



Fundamentals of Statistics

Rakhi Singh



Measures of Central Tendency



They help us to give a number that gives an overview of data.

1. Mean : It is the average of all numbers.
2. Median : It is the central observation of data
3. Mode: It is the observation which occurs maximum number of times.

It is possible for a data set to have same mean but different median/mode or vice versa



Quartile, Decile and Percentiles

1.Quartile- They divide data in four parts.

$Q1=((n+1)/4)^{\text{th}}$ item). There are three quartiles Q1, Q2 and Q3. Q1 is also known as 25 percentile. Q2 is middle quartile equal to median. Q3 is upper quartile can be called as upper quartile

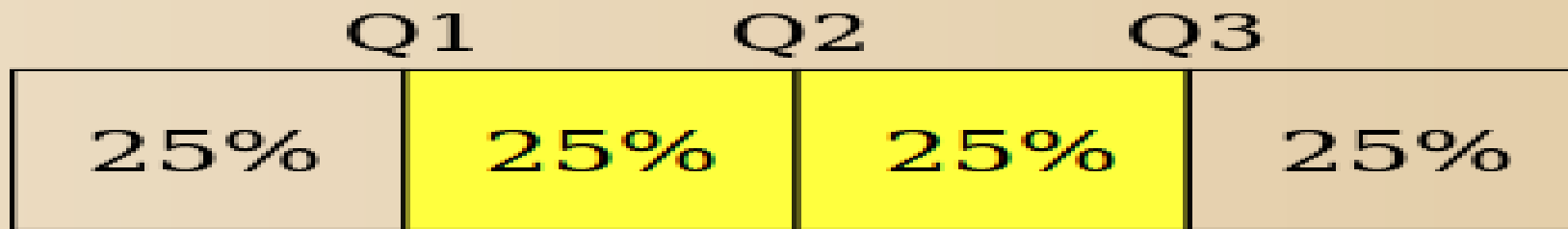
2.Decile- Divide data in 10 parts. $D4=(4(n+1)/10)^{\text{th}}$ item. D1 is 10 percentile, D2 is 20 percentile and so on.

3.Percentile- Divide data in 100 parts. P78 (78 percentile) $= (78(n+1)/100)^{\text{th}}$ item

Quartiles

- Quartiles: There are 3 quartiles, first, second and third. They are the observations which divide data into four equal parts.
- Example: 2,3,6,7,10,11,11,12,12,12,13,14,17,18,19
- $Q1 = (n+1)/4$, $Q2 = 2(n+1)/4$, $Q3 = 3(n+1)/4$

Interquartile Range



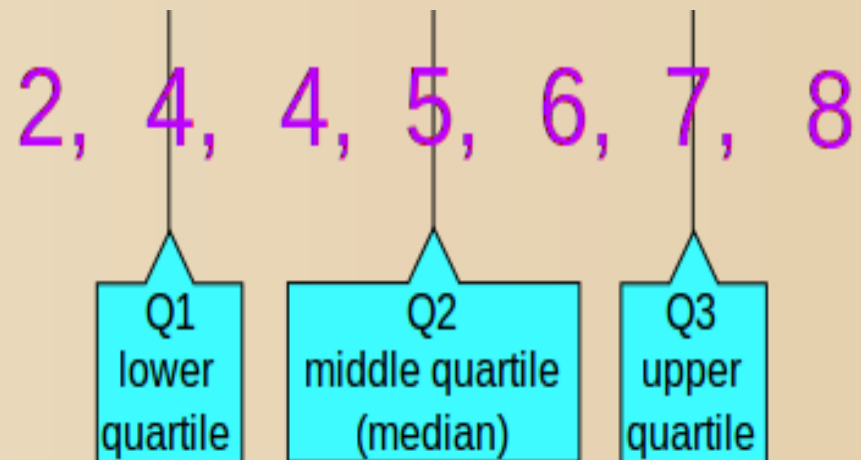
Interquartile Range
= $Q3 - Q1$

Quartiles

Example: 5, 7, 4, 4, 6, 2, 8

Put them in order: 2, 4, 4, 5, 6, 7, 8

Cut the list into quarters:



And the result is:

Quartile 1 (Q1) = 4

Quartile 2 (Q2), which is also the , = 5

Quartile 3 (Q3) = 7

Functions in R

R functions: `x<-c(2,4,6,8,10,12,14,16,18,20)`

`mean(x)`

`[1] 11.72727`

`median(x)`

`[1] 12`

`Summary(x)`

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.00	7.00	12.00	11.73	17.00	20.00

`Quantile(x)`

0%	25%	50%	75%	100%
2	7	12	17	20

Other measures of Central Tendency

- Harmonic mean: It is a specialized average. It is used where rates are to be averaged. Since rate consists of two values and if we want to average the rate in terms of numerator factor then use HM otherwise use AM.

$$1/HM = (1/x_1 + 1/x_2 + \dots + 1/x_n)/n$$

- Geometric mean: It is a specialized average. It is used when data is exponentially rising and declining. Example- (Population growth, compound interest)


$$GM = \text{nth root}(x_1 * x_2 * x_3 * \dots * x_n)$$

- Relationship between AM, H.M and GM $GM =$

- $G.M = \sqrt{(A.M * H.M)}$

Measures of Central Tendency



- Arithmetic , geometric and harmonic mean are called as mathematical average and median and mode are positional average.
 - Partition values: These are measures which divide the data into partition.
- 

Measures of Central Tendency

R functions: `x<-c(2,4,6,8,10,12,14,16,18,20)`

`mean(x)`

`[1] 11.72727`

`median(x)`

`[1] 12`

`Summary(x)`

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
2.00	7.00	12.00	11.73	17.00	20.00

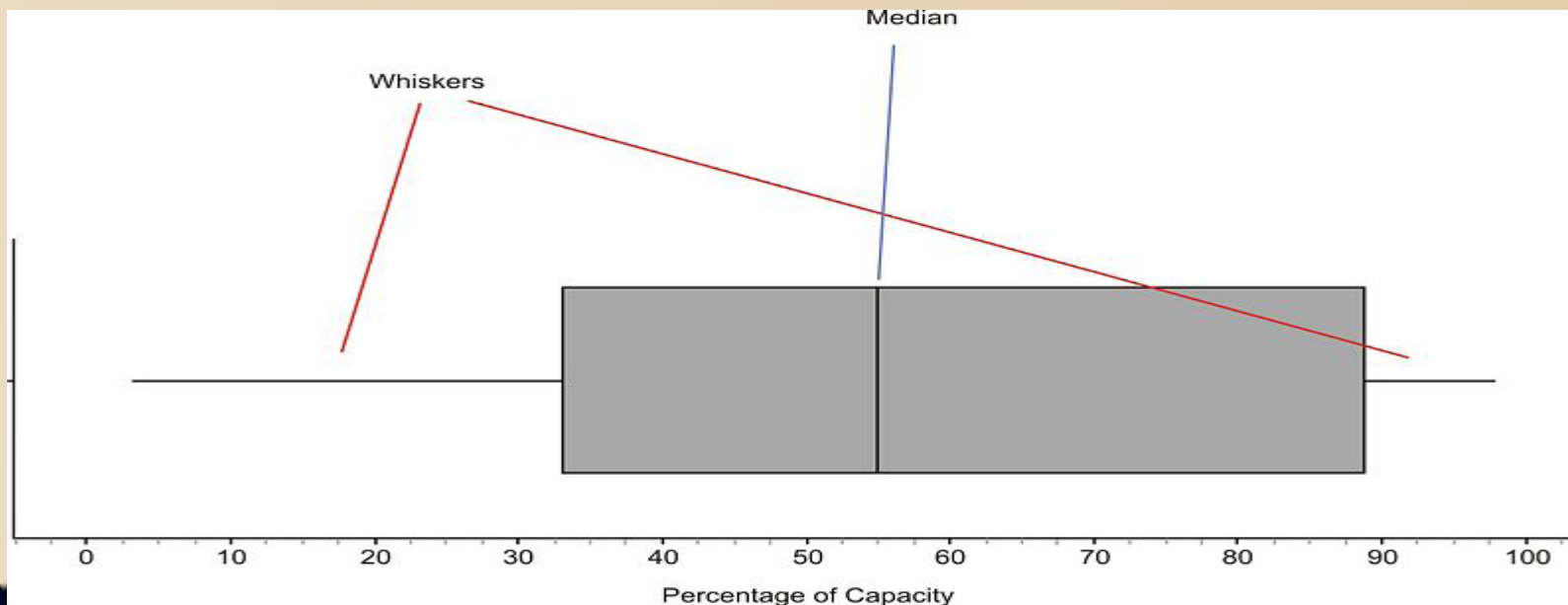
`Quantile(x)`

0%	25%	50%	75%	100%
2	7	12	17	20

BOX AND WHISKER PLOT

(Graphical view of summary)

A box-and-whisker plot is a very convenient and informative way to display the information captured in the five number summary. A box-and-whisker plot shows the center and spread of the values on a single quantitative variable. To create the 'box' part of the plot, first draw a rectangle that extends from the lower (first) quartile to the upper (third) quartile. Then draw a line through the interior of the rectangle at the median. Finally, connect the ends of the box to the minimum and maximum values using line segments to form the 'whiskers'.



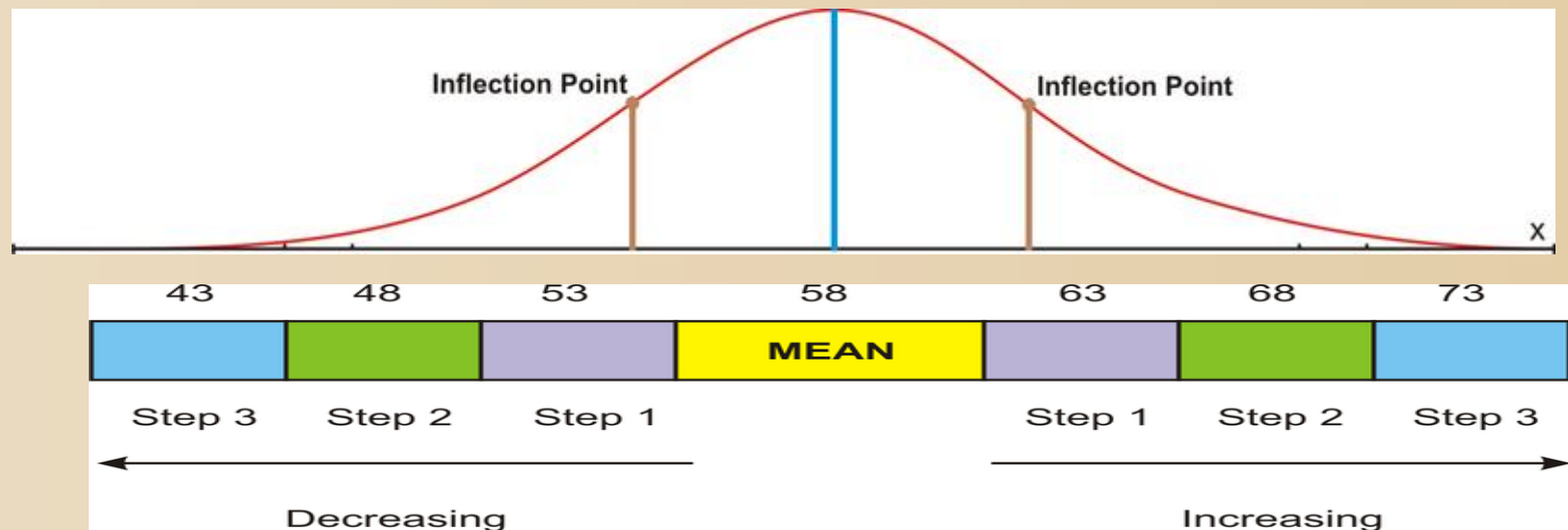
Measures of Dispersion

Variability is universal phenomenon. We may questions like how is growth in shares price, sales growth, investors growth, acadmics growth of students etc. For this we use different measures like

- Range
- Mean Deviation
- Standard Deviation
- Variance
- Quartile Deviation

Measures of Dispersion

When data is normally or nearly normally distributed, there are two preferred measures of center and spread. These are the arithmetic mean and the standard deviation. You are already familiar with the mean. The standard deviation of a data set tells us how it is spread out. The larger the standard deviation is, the more spread out the data is. If we know mean and standard deviation of data then we can calculate how far the value will be from mean.



Mean Deviation

- Mean Deviation: It is the average of absolute deviation from the mean.

$$\mu = \sum |(x(i) - \bar{X})|/n \text{ where } i=1 \text{ to } n$$

- Median Deviation

$$\mu(\text{median}) = \sum |(x(i) - \text{median}(x))|/n \text{ where } i=1 \text{ to } n$$

- Mode Deviation

$$\mu(\text{mode}) = \sum |(x(i) - \text{mode}(x))|/n \text{ where } i=1 \text{ to } n$$

How to calculate mean deviation

•Example: the Mean Deviation of 3, 6, 6, 7, 8, 11, 15, 16

Step 1: Find the mean:

$$\begin{aligned} \bullet M &= \frac{3 + 6 + 6 + 7 + 8 + 11 + 15 + 16}{8} \\ &= 9 \end{aligned}$$

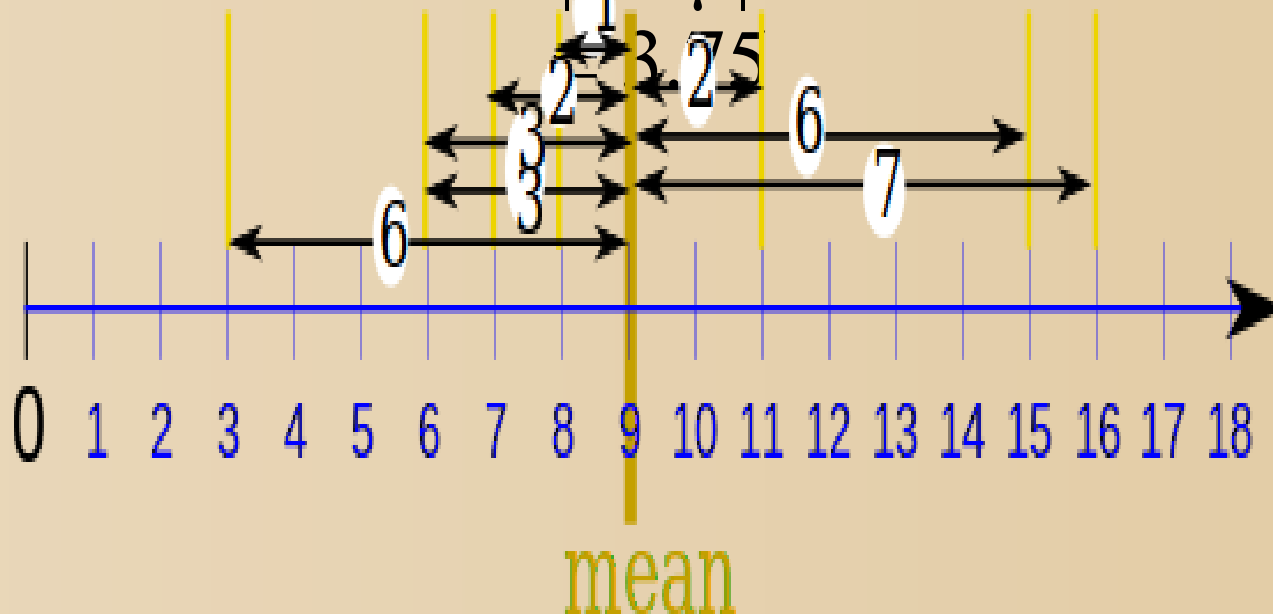
Step 2: Find the Absolute Deviations:

•x	x - μ
•3	6
•6	3
•6	3
•7	2
•8	1
•11	2
•15	6
•16	7
•	$\Sigma x - \mu = 30$

Step 3. Find the Mean Deviation:

Mean Deviation =

$$\Sigma|x - \mu| / 9 = 30$$



Standard Deviation

•**Standard Deviation:** It is the underroot of average of square of absolute deviation from the mean or how far each point is from the mean. How we calculate standard deviation depends on whether data is sample or entire population.

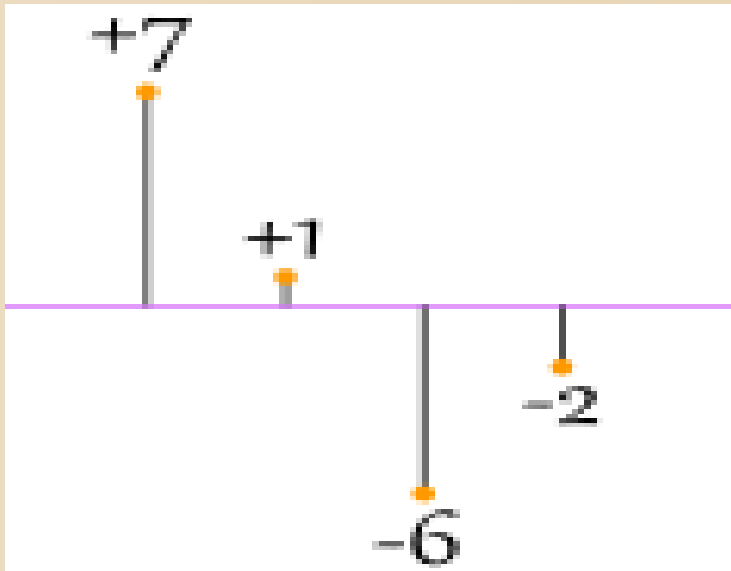
$$\sigma = \sqrt{(\sum (X(i) - \bar{X})^2) / n} \text{ where } i=1 \text{ to } n$$

SD is the measure of dispersion and variability for normally distributed graph.

•**Relationship between Standard deviation and mean deviation**

$$4 \text{ SD} = 5 \text{ MD} = 6 \text{ QD}$$

Why we need to square

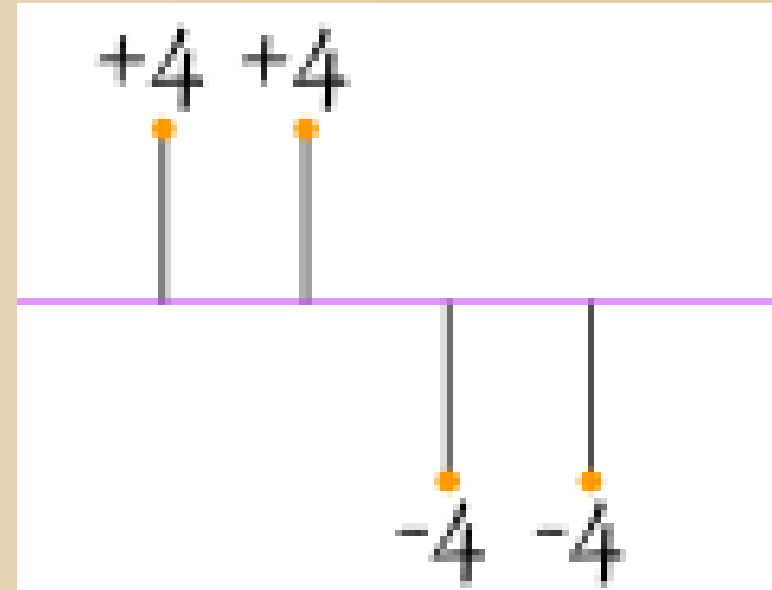


Mean Deviation

$$(|7| + |1| + |-6| + |-2|)/4 = 7 + 1 + 6 + 2 \div 4 = 4$$

Standard Deviation

$$(\sqrt{49 + 1 + 36 + 4}) = \sqrt{(90/4)} = 4.74...$$



Mean Deviation

$$|4| + |4| + |4| + |4| = 16/4 = 4$$

Standard deviation

$$(\sqrt{16 + 16 + 16 + 16})/4 = \sqrt{(64/4)} = 4.0$$

Relation between SD and mean

1) For any distribution 75% of data will lie within two standard deviation of the mean.

Example : Let $n=200$, mean = 80, SD = 10
then 75% of 200 = 150 observation will lie between 60 and 100.

75% data between $\bar{X}-2\sigma$ and $X+2\sigma$

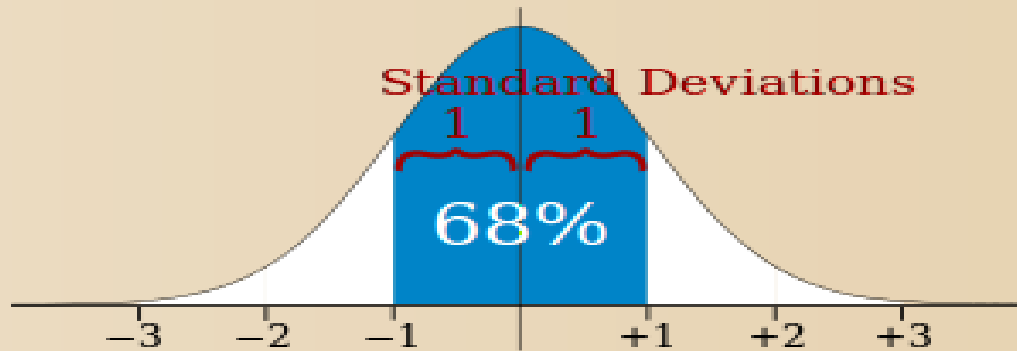
89% data between $\bar{X}-3\sigma$ and $X+3\sigma$

2) For normal distribution this percent increases.

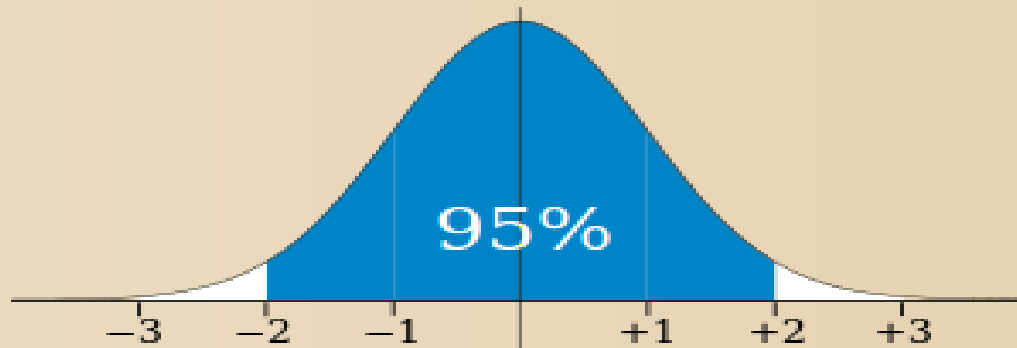
68% data between $\bar{X}-\sigma$ and $X+\sigma$

95% data between $\bar{X}-2\sigma$ and $X+2\sigma$

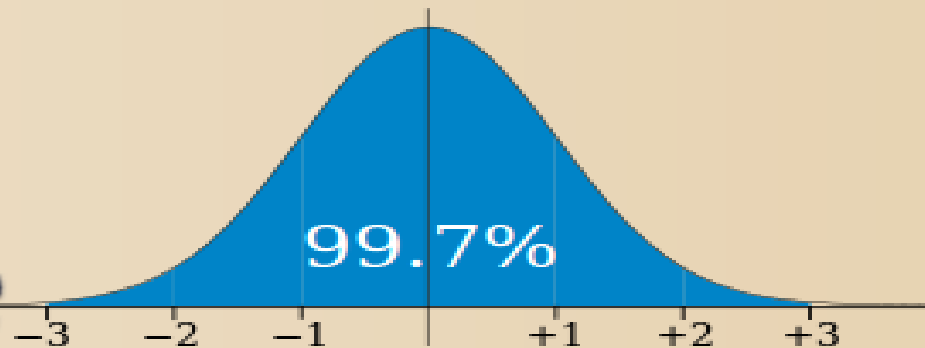
Mean and SD relation



68% of values are within
1 standard deviation of the mean

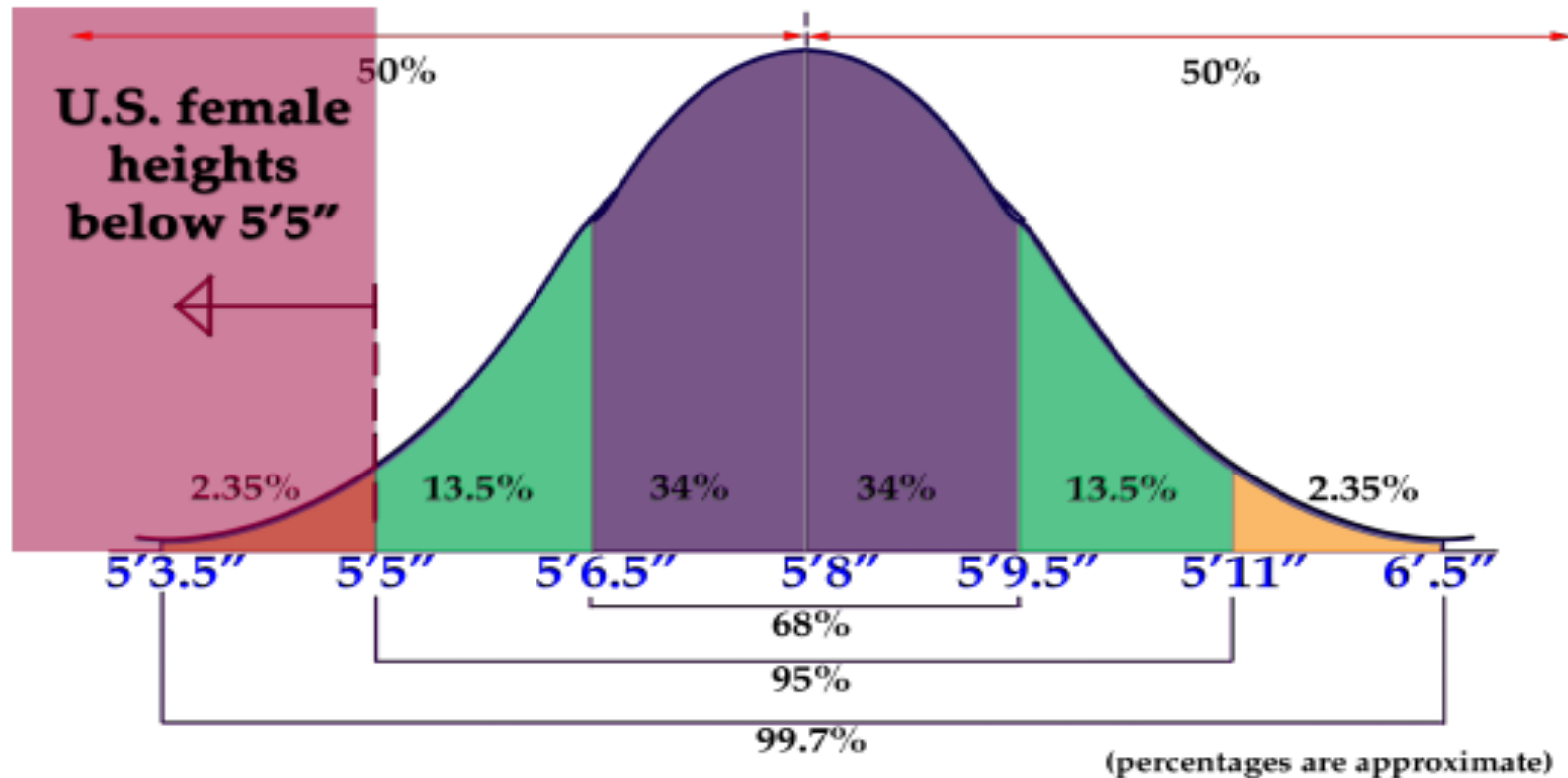


95% of values are within
2 standard deviations of the mean



99.7% of values are within
3 standard deviations of the
mean

EXAMPLE FOR MEAN AND SD RELATION



Standard Error of mean

• In statistics we work on samples and want to know how far is sample mean from other sample mean and from population mean. For this we need to plot the probability distribution of all the sample mean and calculate standard deviation also called standard error of mean. But this is not possible practically so we calculate it by a formula.

• **Standard Error of mean –**

$\text{Standard Deviation} / \sqrt{\text{sample size}}$.

So as the sample size increases standard error decreases.

It is a standard deviation of sample mean wrt true mean. We use it in hypothesis testing of mean.

The probability distribution of sample mean follows normal distribution.

Variance

It is the mean of squared deviations measured from arithmetic mean.

$$\sigma^2 = (\Sigma(X - \bar{X})^2) / n$$

- The unit of variance is square of unit of standard deviation. Variance has lot of applications like risk associated with stock market is measured in terms of variance.
- While calculating variance of population use n and for sample use $n-1$

Covariance

Covariance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together.

$$\text{COV}(X,Y) = (\Sigma(X - \bar{X})(y - \bar{Y})) / n$$

It tells how these two variables are related. A positive covariance means the variables are positively related i.e. they move together in same direction, while a negative covariance means the variables are inversely related i.e. they move in opposite direction. It can range from -infinity to +infinity. When two variables are independent then covariance will be 0 or near to 0.

Coefficient of Variation

•**Coefficient of variation:** It is represented in the form of percentage. It tells how much the standard deviation as a percentage of arithmetic mean.

$$CVD = (\sigma / \bar{X}) * 100$$

•**Coefficient of Standard deviation:** It is just the ratio of SD and Mean.


$$CSD = \text{Standard deviation} / \text{Mean}$$

•**Coefficient of Variance**

$$CV = \text{Variance} / \text{Mean}$$

CV significance



- If mean deviation is equal to mean then CV is 100 . It shows that data shows exponential distribution. If CV is almost 0 then data shows Erlang distribution. For CV greater than 100 data shows hyper exponential distribution.
 - To compare the variability between two datasets it is advisable to use CV or CSD as they might have different units or different skewness.
- 

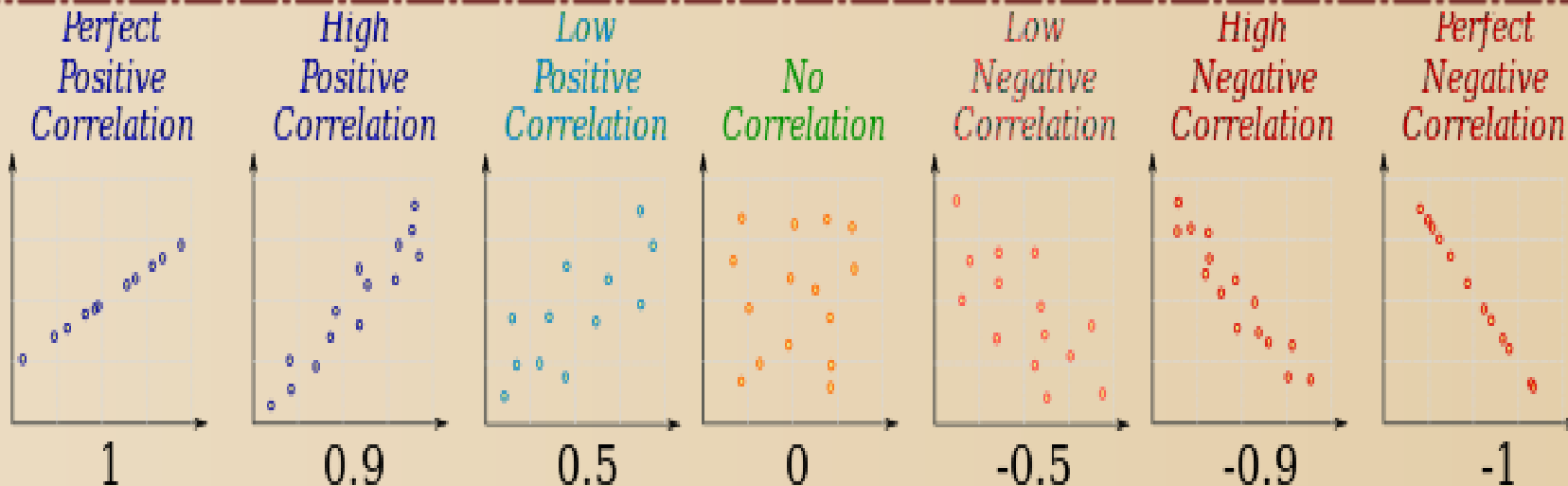
Coefficient of Correlation

In order to standardize the covariance we use correlation coefficient (Pearson correlation coefficient).

$$\text{Corr}(x,y)/r = \text{Cov}(x,y)/\sigma(x) \sigma(y)$$

Now it is easy to understand and show graphically. The correlation between two variables is a number that indicates how closely their relationship follows a straight line. Correlation ranges between -1 and +1. The extreme value indicate a perfect linear relationship while zero indicate the absence of relationship.

Correlation graph



The value of a correlation coefficient ranges between -1 and 1.

The greater the absolute value of a correlation coefficient, the stronger the linear relationship.

The strongest linear relationship is indicated by a correlation coefficient of -1 or 1.

The weakest linear relationship is indicated by a correlation coefficient equal to 0.

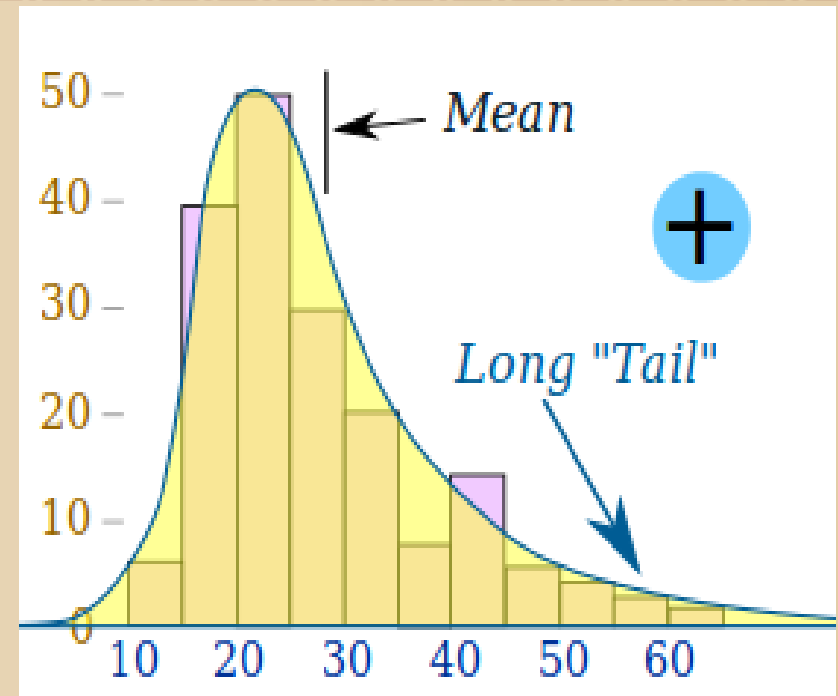
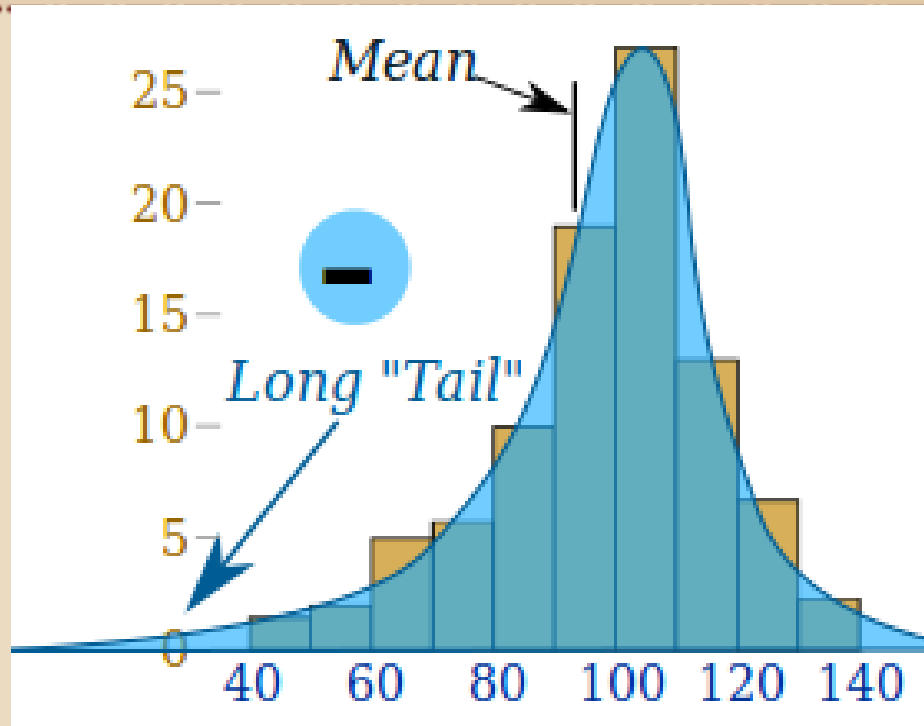
A positive correlation means that if one variable gets bigger, the other variable tends to get bigger.

A negative correlation means that if one variable gets bigger, the other variable tends to get smaller.

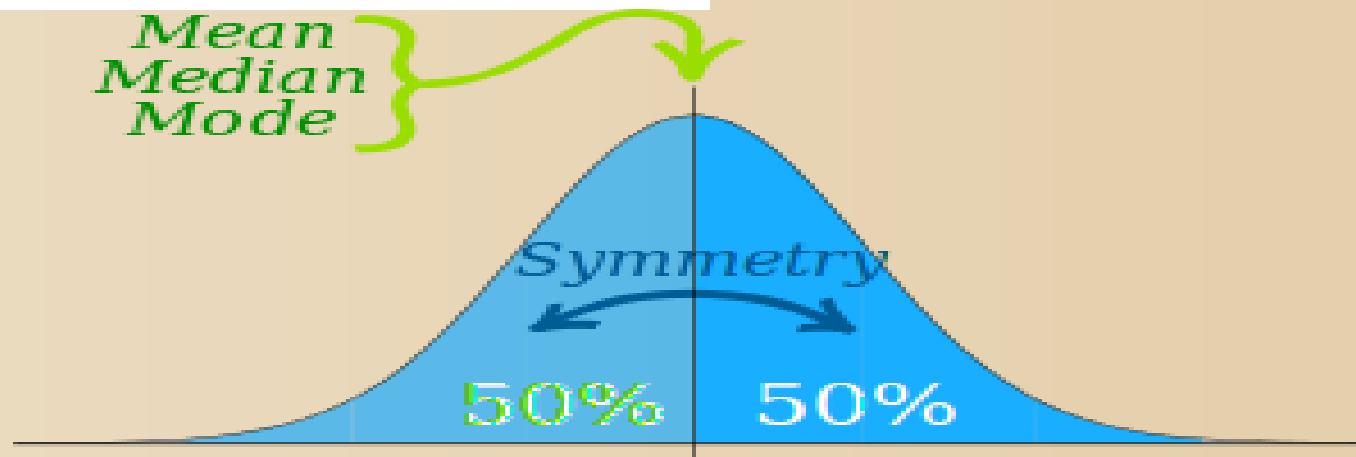
Skewness

- Skewness refers to lack of symmetry
- If $\text{Mean} = \text{Median} = \text{Mode}$ then No skewness
- If $\text{Mode} < \text{Median} < \text{Mean}$ then positive skewness also called right tailed distribution.
- If $\text{Mode} > \text{Median} > \text{Mean}$ then negative skewness also called left tailed distribution.

Skewness

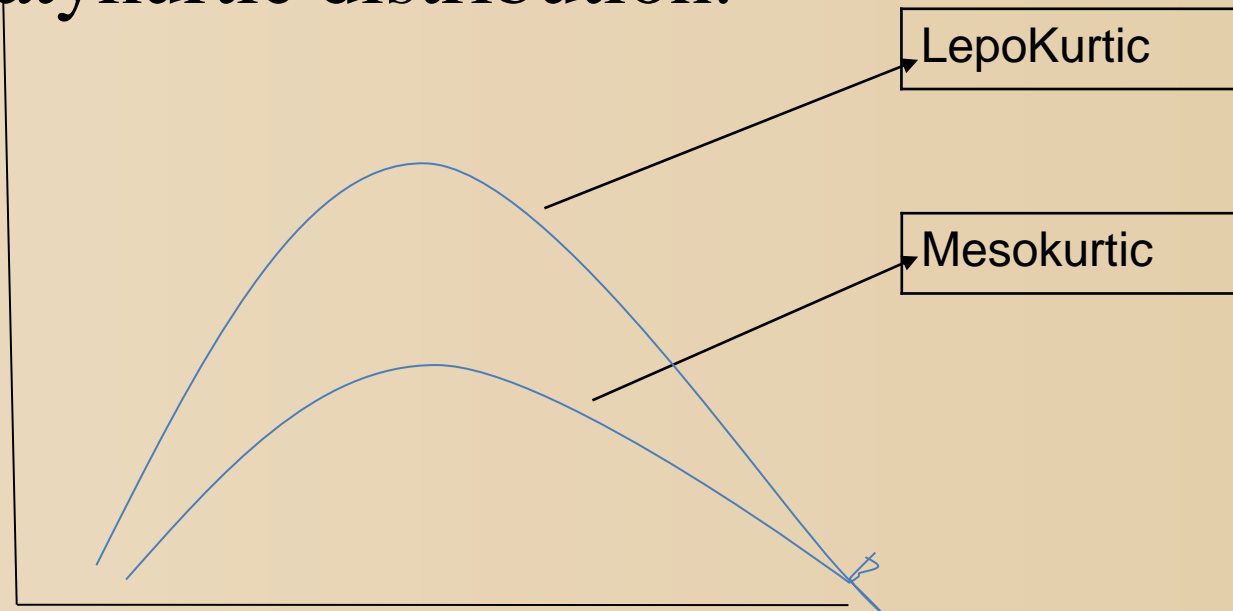


Mean
Median
Mode



Kurtosis

Kurtosis deals with the shape of the distribution. It refers to the degree of the peakedness of a distribution. A normal curve is said to have medium height or mesokurtic distribution. A more dense concentration at centre is called leopkurtic distribution and a flat topped is called platykurtic distribution.



Measurement of Kurtosis

Kelly's Measure

$$k = (Q3 - Q1) / 2(P90 - P10)$$

Measurement of Skewness

Karl Pearson's Measure

Skewness = Mean - Mode

Co-efficient of skewness, $J = (\text{Mean} - \text{Mode}) / \text{SD}$

Kelly's Measure

Skewness = $(P(90) + P(10) - 2P(50)) / 2$

Co-efficient of skewness,

- $J = (P(90) + P(10) - 2P(50)) / (P(90) - P(10))$

Linear/Non linear

Pearson product-moment correlation coefficient only measures linear relationships. Therefore, a correlation of 0 does not mean zero relationship between two variables; rather, it means zero linear relationship. (It is possible for two variables to have zero linear relationship and a strong curvilinear relationship at the same time.)

Linear transformation. A linear transformation preserves linear relationships between variables. Therefore, the correlation between x and y would be unchanged after a linear transformation. Examples of a linear transformation to variable x would be multiplying x by a constant, dividing x by a constant, or adding a constant to x .

Nonlinear transformation. A nonlinear transformation changes (increases or decreases) linear relationships between variables and, thus, changes the correlation between variables. Examples of a nonlinear transformation of variable x would be taking the square root of x or the reciprocal of x .

Linear Transformation

A linear transformation is a change to a variable characterized by one or more of the following operations: adding a constant to the variable, subtracting a constant from the variable, multiplying the variable by a constant, and/or dividing the variable by a constant.

When a linear transformation is applied to a random variable, a new random variable is created. To illustrate, let X be a random variable, and let m and b be constants. Each of the following examples show how a linear transformation of X defines a new random variable Y .

Adding a constant: $Y = X + b$

Subtracting a constant: $Y = X - b$

Multiplying by a constant: $Y = mX$

Dividing by a constant: $Y = X/m$

Multiplying by a constant and adding a constant: $Y = mX + b$

Dividing by a constant and subtracting a constant: $Y = X/m - b$



REGRESSION



Regression

Mathematical models describes many aspects of everyday life. Eg A person's weight can be described in terms of his/her calorie intake.

- Regression is concerned with specifying the relationship between a single numeric dependent variable and one or more numeric independent variable.

- Regression is rooted in a study of genetics by Sir Galton, who established the relationship between height of son and father, in 19th century.

Simple/Multiple Regression

- If there is one variable, it is called simple regression.

$$Y = ax + b \quad (1)$$

$$X = ay + b$$

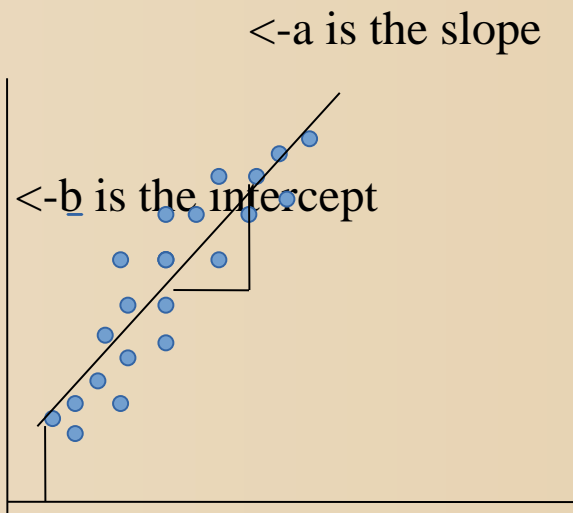
In (1)

- If more than one variable then it is called multiple regression.

$$Y = ax_1 + bx_2 + cx_3 + d$$

Relationship

- A equation of line is defined in the form of
$$y = ax + b$$
where a is intercept and b is slope. Y is dependent variable and x in independent variable.
- In Regression also we try to define the relationship with the help of deriving such equation.



Value of b and a

For $y=ax+b$

- $a(y \text{ on } x) = \text{Cov}(x, y) / \text{var}(x)$ or $r^*(\sigma(y) / \sigma(x))$
- $b = \text{mean}(y) - a * \text{mean}(x)$

For $x=ay+b$

- $a(x \text{ on } y) = \text{Cov}(x, y) / \text{var}(y)$ or $r^*(\sigma(x) / \sigma(y))$
- $b = \text{mean}(x) - a * \text{mean}(y)$

where r is Coefficient of Correlation.

Regression in R

- `lm` function is for Regression analysis in R.
- **`fit<-lm(Prestige$income~Prestige$education)`**
`scatterplot(Prestige$income~Prestige$education)`
- **`fit`** ## It will give coefficients of equation $y=ax+b$
we get intercept(b) and slope(a) of line.
- **`summary(fit)`**
It will give all the details of the Regression analysis

Residuals of Regression

Residuals: It is the most important thing of regression. The difference between the observed value of the dependent variable (y) and the predicted value (\hat{y}) is called the residual (e). $e = y - \hat{y}$

- Each data point has one residual. Both the sum and the mean of the residuals are equal to zero.

Residuals refer to the amount of variability in a dependent variable that is left over after accounting for the variability explained by the regression.

- `plot(Prestige$education, resid(fit), ylab="Residuals", xlab="education")`


- If in plot residuals are randomly dispersed around horizontal line then linear regression model is appropriate for data, otherwise non linear model is better

- Standard deviation of residuals is not zero and is reported as "residual standard error" (a measure given by most statistical softwares when running regression) is an estimate of this standard deviation, and substantially expresses the variability in the dependent variable "unexplained" by the model.

Coefficient of Determination

The coefficient of determination (denoted by R^2) is a key output of regression analysis (Square of correlation coefficient). It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.

- The coefficient of determination ranges from 0 to 1.
- An R^2 of 0 means that the dependent variable cannot be predicted from the independent variable.
- An R^2 of 1 means the dependent variable can be predicted without error from the independent variable.
- An R^2 between 0 and 1 indicates the extent to which the dependent variable is predictable. An R^2 of 0.10 means that 10 percent of the variance in Y is predictable from X; an R^2 of 0.20 means that 20 percent is predictable; and so on.

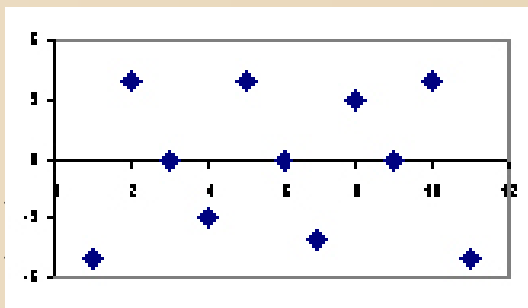


The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis.

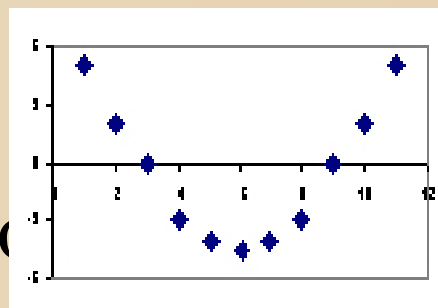
A predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

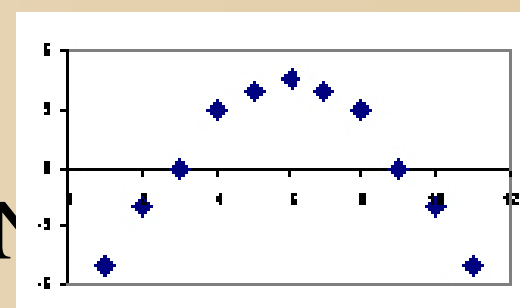
Plot of Residuals



h No



N



Regression Types

Standard linear regression	None	$y = a + bx$	$\hat{y} = a + bx$	Linear transformation
Exponential model	Dependent variable = $\log(y)$	$\log(y) = a + bx$	$\hat{y} = 10^{(a + bx)}$	Nonlinear transformation
Quadratic model	Dependent variable = \sqrt{y}	$\sqrt{y} = a + bx$	$\hat{y} = (a + bx)^2$	Nonlinear transformation
Reciprocal model	Dependent variable = $1/y$	$1/y = a + bx$	$\hat{y} = 1 / (a + bx)$	Nonlinear transformation
Logarithmic model	Independent variable = $\log(x)$	$y = a + b \log(x)$	$\hat{y} = a + b \log(x)$	Nonlinear transformation
Power model	Dependent variable = $\log(y)$ Independent variable = $\log(x)$	$\log(y) = a + b \log(x)$	$\hat{y} = 10^{(a + b \log(x))}$	Nonlinear transformation

How to find which regression is best

- Transforming a data set to enhance linearity is a multi-step, trial-and-error process.
 - Conduct a standard regression analysis on the raw data.
 - Construct a residual plot.
 - If the plot pattern is random, do not transform data.
 - If the plot pattern is not random, continue.
 - Compute the coefficient of determination (R^2).
 - Choose a transformation method (see above table).
 - Transform the independent variable, dependent variable, or both.
 - Conduct a regression analysis, using the transformed variables.
 - Compute the coefficient of determination (R^2), based on the transformed variables.
 - If the transformed R^2 is greater than the raw-score R^2 , the transformation was successful. Congratulations!
 - If not, try a different transformation method.
- The best transformation method (exponential model, quadratic model, reciprocal model, etc.) will depend on nature of the original data. The only way to determine which method is best is to try each and compare the result (i.e., residual plots, correlation coefficients).



END OF REGRESSION





PROBABILITY



Probability

Probability of an event happening =
Number of ways it can happen
Total number of outcomes

Example 1: the chances of rolling a "4" with a die

Number of ways it can happen: 1 (there is only 1 face with a "4" on it)

Total number of outcomes: 6 (there are 6 faces altogether)

So the probability = $1/6$

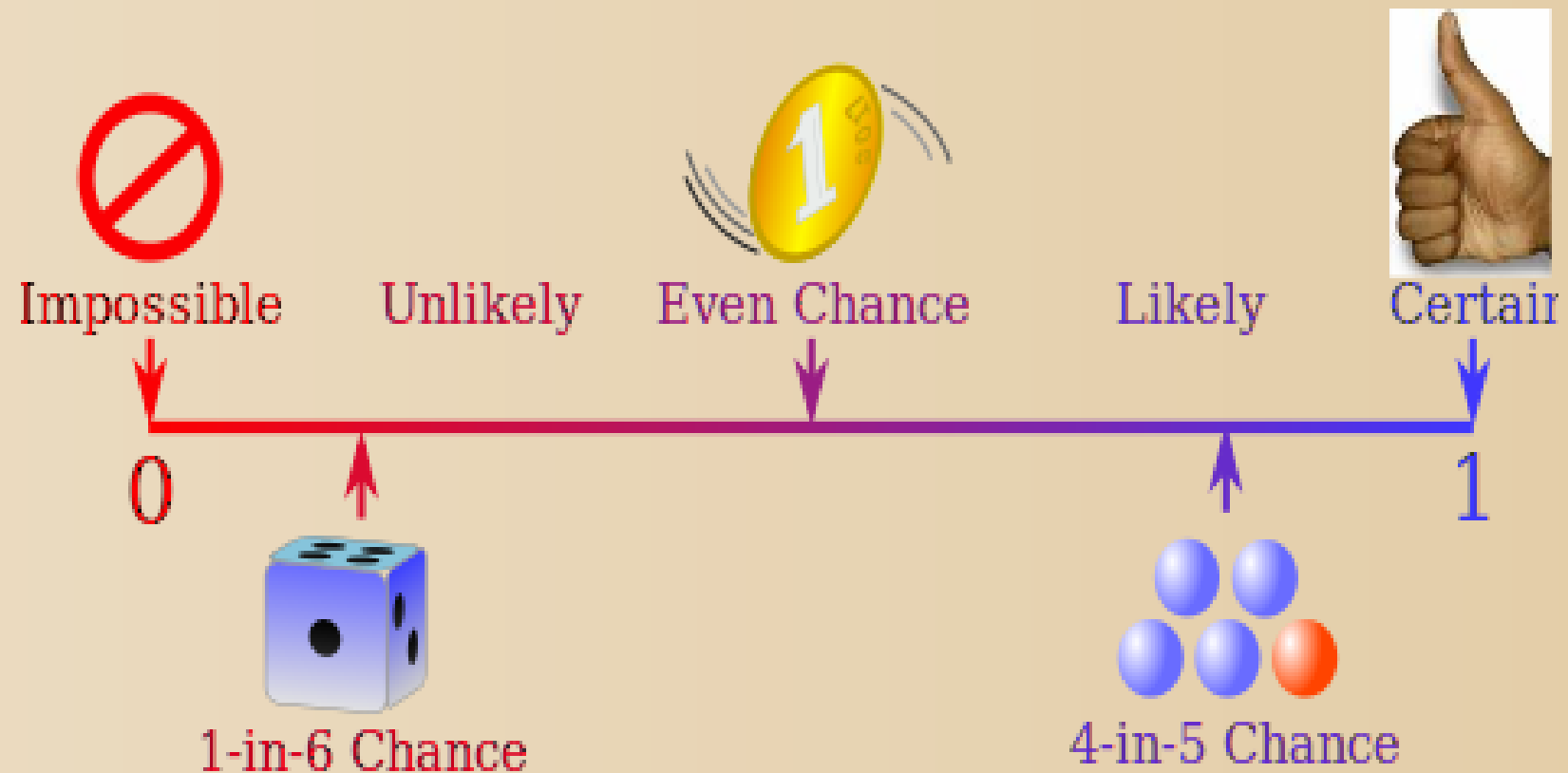
Example 2 : The chances of getting a King from deck of cards.

Number of Kings=4

Number of Cards=52

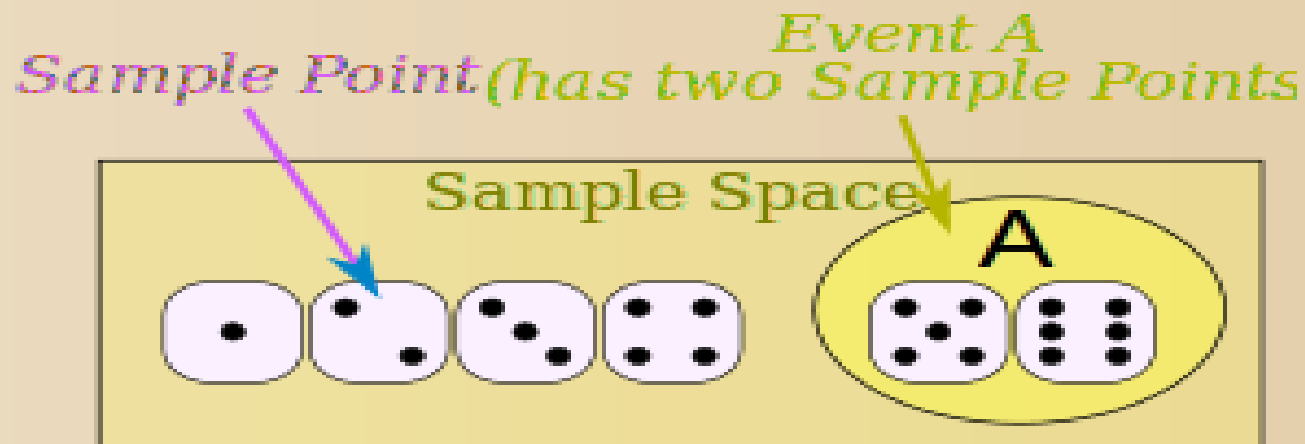
Probability= $4/52=1/13$

Probability line



Definition

- Experiment or Trial: an action where the result is uncertain.
- Sample Space: all the possible outcomes of an experiment
- Sample Point: just one of the possible outcomes
- Event: a single result of an experiment.



Example

Example: Alex wants to see how many times a "double" comes up when throwing 2 dice.

- Each time Alex throws the 2 dice is an Experiment.
- It is an Experiment because the result is uncertain.
- The Event Alex is looking for is a "double", where both dice have the same number. It is made up of these 6 Sample Points:
 - {1,1} {2,2} {3,3} {4,4} {5,5} and {6,6}
- The Sample Space is all possible outcomes (36 Sample Points):
 - {1,1} {1,2} {1,3} {1,4} ... {6,3} {6,4} {6,5} {6,6}
- These are Alex's Results:

Experiment	Is it a Double?
• {3,4}	No
• {5,1}	No
• {2,2}	Yes
• {6,3}	No
•

Types of Event

Complementary Events-The complement of an event is the event not occurring. The probability that Event A will not occur is denoted by $P(A')$.

Mutually Exclusive Events- Two events are mutually exclusive or disjoint if they cannot occur at the same time.

Independent Event- If the occurrence of Event A does not change the probability of Event B, then Events A and B are independent.

Dependent Event-If the occurrence of Event A changes the probability of Event B, then Events A and B are dependent

Theorem of Addition

- When two events are mutually exclusive their joint probability is calculated by theorem of addition.

$$P(A \cup B) = P(A) + P(B)$$

- When two events are overlapping i.e they have some instances in common then joint probability is.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Theorem of Multiplication

Rule of Multiplication :

- Case1 (Independent Event)The probability that Events A and B both occur is equal to the probability that Event A occurs times the probability that Event B occurs, given that A has occurred.

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

- Case 2(Dependent Event)

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

Bayes Theorem

From theorem of multiplication

$$P(A \cap B) = P(A) P(B|A) \text{ ---- (1)}$$

$$P(B|A) = P(A \cap B)/P(A) \text{ ---- (2)}$$

Similarly

$$P(A|B) = P(A \cap B)/P(B) \text{ -----(3)}$$

by 2 and 3

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


$$P(B|A)/P(A|B) = P(B)/P(A)$$

Application of Bayes theorem



So now you know how search engines can guess what you want: they simply keep track of what lots of people type in and what websites they eventually click on. Then using Bayes they figure which ones are probably the best to show first. It makes them look like they can read your mind!

Bayes theorem helps in classifying lot of things like mails, customers, clients, students, weather forecasting and on and on.....





PROBABILITY DISTRIBUTIONS



Discrete Probability Distribution

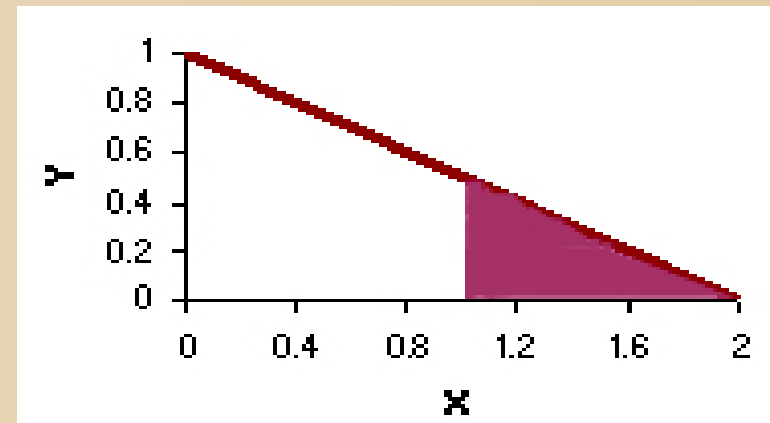
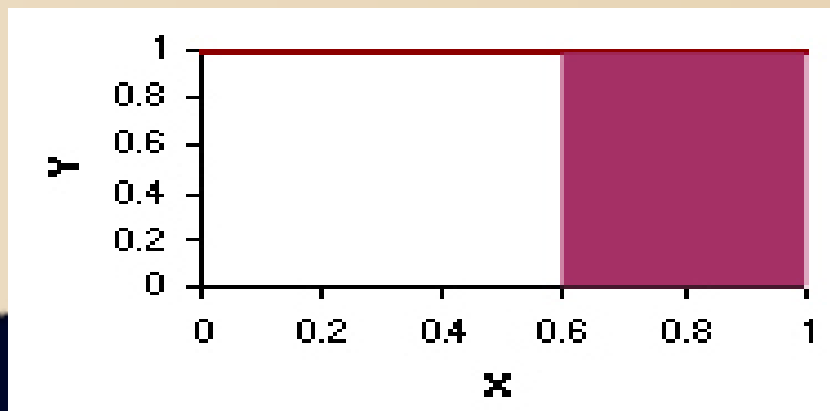
A probability distribution is a table or an equation that links each possible value that a random variable can assume with its probability of occurrence. The following table is an example of a probability distribution for a discrete random variable.

Number of head	Probability
0	0.25
1	0.50
2	0.25

Continuous Probability Distributions

The probability distribution of a continuous random variable is represented by an equation, called the probability density function (pdf). All probability density functions satisfy the following conditions:

- The random variable Y is a function of X ; that is, $y = f(x)$.
- The value of y is greater than or equal to zero for all values of x .
- The total area under the curve of the function is equal to one.
- Left is probability distribution graph for $y=1$ and right is for $y=1-0.5x$



Binomial distribution

A binomial random variable is the number of successes x in n repeated trials of a binomial experiment. The probability distribution of a binomial random variable is called a binomial distribution.

The binomial experiment has four properties.

- The experiment consists of n repeated trials.
- Each trial can result in just two possible outcomes. We call one of these outcomes a success and the other, a failure.
- The probability of success, denoted by P , is the same on every trial.
- The trials are independent; that is, the outcome on one trial does not affect the outcome on other trials.

The binomial distribution has the following properties:

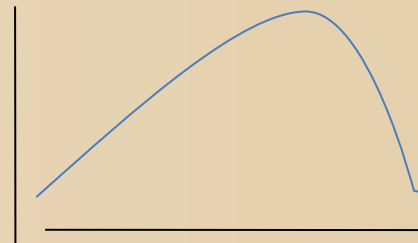
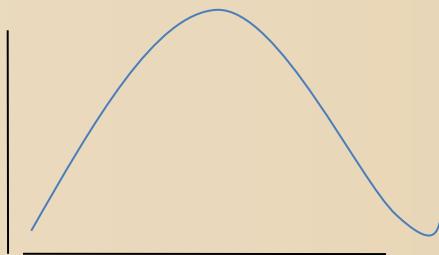
- The mean of the distribution (μ_x) is equal to $n * P$.
- The variance (σ^2_x) is $n * P * (1 - P)$.
- The standard deviation (σ_x) is $\sqrt{n * P * (1 - P)}$.

Normal Distribution

The normal distribution refers to a family of continuous probability distributions described by the normal equation.

Normal Curve

The graph of the normal distribution depends on two factors - the mean and the standard deviation. The mean of the distribution determines the location of the center of the graph, and the standard deviation determines the height and width of the graph. When the standard deviation is large, the curve is short and wide; when the standard deviation is small, the curve is tall and narrow. All normal distributions look like a symmetric, bell-shaped curve, as shown below.



Z-score

Standard Score/ z score

The normal random variable of a standard normal distribution is called a standard score or a z-score.

Every normal random variable X can be transformed into a z score via the following equation:

$$z = (X - \mu) / \sigma$$

Z score > 0 tells that element greater than mean.

Z score $= 0$ tells that element is equal to mean

Z score $= 1$ tells that element is 1 standard deviation from mean.

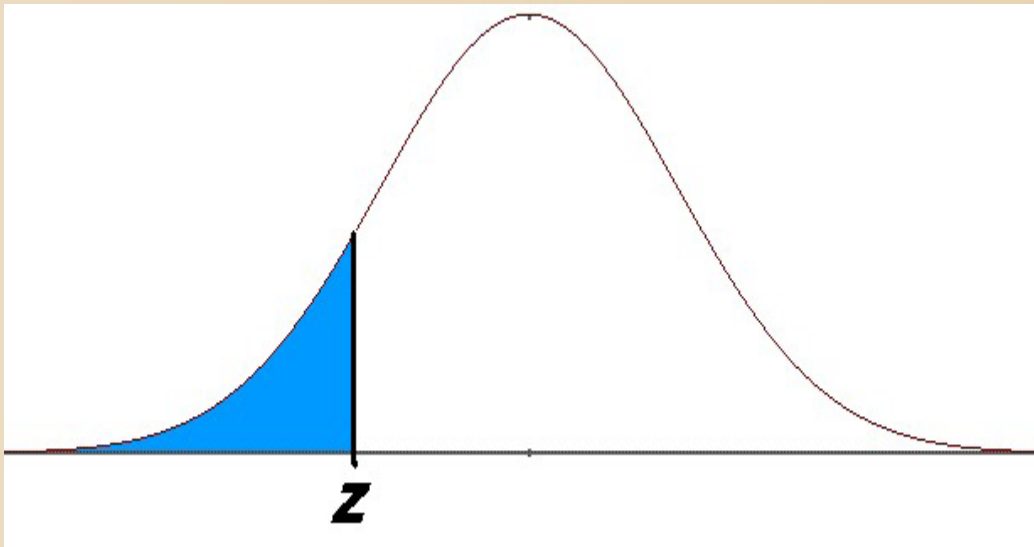
Z-score and Probability

- Once we found the z score, its probability can be calculated by z-score. For example what is the probability of z-score 2.47
- Z-table is used to find probabilities for a statistical sample with a standard normal (Z) distribution.
- Go to the row that represents the ones digit and the first digit after the decimal point (the tenths digit) of your z-value.

z	0.01	0.02	0.03
-1.0	0.1587	0.1563	0.1539
0.0	0.5000	0.496	0.4920
1.0	0.8413	0.8438	0.8461

Finding the Probability Associated with a z-Score

The **z-score table** above provides the area under the standard normal distribution that falls to the left of each particular z value. That is the value shaded in the diagram below. The area can be interpreted as the probability that a score in the distribution is less than the score that corresponds to z .

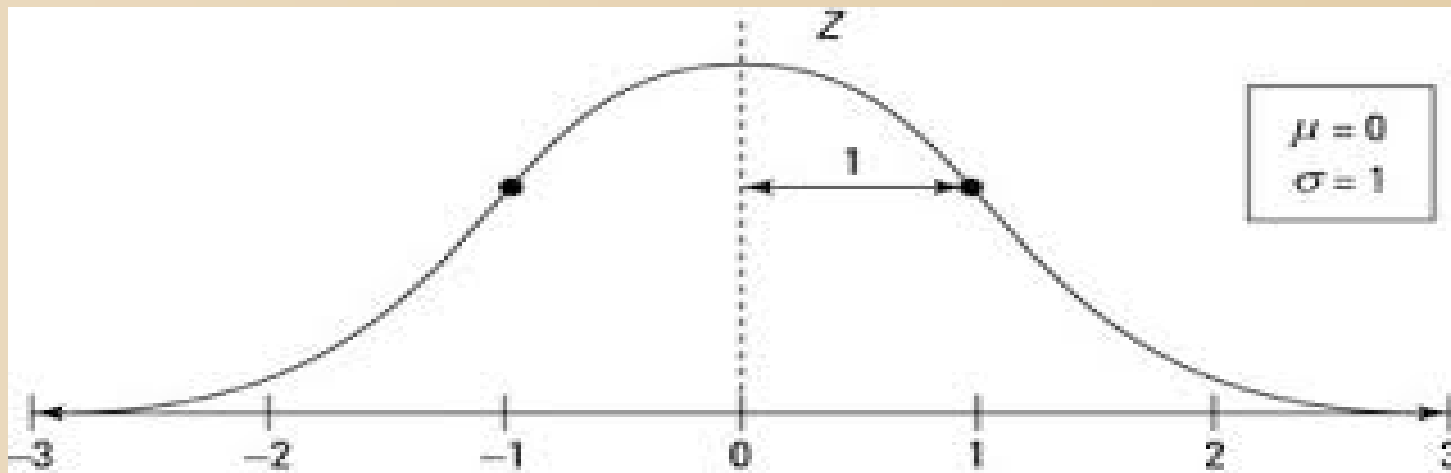


For example, a z -score of zero (remember that is the z -score that corresponds to the mean), has a probability of 0.5 because half of the scores in the normal distribution are lower than the mean.

Z-distribution

In statistics, the Z-distribution is used to help find probabilities and percentiles for regular normal distributions (X). It serves as the standard by which all other normal distributions are measured. The Z-distribution is a normal distribution with mean zero and standard deviation 1; its graph is shown here.

- According to the central limit theorem, the sampling distribution of a statistic (like a sample mean) will follow a normal distribution, as long as the sample size is sufficiently large. Therefore, when we know the standard deviation of the population, we can compute a z-score, and use the normal distribution to evaluate probabilities with the sample mean.



Example

Molly earned a score of 940 on a national achievement test. The mean test score was 850 with a standard deviation of 100. What proportion of students had a lower and higher score than Molly? (Assume that test scores are normally distributed.)

- Find z score

$$z = (X - \mu) / \sigma = (940 - 850) / 100 = 0.90$$

- From z table, $P(Z < 0.90) = 0.8159$.

- 81% students scored less than Molly.

- So $P(Z > 0.90) = 1 - P(Z < 0.90) = 1 - 0.8159 = 0.1841$.

- 18% student scored more than Molly

t-score

When sample sizes are sometimes small, and often we do not know the standard deviation of the population. When either of these problems occur, statisticians rely on the distribution of the t statistic (also known as the t score), whose values are given by:

$$t = [x - \mu] / [s / \text{sqrt}(n)]$$

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size. The distribution of the t statistic is called the t distribution or the Student t distribution.

- The t distribution allows us to conduct statistical analyses on certain data sets that are not appropriate for analysis, using the normal distribution.

Properties of t-score

- The t distribution has the following properties:

- The mean of the distribution is equal to 0 .
- The variance is equal to $v / (v - 2)$, where v is the degrees of freedom (see last section) and $v > 2$.
- The variance is always greater than 1, although it is close to 1 when there are many degrees of freedom. With infinite degrees of freedom, the t distribution is the same as the standard normal distribution.

When to Use the t Distribution

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). The sampling distribution of a statistic should be bell-shaped if any of the following conditions apply.

- The population distribution is normal.
- The population distribution is symmetric, unimodal, without outliers, and the sample size is at least 30.
- The population distribution is moderately skewed, unimodal, without outliers, and the sample size is at least 40.
- The sample size is greater than 40, without outliers.

t-table


T-table is a matrix of degree of freedom and percentile score.

F-Statistics



An F statistic is a value you get when you run an ANOVA test or a regression analysis to find out if the means between two populations are significantly different.

In an ANOVA, the F-ratio is the statistic used to test the hypothesis that the effects are real: in other words, that the means are significantly different from one another.



F-Test



An F-test is any statistical test in which the test statistic has an F-distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled.



F-Value

The F value is the ratio of the mean regression sum of squares divided by the mean error sum of squares. Its value will range from zero to an arbitrarily large number. The value of $\text{Prob}(F)$ is the probability that the null hypothesis for the full model is true (i.e., that all of the regression coefficients are zero).

A high F value means that your data does not well support your null hypothesis. Or in other words, the alternative hypothesis is compatible with observed data. In regression there are typically two types of F values.

Confidence Interval of sample mean

Z -Statistic

Case1 :When we select a sample from population and we know the SD of population then we can calculate the confidence interval of mean or how reliable is the mean.

$$(\bar{x} - z(\alpha/2) * SE) < \mu < (\bar{x} + z(\alpha/2) * SE)$$

- where \bar{x} is the sample mean
- SE is the standard error = σ/\sqrt{n}
- Level of confidence is $1-\alpha$ (For 99% , $\alpha=.01$)

Case2: When population SD is not known then confidence interval is calculated as follows we replace SE by sample standard deviation/ \sqrt{n}

$$(\bar{x} - z(\alpha/2) * s/\sqrt{n}) < \mu < (\bar{x} + z(\alpha/2) * s/\sqrt{n})$$

Example of z statistic

A sample of 100 people is selected and mean time spent watching TV is 32. SD of TV watching time is 12 hours. Find the range of 95 percent confidence interval.

SD=12, $n=100$, $SE=12/\sqrt{100}=1.2$

for 95%, $\alpha = 0.05/2 = 0.025$

$z(\alpha/2)=z(0.025) = 1.96$ (from z value and confidence table (standard table))

$32(+/-)1.96*1.2 = 29.648 \text{ to } 34.352$

t-Statistics

Case3: when these three condition are satisfied we use t-statistics to find the confidence interval.

1. When population is normally distributed
2. Population standard deviation is not known.
3. Sample is small in size. $N < 30$

$$t = (\bar{x} - \mu) / (s / \sqrt{n})$$

where \bar{x} = sample mean

μ = population mean

s = standard deviation of sample

n = sample size

STATISTICAL INFERENCE



Statistical Inference

One of the major applications of statistics is estimating population parameters from sample statistics.

In statistics, estimation refers to the process by which one makes inferences about a population, based on information obtained from a sample.

For example, sample means are used to estimate population means; sample proportions, to estimate population proportions. An estimate of a population parameter may be expressed in two ways:

Point estimate. A point estimate of a population parameter is a single value of a statistic. For example, the sample mean \bar{x} is a point estimate of the population mean μ . Similarly, the sample proportion p is a point estimate of the population proportion P . Point estimates are usually supplemented by interval estimates called confidence intervals.

Interval estimate. An interval estimate is defined by two numbers, between which a population parameter is said to lie. For example, $a < x < b$ is an interval estimate of the population mean μ . It indicates that the population mean is greater than a but less than b .

Confidence Interval

Statisticians use a **confidence interval** to describe the amount of uncertainty associated with a sample estimate of a **population parameter**.

Confidence Interval Data Requirements

To express a confidence interval, you need three pieces of information.

- Confidence level

- Statistic

- Margin of error

Given these inputs, the range of the confidence interval is defined by the sample statistic + margin of error. And the uncertainty associated with the confidence interval is specified by the confidence level.

How to Interpret Confidence Intervals



Suppose that a 90% confidence interval states that the population mean is greater than 100 and less than 200. How would you interpret this statement?

Some people think this means there is a 90% chance that the population mean falls between 100 and 200. This is incorrect.

The confidence level describes the uncertainty associated with a sampling method. Suppose we used the same sampling method to select different samples and to compute a different interval estimate for each sample.

Some interval estimates would include the true population parameter and some would not. A 90% confidence level means that we would expect 90% of the interval estimates to include the population parameter. A 95% confidence level means that 95% of the intervals would include the parameter; and so on.



How to calculate Confidence Interval

- 1) Identify a sample statistic. Choose the statistic (e.g, sample mean, sample proportion,) that you will use to estimate a population parameter.
- 2) Select a confidence level. As we noted in the previous section, the confidence level describes the uncertainty of a sampling method. Often, researchers choose 90%, 95%, or 99% confidence levels; but any percentage can be used.
- 3) Find the margin of error. If you are working on a homework problem or a test question, the margin of error may be given. Often, however, you will need to compute the margin of error, based on one of the following equations.
$$\text{Margin of error} = \text{Critical value} * \text{Standard deviation of statistic}$$
$$\text{Margin of error} = \text{Critical value} * \text{Standard error of statistic}$$

(Critical value is calculated through t-statistics, z-statistics)
- 4) Specify the confidence interval. The uncertainty is denoted by the confidence level. And the range of the confidence interval is defined by the following equation.
$$\text{Confidence interval} = \text{sample statistic} + \text{Margin of error}$$

Example

A random sample of 1,000 men from a population of 1,000,000 men and weigh them. We find that the average man in our sample weighs 180 pounds, and the standard deviation of the sample is 30 pounds. What is the 95% confidence interval.

1. Identify a sample statistic. Since we are trying to estimate the mean weight in the population, we choose the mean weight in our sample (180) as the sample statistic.
2. Select a confidence level. In this case, the confidence level is defined for us in the problem. We are working with a 95% confidence level.

Example-continued

Find the margin of error = SE * Critical value.

2.1 Find standard error. The SE of the mean is

$$SE = s / \sqrt{n} = 30 / \sqrt{1000} = 30/31.62 = 0.95$$

2.2 Critical value can be expressed as t-score or z-score. Here we use t-score

Compute alpha (α): $\alpha = 1 - (\text{confidence level} / 100) = 0.05$

Find the critical probability (p^*): $p^* = 1 - \alpha/2 = 1 - 0.05/2 = 0.975$

Find the degrees of freedom (df): $df = n - 1 = 1000 - 1 = 999$

The critical value is the t statistic having 999 degrees of freedom and a cumulative probability equal to 0.975. From the t Distribution Calculator, we find that the critical value is 1.96.

2.3 margin of error (ME): $ME = \text{critical value} * \text{standard error} = 1.96 * 0.95 = 1.86$

Example-continued

Specify the confidence interval. The range of the confidence interval is defined by the sample statistic + margin of error.

And the uncertainty is denoted by the confidence level. Therefore, this 95% confidence interval is 180 ± 1.86 .

Test of Hypothesis

Very often we make decision about population based on sample information. Such decision are called statistical decision.

Example we may wish to decide on the basis of sample data whether a new serum is really effective in curing a disease or we test sample from production to consider it good or bad.

- Procedure which help us to decide to accept or reject a hypothesis is called test of hypothesis or tests of significance.

NULL HYPOTHESIS



- Research Hypothesis- The hypothesis which determines the research is called research hypothesis or alternate hypothesis.

- Null Hypothesis- The negation of research Hypothesis is called Null Hypothesis.

The purpose of research is to nullify the research hypothesis and prove the research hypothesis.

Example:

H₀- World is flat

H₁- World is not flat.




Example


- $H_0: \mu = \$155$ (the mean sales per order this year is \$155)
- $H_a: \mu \neq \$155$ (the mean sales per order this year is not \$155)
- To test this hypothesis, a sample of 100 orders for the current year are selected and the mean of the sample is used to decide whether to reject the null hypothesis or not. A statistic, which is used to decide whether to reject the null hypothesis or not is called a test statistic

Example

If the mean of the sample, $\text{mean}(x)$, is “close” to \$155, then we would likely not reject the null hypothesis. However, if the computed value of $\text{mean}(x)$ is considerably different from \$155, we would reject the null hypothesis since this outcome supports the truth of the research hypothesis. The critical decision is how different does X need to be from \$155 in order to reject the null hypothesis.



•So in layman terms we need to decide how much variation from 155 is needed to reject the null hypothesis. Let say the sample mean comes out to be between 145 and 165 (2 standard error from mean, $SE=SD/\sqrt{n}=50/\sqrt{100}$) then we do not reject the null hypothesis. Beyond this we will reject the null hypothesis.



Significance Level

conclusion	H0 True	H0 False
Accept H0	Correct Conclusion	Type 2 error
Reject Ho	Type 1 error	Correct Conclusion

Now we calculate the probability of making error.

α = Probability of making type1 error or
 $\alpha = P(\text{rejecting null hypothesis when it is true})$

β = Probability of making type2 error
 $\beta = P(\text{not rejecting null hypothesis when it is false})$

Example

Let us assume sample mean comes out to be much higher than 155 (population mean = 155) we reject the null hypothesis and alternate hypothesis is accepted. This is the case of type1 error.

Let us assume sample mean comes out to be almost 155 (population mean > 155) we do not reject the null hypothesis and alternate hypothesis is rejected. This is the case of type2 error.

Significance level/ α

• In testing a given hypothesis, we calculate the probability of the outcome of hypothesis. If probability is low we conclude that null hypothesis is wrong. But how low should be this probability value. Some researcher say that reject the null hypothesis only if probability is less than 0.05 and some set to 0.01. The probability value below which null hypothesis is rejected is called significance level or α (alpha) level or simply α . In other terms we can say it is the probability of type 1 error given null hypothesis is true. As per Pearson this probability is often specified before any samples are drawn so that results obtained will not influence our decision.

Significance testing/ β

The second type of error that can be made in significance testing is failing to reject a false null hypothesis. It happens when data does not provide strong evidence that null hypothesis is false.

A Type II error can only occur if the null hypothesis is false. If the null hypothesis is false, then the probability of a Type II error is called β (beta).

One tailed/two tailed

A test of a statistical hypothesis, where the region of rejection is on only one side of the sampling distribution, is called a one-tailed test. For example, suppose the null hypothesis states that the mean is less than or equal to 10. The alternative hypothesis would be that the mean is greater than 10. The region of rejection would consist of a range of numbers located on the right side of sampling distribution; that is, a set of numbers greater than 10.

A test of a statistical hypothesis, where the region of rejection is on both sides of the sampling distribution, is called a two-tailed test. For example, suppose the

Steps in Hypothesis testing

1. The first step is to specify the null hypothesis. For a one-tailed test, the null hypothesis is either that a parameter is greater than or equal to zero. For a two-tailed test, the null hypothesis is typically that a parameter equals zero.

The second step is to specify the α level which is also known as the significance level. Typical values are 0.05 and 0.01.

The third step is to compute the probability value (also known as the p value).

Finally, compare the probability value with the α level

Selecting the test statistic

In most cases we calculate the z-statistic for the hypothesis testing. From z-distribution we calculate the probability of z value. If it lies in acceptance region according to significance level then we accept null hypothesis otherwise we reject it.

Example: It is hypothesized for a sample with $n=64$ of normal distribution that, $\sigma(\text{SD})=16, \mu(\text{Mean})=80$.

Now test this hypothesis at $\alpha = 0.05$ if mean value is supposed to be 82

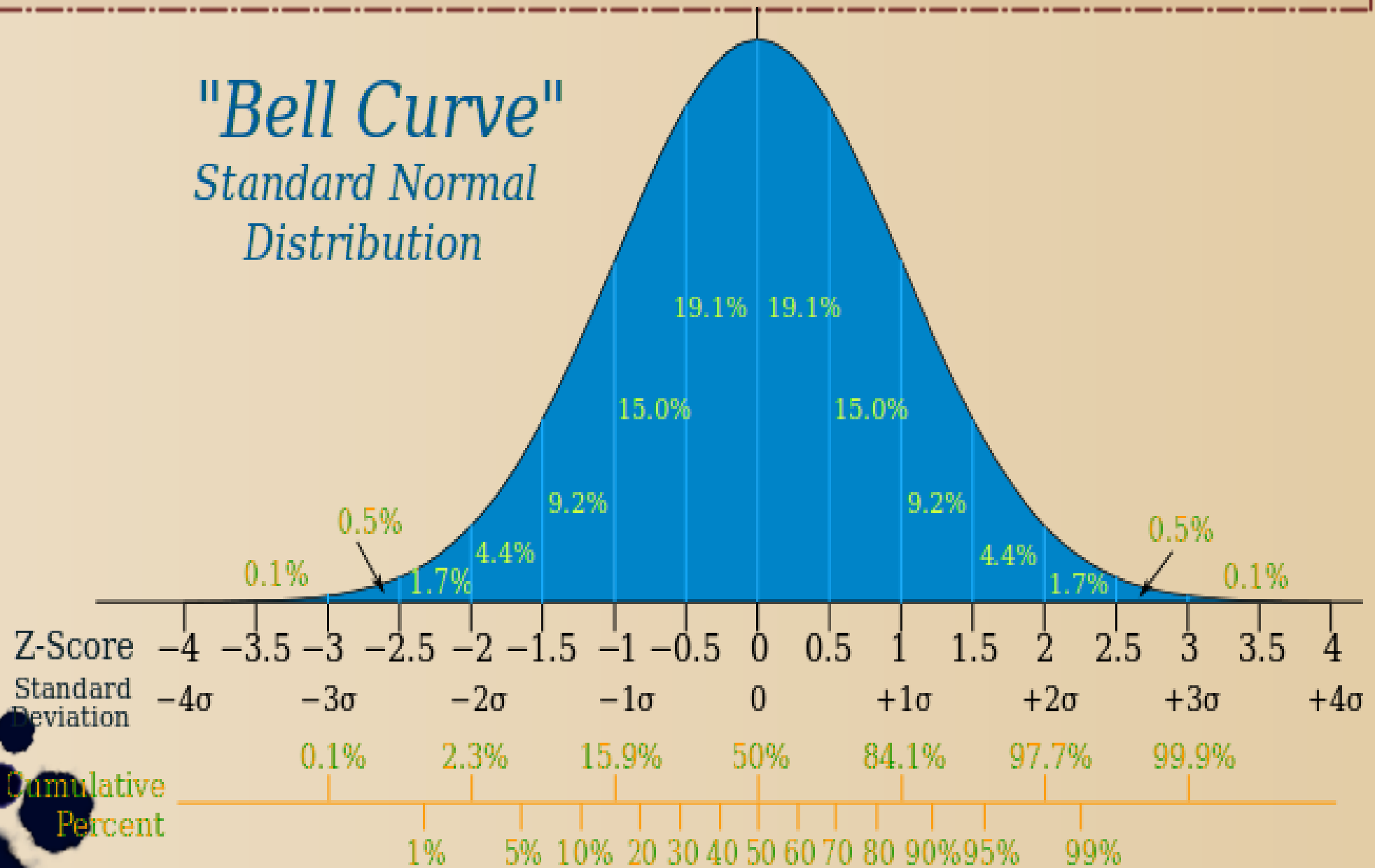
Solution :

- H_0 = Mean is 80.
- H_1 = mean is not 80.

First we apply z statistic

Normal Distribution

"Bell Curve"
Standard Normal
Distribution



Why we standardize

- Example: Professor Willoughby is marking a test.
- Here are the students results (out of 60 points):
- 20, 15, 26, 32, 18, 28, 35, 14, 26, 22, 17
- Most students didn't even get 30 out of 60, and most will fail.
- The test must have been really hard, so the Prof decides to Standardize all the scores and only fail people 1 standard deviation below the mean.
- The Mean is 23, and the Standard Deviation is 6.6,

Symbols

- Ω - Sigma
- Π - Pi
- α = alpha
- β = Beta
- \sqrt{X} = underoot
- \bar{X} = mean of x
- σ_{μ} = Sigma
-