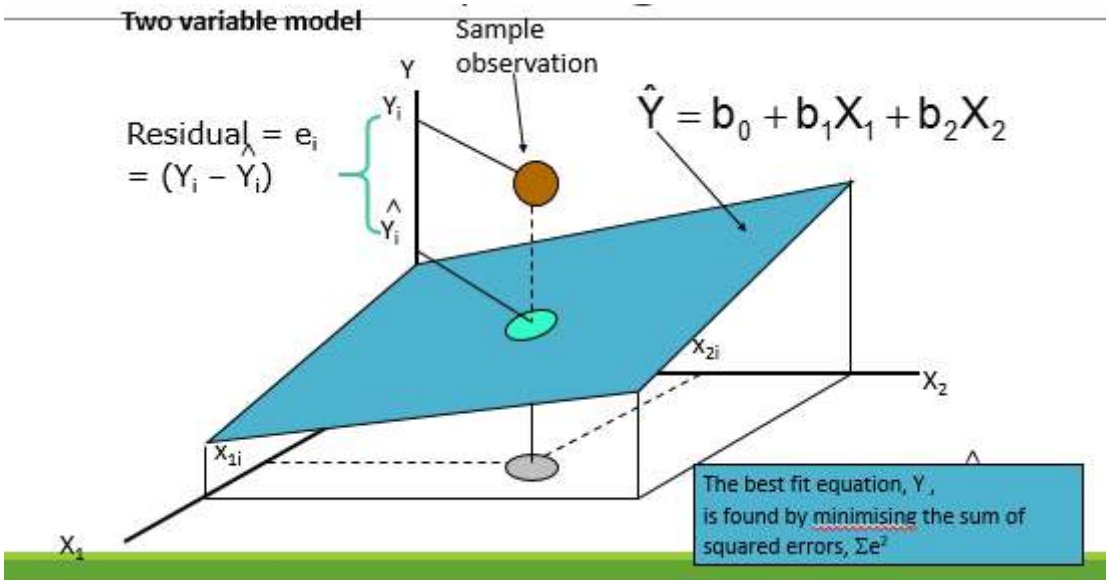
Decision:

Since F test statistic is in the rejection region (p-value < .05), reject H_0

Conclusion:

There is evidence that at least one independent variable affects sales.

Residuals in Multiple Regression



Multiple Regression Assumptions

Errors (residuals):

$$e_i = (Y_i - \hat{Y}_i)$$

Assumptions:

- The errors are normally distributed
- Errors have a constant variance
- The model errors are independent

Residual Plots Used in Multiple Regression

These residual plots are used in multiple regression:

- Residuals vs. Y_i (Predicted Sales)
- Residuals vs. X_{1i}
- Residuals vs. X_{2i}
- Residuals vs. time (if time series data)

Use the residual plots to check for violations of regression assumptions.

-> Should not see any patterns

Interaction Effects

Sometimes our model will predict that in addition to individual variables influencing our dependent variable, some combination of these variables will differentially effect the dependent variable.