Statistics

Data

- The collected observations we have about something.
- It can be continuous e.g. stock price or categorical i.e., car having best mileage.
- It helps us in understanding the things as they are or the relationships between any two events.
- Helps us to predict future behaviour.
- Textual and numerical data is hard to read or understand but visualising data by plotting charts or trends helps us in analysing the trends much faster.

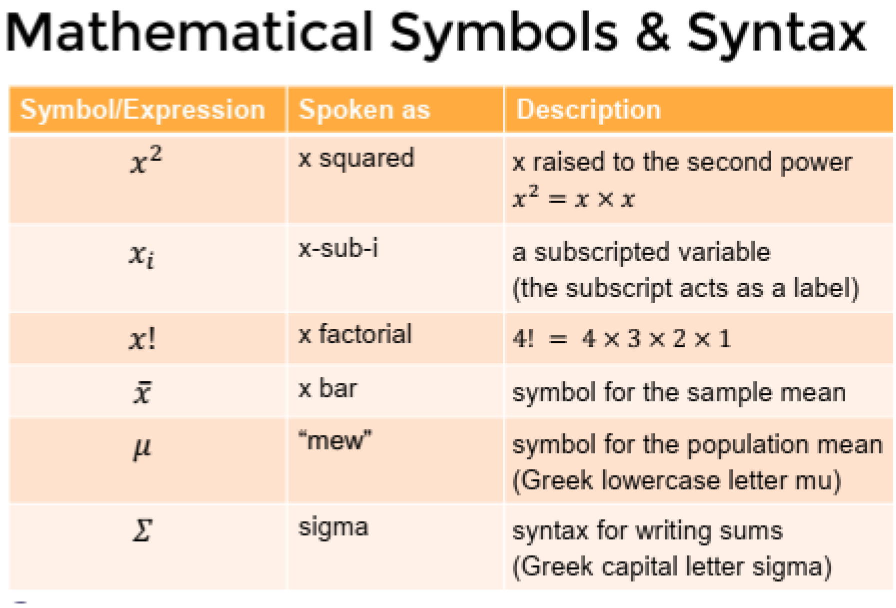
Measurement of data

- There are four levels of measurement.
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Nominal data is predefined categories which can’t be sorted.
  - Gender
  - Political parties
  - Animal classification
- Ordinal data can be sorted but lacks scale.
  - Survey response like often, sometimes, seldom or never.
- Interval data provides scale but lacks zero point.
  - Temperature in Celsius.
- Ratio data have a true zero point
  - Age, weight, salary

Population vs Sample

- Population represents every member of the group
- Sample represents a subset of the members that time and resources allow us to measure.

Mathematical Symbols



Mathematical Symbols & Syntax

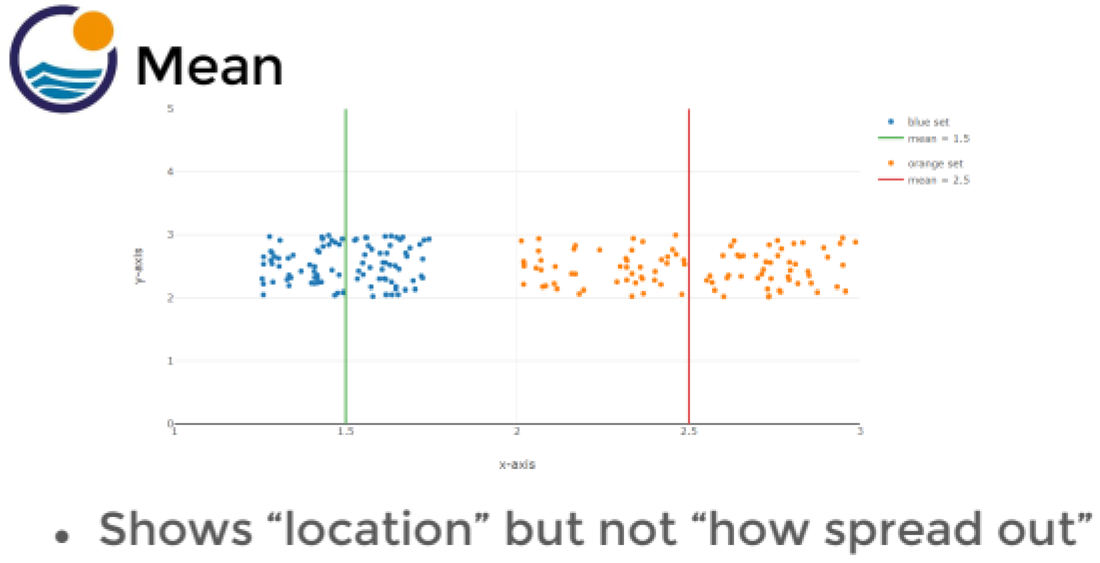
Symbol/Expression	Spoken as	Description
x^2	x squared	x raised to the second power $x^2 = x \times x$
x_i	x-sub-i	a subscripted variable (the subscript acts as a label)
$x!$	x factorial	$4! = 4 \times 3 \times 2 \times 1$
$\bar{x}$	x bar	symbol for the sample mean
μ	"mew"	symbol for the population mean (Greek lowercase letter mu)
Σ	sigma	syntax for writing sums (Greek capital letter sigma)

Sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Measurement of Central Tendency

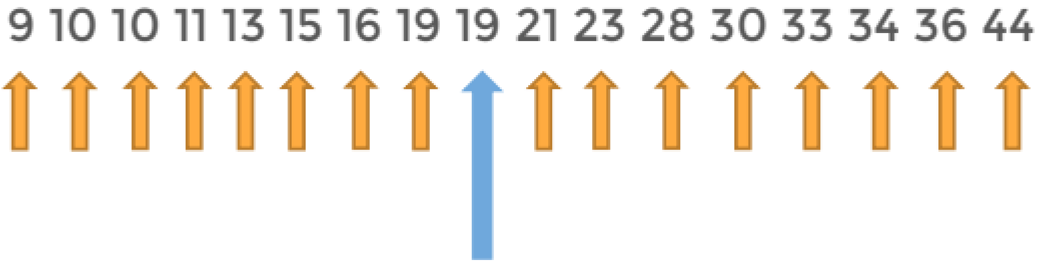
- Describe the location of the data
- Fail to describe the shape of the data
- Mean = Calculated average
- Median = Middle value
- Mode = Most occurring value



Mean

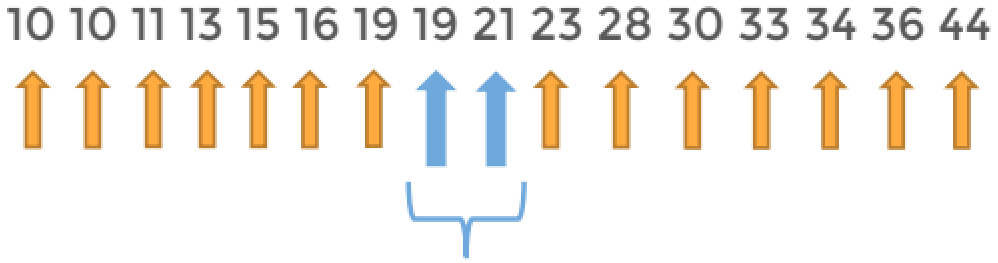
- Shows "location" but not "how spread out"

Median – *odd number of values*


9 10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44

$= 19$

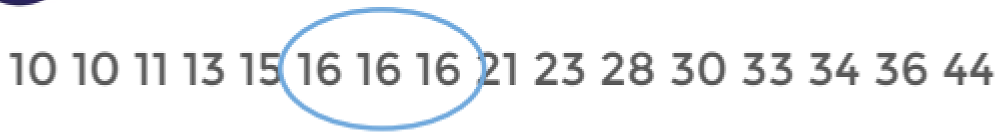
Median - *even number of values*


10 10 11 13 15 16 19 19 21 23 28 30 33 34 36 44

$$\frac{19+21}{2} = 20$$

- While calculating the median, it is important to sort the data in the ascending order.
- Mean can be influenced by the outliers whereas mode and median won’t.

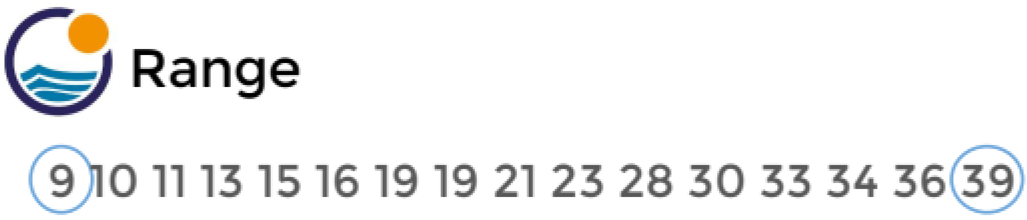
Mode


10 10 11 13 15 16 16 16 21 23 28 30 33 34 36 44

$= 16$

Measurement of Dispersion

- Will describe how spread out the data is.



Range

9 10 11 13 15 16 19 19 21 23 28 30 33 34 36 39

$$\begin{aligned} Range &= max - min \\ &= 39\ - 9 \\ &= 30 \end{aligned}$$

- **Variance**
  - Calculated as the sum of square distances from each point of the mean.
  - There is difference between the sample variance and the population variance as sample variance is subject to Bessel's correction (n-1)
  - Problem with variance is that it is square of the units of the measurement which is why standard deviation is used.

SAMPLE VARIANCE: $s^2 = \frac{\Sigma(x-\bar{x})^2}{n-1}$

POPULATION VARIANCE: $\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}$

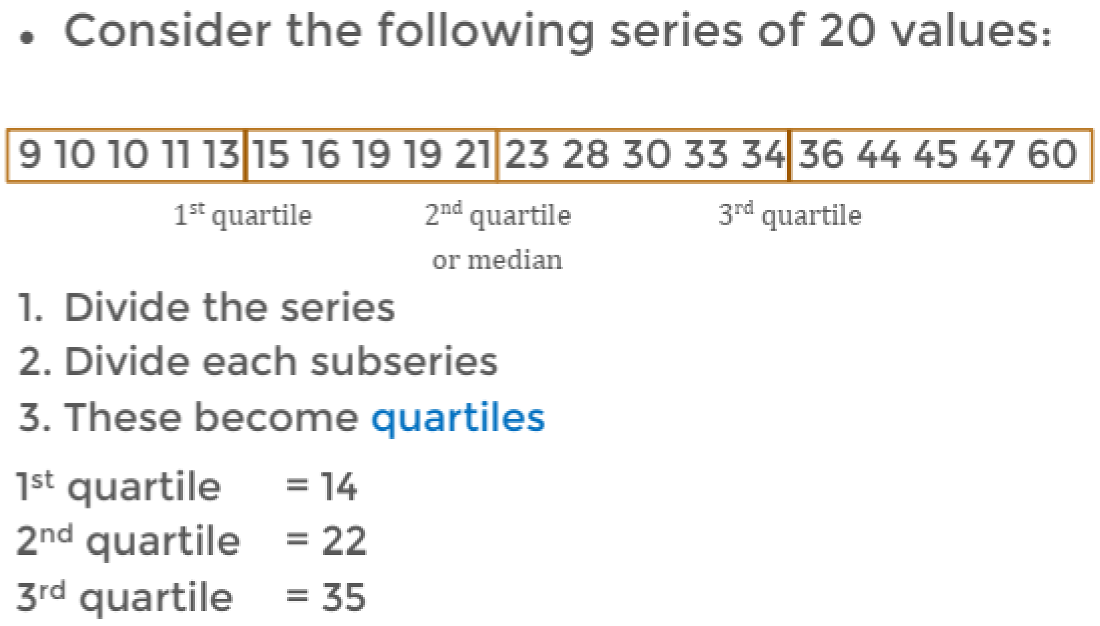
- **Standard Deviation**
  - Square root of the variance.
  - Benefit: Same unit as that of the sample
  - Meaningful to about values that lie within one standard deviation of the mean.

Sample Standard Deviation $\boxed{s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}}$

Population Standard Deviation $\boxed{\sigma = \sqrt{\frac{\Sigma(X-\mu)^2}{N}}}$

Measurement Types: Quartiles

- Has the advantage that every data point is considered, not aggregated.
- Sort the data in ascending order.



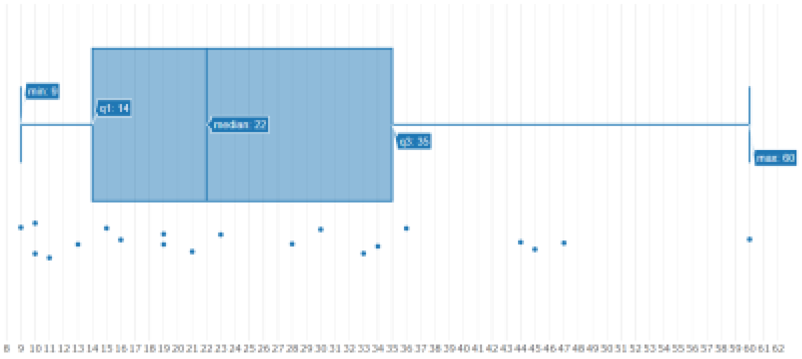
- Consider the following series of 20 values:

9 10 10 11 13 | 15 16 19 19 21 | 23 28 30 33 34 | 36 44 45 47 60

1st quartile　　2nd quartile or median　　3rd quartile

1. Divide the series
2. Divide each subseries
3. These become quartiles

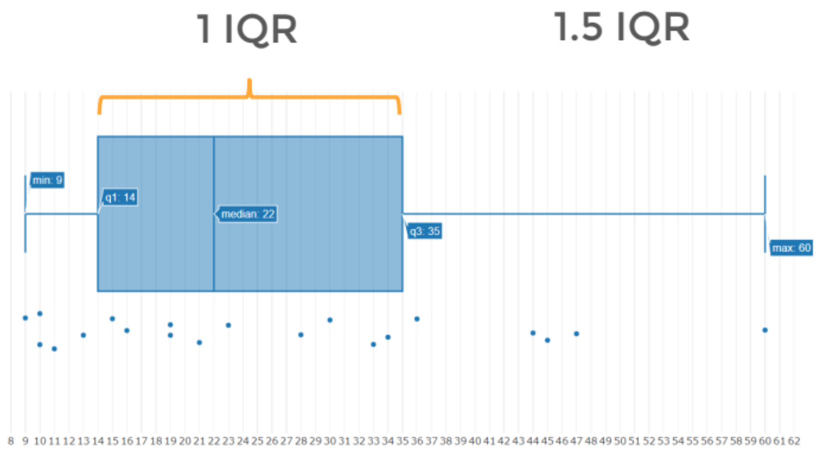
1st quartile = 14
2nd quartile = 22
3rd quartile = 35



Quartile ranges are seldom the same size!

- Box plot helps in visualising the dispersion in our dataset.
- Box plot shows the actual distribution of our data.
- Box represents the middle half of our data.
- Whiskers extend from the box to the maximum and minimum value.
- To consider an outlier we set ‘fence’ which is 1.5 times the width of the IQR.
    - Anything outside the fence is an outlier.
    - Outliers are determined by the data not some arbitrary percentage.

- Inter quartile range is the distance between quartile one and quartile 3.
- While taking a business decision, we can discard the outliers or we can examine them more closely.



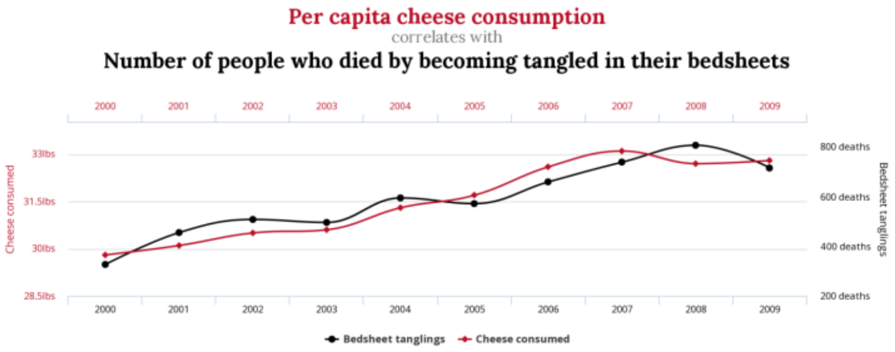
In this set, 60 is *not* an outlier, but 70 would be

Box plot contains 50% of the total data

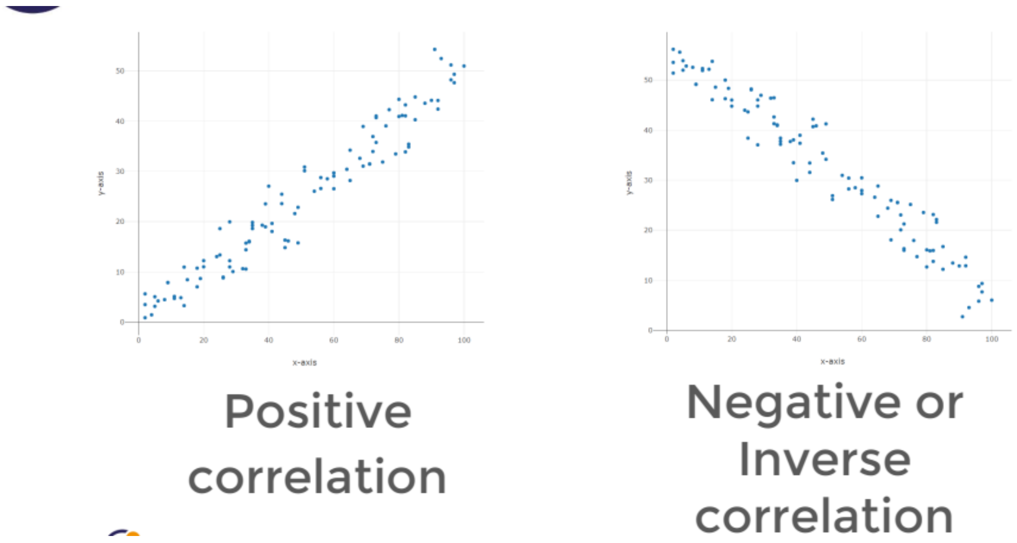
Bivariate Data

- Compares two data
- By convention, x-axis is set to the independent variable and y-axis is set as dependent variable.
- Scatter plot may uncover the correlation between the two variables but they can't show causality.

- **Correlation** between two variables
- Doesn't prove **causality**!



- More statistical analysis is required to determine the causality.



Positive correlation

Negative or Inverse correlation

- Two variables are compared with the help of their variances, but to do that, we have to match their scale i.e., we can't weight in kgs. to height in inches.
    - To do that we have to develop a standard score to normalise the score.
    - We consider population co-variance.

$$cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

- When given a dataset
    - Plot it
    - Calculate mean for both variables
    - Calculate (x -x̅) and (y -y̅).
    - Calculate (x -x̅)*(y -y̅)
    - Calculate sum of (x -x̅)*(y -y̅)
    - Calculate covariance
        - Positive value shows the positive co relation between these two variables and negative value shows the negative co relation.
        - The further the data from the zero, more it represents some sort of relationship between these two variables and closer to zero, represents the more scattering the nature of the data.

Pearson Correlation coefficient

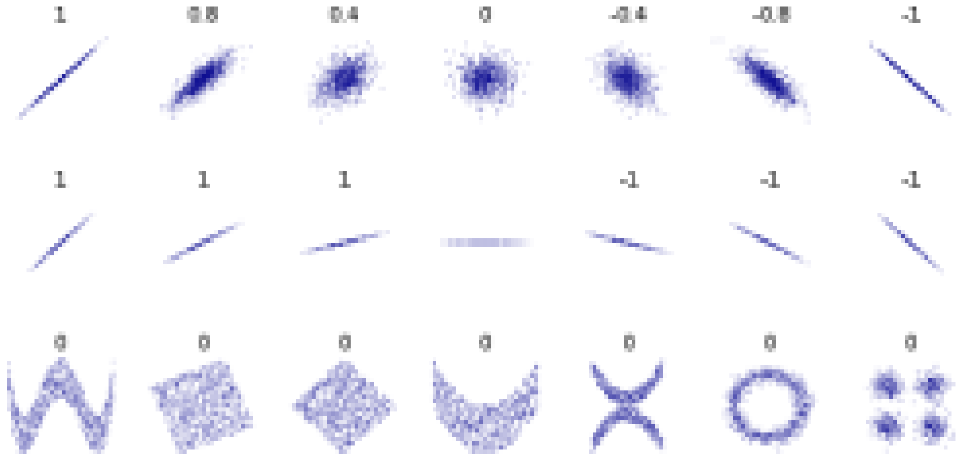
- In order to normalise values coming from two different distributions, we use:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\frac{1}{n}\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\frac{\Sigma(x-\bar{x})^2}{n}}\sqrt{\frac{\Sigma(y-\bar{y})^2}{n}}} = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\sqrt{\Sigma(x-\bar{x})^2}\sqrt{\Sigma(y-\bar{y})^2}}$$

$\rho =$ Greek letter "rho" $\quad \sigma =$ standard deviation

$cov =$ covariance $\quad \bar{x} =$ mean of X

- Values fall between -1 to 1 where 1 represents total positive linear co-relation, 0 no linear co-relation and -1 to total negative linear correlation



- To calculate the coefficient:
  - Calculate the mean of x and y
  - Calculate $(x - \bar{x})$ and $(y - \bar{y})$.
  - Calculate $(x - \bar{x})*(y - \bar{y})$
  - Calculate $(x - \bar{x})^2$ and $(y - \bar{y})^2$.
  - Compute the sum of $(x - \bar{x})*(y - \bar{y})$, $(x - \bar{x})^2$ and $(y - \bar{y})^2$.
  - Put values in the formula and calculate the coefficient.

Probability

- Probability is a value between 0 and 1, where 0 represents the impossibility of event happening and 1 represents the certainty of occurrence of an event
- The act of conducting an experiment i.e., act of flipping a coin is called a trial.
- For flipping of a coin, each trial is independent of the other.
- Each trial can be called an experiment.
  - Consider rolling a dice where each roll is called as an experiment.
- Each mutually exclusive outcome is called a simple event.
  - The probability that a fair dice will roll a six is a simple event.
- The sample space is the sum of every possible simple event.
  - There are six possible outcomes, so sample space has six possible outcomes

Permutations

- A permutation of a set of objects is an arrangement of the objects in a certain order.
  - Here order is important.
- The number of permutations of a set of n objects taken r at a time without repetition is given by the following formula:

$$_{n}P_{r} = \frac{n!}{(n-r)!}$$

- The number of arrangements of n objects taken r at a time, with repetition is given by:

$$n^{r}$$

Combinations

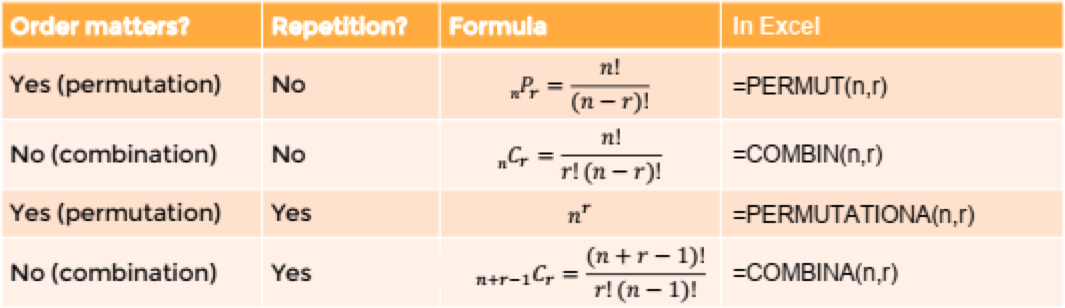
- Unordered arrangements of objects are called combinations.
  - A group of people selected for a team are the same group, no matter the order.
- The number of combinations of a set of n objects taken r at a time is given by:

$$_{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

- The number of combinations taken r at a time from a set n and following the repetition:

$$_{n+r-1}C_{r} = \frac{(n+r-1)!}{r!\,(n-1)!}$$

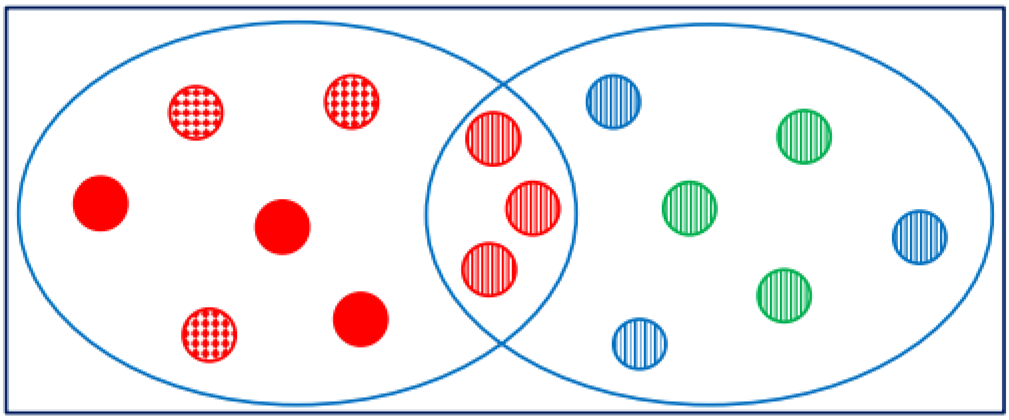
- Permutations and combinations in Excel:



Order matters?	Repetition?	Formula	In Excel
Yes (permutation)	No	$_{n}P_{r} = \frac{n!}{(n-r)!}$	=PERMUT(n,r)
No (combination)	No	$_{n}C_{r} = \frac{n!}{r!\,(n-r)!}$	=COMBIN(n,r)
Yes (permutation)	Yes	n^{r}	=PERMUTATIONA(n,r)
No (combination)	Yes	$_{n+r-1}C_{r} = \frac{(n+r-1)!}{r!\,(n-1)!}$	=COMBINA(n,r)

Intersections, Unions and Complements

- In probability, an intersection describes the sample space where two events both occur.
  - 3 of the balls are both red and striped:



- The intersection of event A and B is given as $A \cap B$.
  - Note that order doesn't matter:

$$A \cap B = B \cap A$$

- The probability pf A and B is given as

$$P(A \cap B)$$

- In the above case the probability of getting both red and striped balls is

$$P(A \cap B) = \frac{3}{15} = \mathbf{0.2}$$

- The union of two events considers if A or B occurs, and is given by $A \cup B$
  - Note that order doesn't matter:

$$A \cup B = B \cup A$$

- The probability of A or B is given as:

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

- In the above case

$$P(A \cup B) = \frac{9}{15} + \frac{9}{15} - \frac{3}{15} = \frac{15}{15} = \mathbf{1.0}$$

- The complement of an event considers everything outside of the event given by $\bar{A}$.
- The probability of not A is:

$$P(\bar{A}) = 1 - P(A) = \frac{15}{15} - \frac{9}{15} = \frac{6}{15} = \mathbf{0.4}$$

Independent and Dependent Events

- An independent series of events occur when the outcome of one event has no effect on the outcome of the other.
  - Flipping a fair coin
  - Rolling a fair dice.
- The probability of seeing two heads with two flips of a fair coin is

$$P(H_1H_2) = P(H_1) \times P(H_2)$$
$$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

- A dependent event occurs when the outcome of a first event does affect the probability of a second event.
  - Draw coloured marbles from a bag without replacement.
  - Drawing two red balls from a bag that contains 2 blue marbles and 3 red marbles.
    - The probability of drawing a first red marble is

$$P(R_1) = \frac{3}{5}$$

    - The probability of drawing a second red marble given that the first marble was red is

$$P(R_2|R_1) = \frac{2}{4}$$

- So, the probability of two red marbles is

$$P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1)$$

$$= \frac{3}{5} \times \frac{2}{4} = \frac{6}{20} = \mathbf{0.3}$$

Conditional Probability

- The idea that we want to know the probability of an event A, given that event B has occurred is called conditional probability and is written as $P(A|B)$.
  - The probability of drawing two red marbles is

    $$P(R_1 \cap R_2) = P(R_1) \cdot P(R_2|R_1)$$

  - The conditional probability in this equation is $P(R_2|R_1)$.
- Rearranging the formula

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

  - Where the probability of **A given B** equals the probability of **A and B** divided by probability of **B.**

Addition and Multiplication rules

- Addition rule

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

- Multiplication rule

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

  - Probability of drawing for aces in a deck of 52 cards is

$$P(A \cap B \cap C \cap D) = P(A) \cdot P(B|A) \cdot P(C|AB) \cdot P(D|ABC)$$

$$= \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} \times \frac{1}{49} = \frac{24}{6{,}497{,}400} = \frac{1}{270{,}725}$$

Bayes Theorem

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \; provided\ that\ P(A), P(B) > 0$$

- It is used to determine the probability of a parameter, given a certain event.

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

- A company learns that 1 out of 500 their products are defective. The company buys a diagnostic tool that correctly identifies a defective part 99% of the time. If the part is diagnosed as defective, what is the probability that it is really defective?

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|-A) \cdot P(-A)}$$

$$= \frac{0.99 \times 0.002}{0.99 \times 0.002 + 0.01 \times 0.998}$$

$$= 0.165$$

  - ○ What if we perform a second test, by that we mean we perform the diagnostic test again on the same product that was tested earlier, and that also comes up positive?

$$P(A|B) = \frac{P(B|A) \cdot \enclose{circle}{P(A)}}{P(B|A) \cdot \enclose{circle}{P(A)} + P(B|-A) \cdot \enclose{circle}{P(-A)}}$$

$$= \frac{0.99 \times \cancel{0.002}\ 0.165}{0.99 \times \underset{0.165}{\cancel{0.002}} + 0.01 \times \cancel{0.998}\ \ 0.835}$$

$$= \cancel{0.165}\ \ 0.951$$

Distributions

- A distribution describes all of the probable outcomes of a variable.
- In discrete distribution, the sum of all individual probabilities must equal 1.
- In continuous distribution, the area under the probability curve equals 1.

Discrete Probability Distributions

- Discrete probability distributions are also called probability mass functions.
- It can be of three types:
    - Uniform distribution
    - Binomial Distribution
    - Poisson Distribution

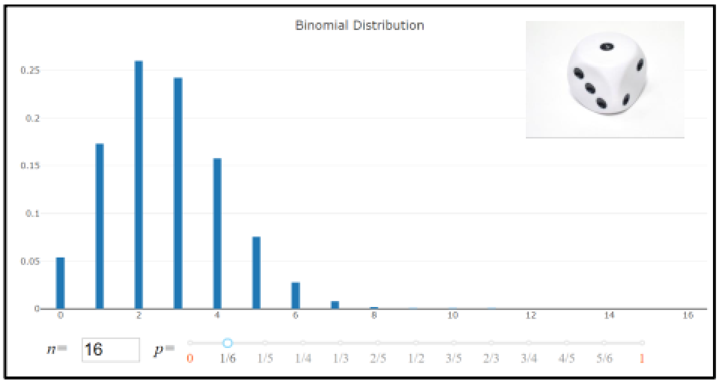
Uniform Distribution

- The probability of each outcome is evenly distributed across the sample.
    - Rolling a fair dice has 6 discrete equally probable outcomes.
    - All probabilities add to 1.

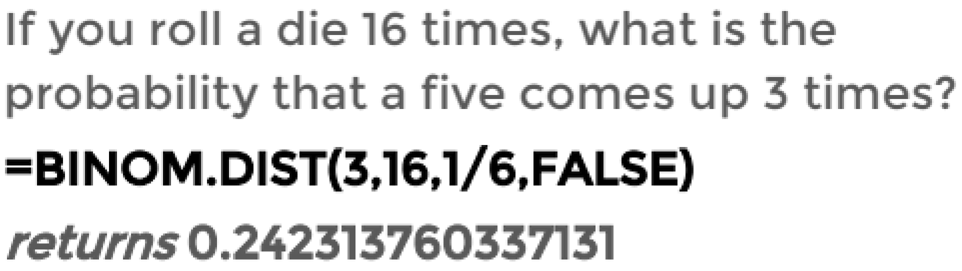
Binomial Distribution

- Binomial means there are two discrete mutually exclusive outcomes of a trial.
    - Heads or tails
    - On or off
    - Success or failure
- A **Bernoulli Trial** is a random experiment in which there are only two possible outcomes.
- A series of trials n will follow a binary distribution so long as
    - The probability of success p is constant
    - Trials are independent of each other.
- Probability of observing ***x*** successes in ***n*** trials if the probability of success on a single trial is denoted by ***p*** and p is fixed for all the trials

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



- If we roll a fair dice 16 times then probability of having a five for 3 times is 0.242.
  - This is the probability of having 5 exactly 3 times
- Using Excel

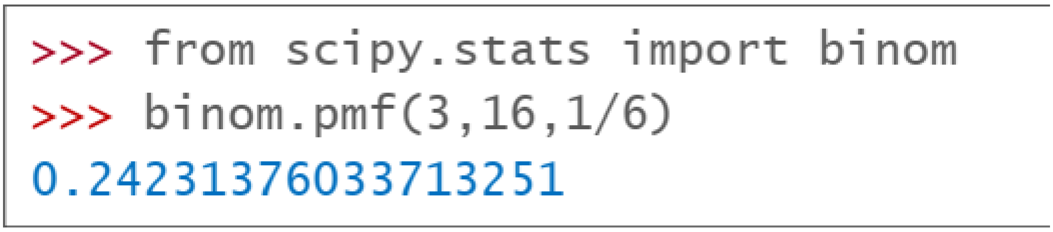


If you roll a die 16 times, what is the probability that a five comes up 3 times?

=BINOM.DIST(3,16,1/6,FALSE)

returns **0.242313760337131**

- Using Python



```
>>> from scipy.stats import binom
>>> binom.pmf(3,16,1/6)
0.24231376033713251
```

Poisson Distribution

- A Poisson Distribution considers the number of successes per unit of time* over the course of many units
  - * or any other continuous unit, e.g. distance
- Calculation of the Poisson probability mass function starts with a mean expected value.

$$E(X) = \mu$$

- This is then assigned to “Lambda”.

$$\lambda = \frac{\#\, occurrences}{interval} = \mu$$

- Probability for this distribution becomes

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

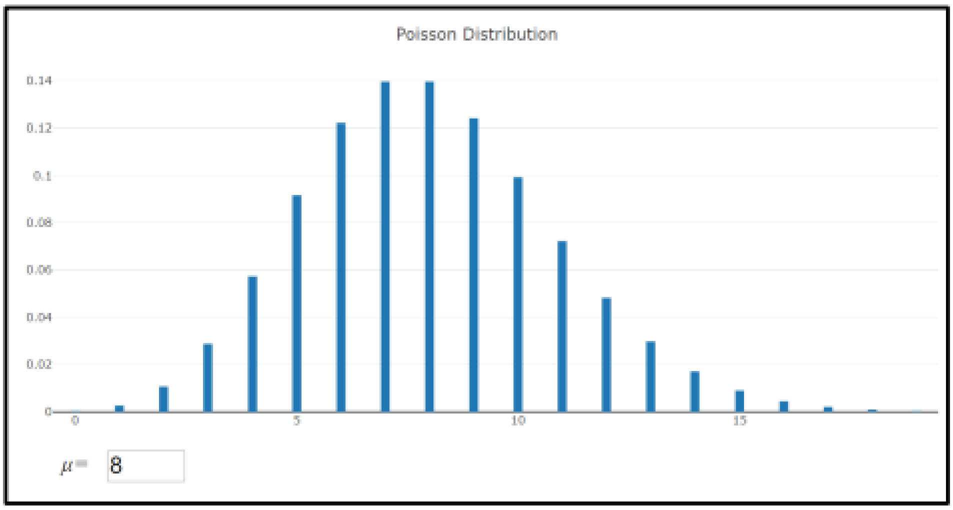
$$\text{where } e = Euler's\ number = 2.71828\ldots$$

- A warehouse typically receives 8 deliveries between 4 and 5pm on Friday. What is the probability that only 4 deliveries will arrive between 4 and 5pm this Friday?

$$x = 4 \quad \lambda = 8$$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{8^4 \cdot 2.71828^{-8}}{4!}$$

$$= \frac{4096 \cdot \left(\frac{1}{2980.96}\right)}{24} = \mathbf{0.0572}$$



- The cumulative mass function is simply the sum of all the discrete probabilities.
  - The probability of seeing fewer than 4 events in Poisson Distribution is:

$$P(X: x < 4) = \sum_{i=0}^{3} \frac{\lambda^i e^{-\lambda}}{i!}$$

$$= \frac{\lambda^0 e^{-\lambda}}{0!} + \frac{\lambda^1 e^{-\lambda}}{1!} + \frac{\lambda^2 e^{-\lambda}}{2!} + \frac{\lambda^3 e^{-\lambda}}{3!}$$

  - Sum of all probabilities equals 1. So, the probability of seeing at least 1 event is one minus the probability of seeing none:

$$P(X: x \geq 1) = 1 - P(X: x = 0)$$

$$= 1 - \frac{\lambda^0 e^{-\lambda}}{0!} = 1 - e^{-\lambda}$$

- The Poisson Distribution assumes that the probability of success during a small-time interval is proportional to the entire length of the interval. So, if know the expected value λ over an hour, then the expected value over one minute in the hour is

$$\lambda_{minute} = \frac{\lambda_{hour}}{60}$$

  - So, the probability that no deliveries arrive between 4:00 and 5:00 this Friday?

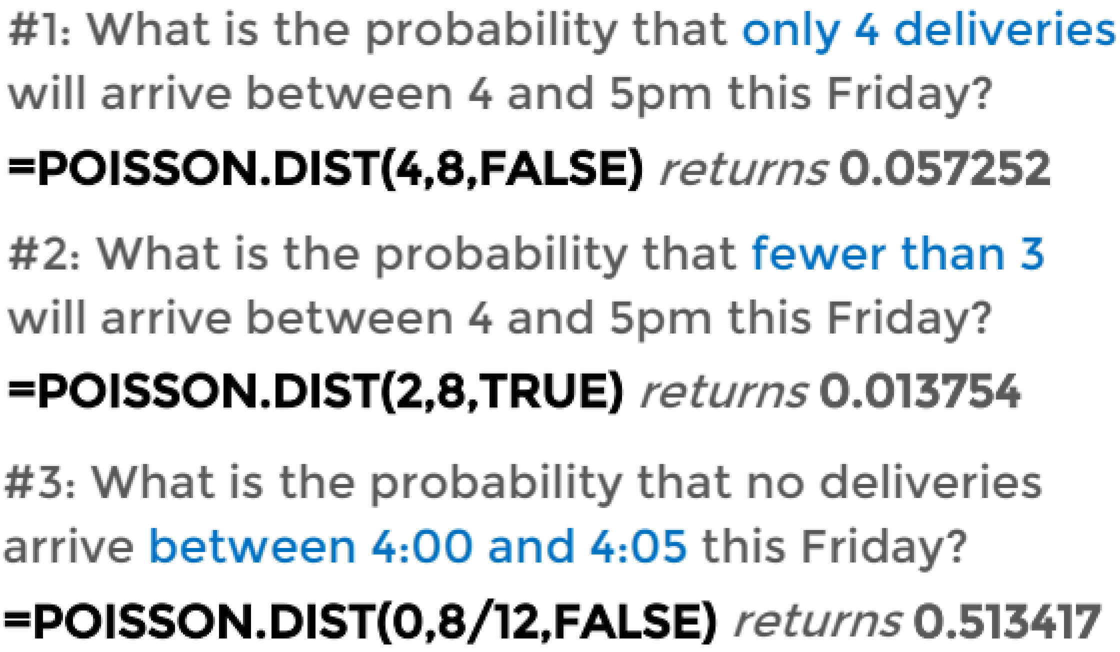
$$x = 0 \quad \lambda_{1\ hour} = 8$$

$$\lambda_{5\ minutes} = \frac{\lambda_{1\ hour}}{60/5} = \frac{8}{12} = 0.6667$$

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{0.67^0 \cdot 2.71828^{-0.6667}}{0!}$$

$$= \mathbf{0.5134}$$

- Using Excel



#1: What is the probability that only 4 deliveries will arrive between 4 and 5pm this Friday?

=POISSON.DIST(4,8,FALSE) *returns* **0.057252**

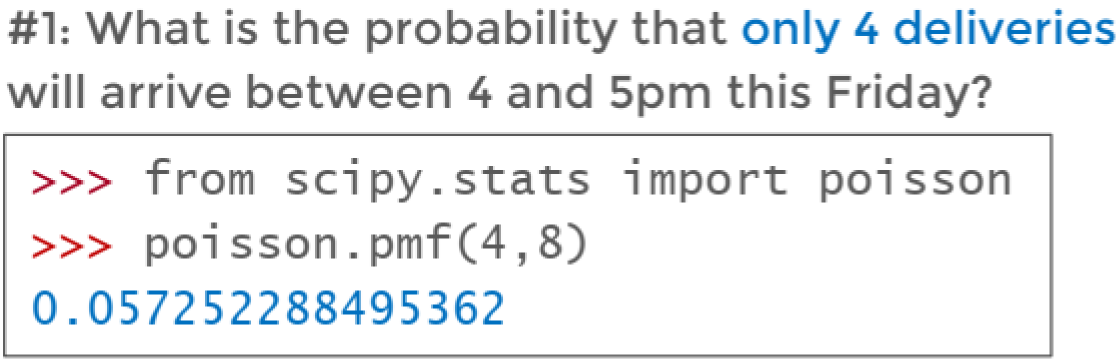
#2: What is the probability that fewer than 3 will arrive between 4 and 5pm this Friday?

=POISSON.DIST(2,8,TRUE) *returns* **0.013754**

#3: What is the probability that no deliveries arrive between 4:00 and 4:05 this Friday?

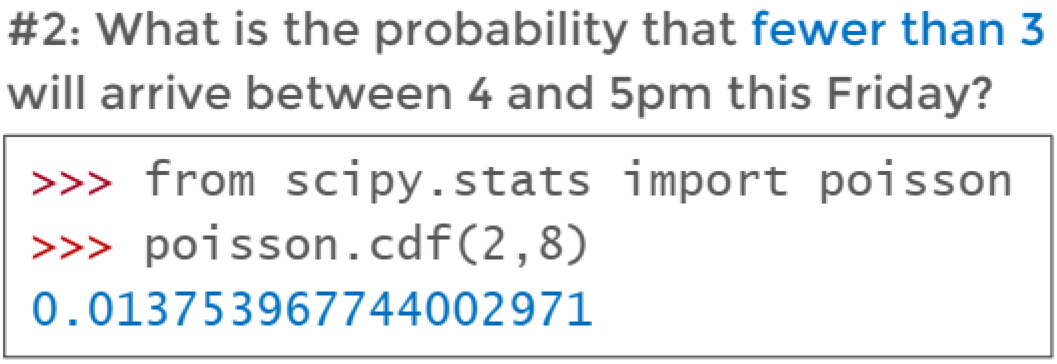
=POISSON.DIST(0,8/12,FALSE) *returns* **0.513417**

- Using Python



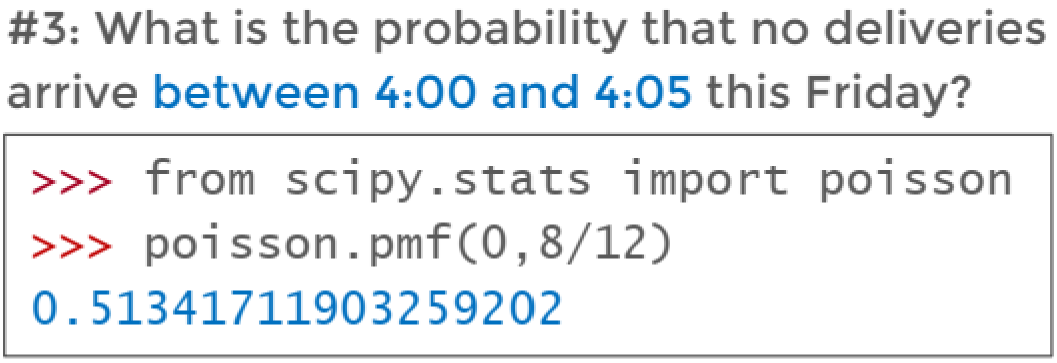
#1: What is the probability that only 4 deliveries will arrive between 4 and 5pm this Friday?

```
>>> from scipy.stats import poisson
>>> poisson.pmf(4,8)
0.057252288495362
```



#2: What is the probability that fewer than 3 will arrive between 4 and 5pm this Friday?

```
>>> from scipy.stats import poisson
>>> poisson.cdf(2,8)
0.013753967744002971
```



#3: What is the probability that no deliveries arrive between 4:00 and 4:05 this Friday?

```
>>> from scipy.stats import poisson
>>> poisson.pmf(0,8/12)
0.51341711903259202
```

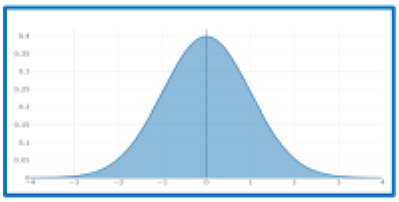
Continuous Probability Distributions

- Continuous probability distributions are also called probability density functions.
- It can be of types:
  - Normal Distribution
  - Exponential Distribution
  - Beta Distribution

Normal Distribution

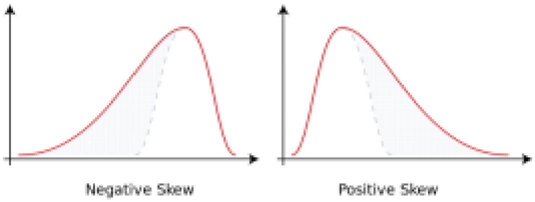
- Many real-life data points follow a normal distribution:
  - People's Heights and Weights
  - Population Blood Pressure
  - Test scores
  - Measurement Errors

- These data sources tend to be around a central value with no bias left or right, and it gets close to a Normal Distribution like this:

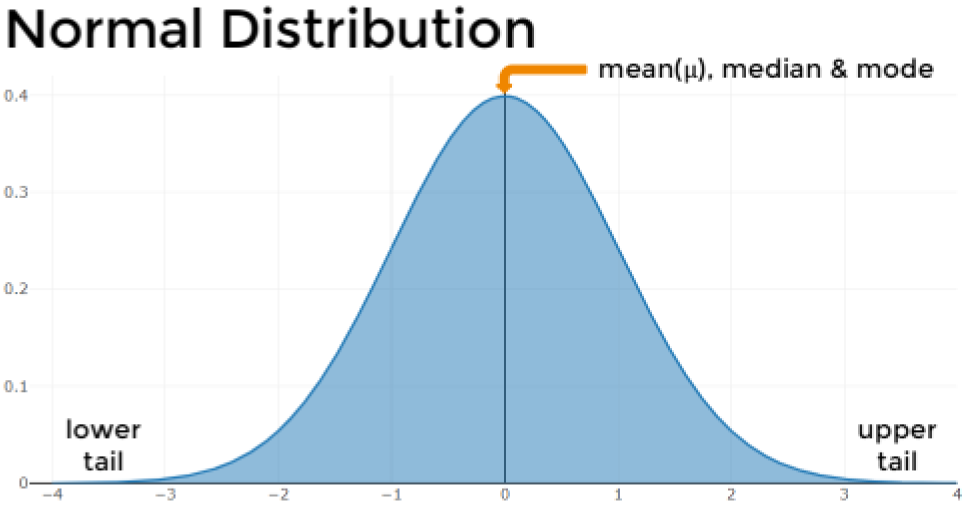


Normal Distribution

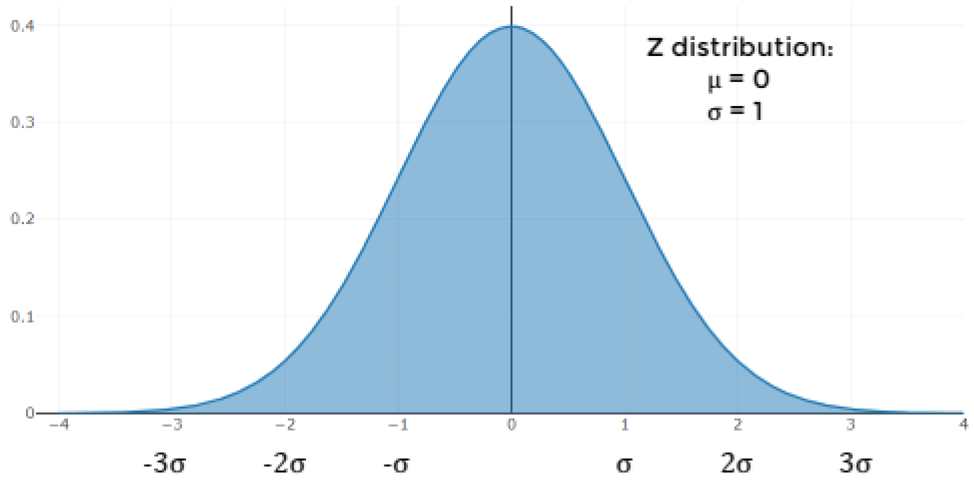
- We use a continuous distribution to model the behaviour of these data sources. Notice the continuous line and area in this diagram.
- In a normal distribution the area under the curve equals one.
- It is also called Bell Curve or Gaussian Distribution.
- Normal distribution is always symmetrical.
- Asymmetrical curves display skew and are not normal.



- The probability of a specific outcome is zero.
- We can only find probabilities over a specified interval or range of outcomes.

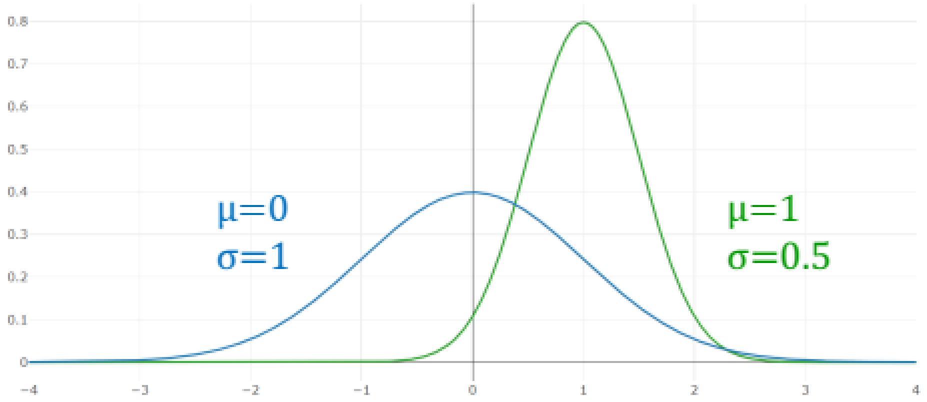


- For a standard normal distribution

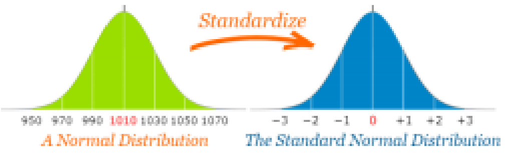


    - For a normal distribution mean does not need to be zero and standard deviation does not need to be 1 but for standard normal distribution they must have that value.
    - 68.27% of values lie between one standard deviation.
    - 95.45% of values lie between two standard deviation.
    - 99.73% of values lie between two standard deviation.

Other populations can be normal as well:



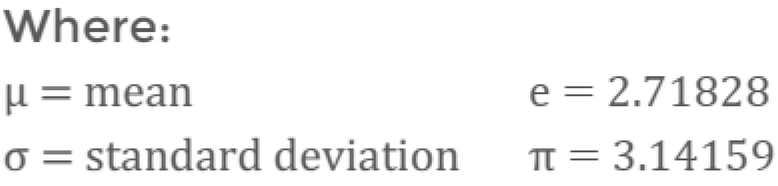
- If a population approximates a normal distribution, then powerful inferences about it can be drawn, if we know its mean and standard deviation.
- We can take any normal distribution and standardize it to a standard normal distribution through Z score.



    - Using Z score, we can calculate a particular x value's percentile.
        - A percentile is a way of saying "What percentage falls below this value", e.g., 90 percentile in CAT means 90% of students scored less than that student.

- Percentile can be calculated by converting an approximate normal distribution to standard normal distribution.
    - Formula for converting a normal distribution to standard normal distribution is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$



Where:

μ = mean	e = 2.71828
σ = standard deviation	π = 3.14159

where the new plot has mean equals to zero and standard deviation equals to 1.

- Z score for a particular x value in a normal distribution can be calculated by

$$z = \frac{x-\mu}{\sigma}$$

    - then the percentile of x can be determined by looking at a z-table.
- A z-table of Standard Normal Probabilities maps a particular z-score to the area under a normal distribution curve to the left of the score.
    - As the total area under the curve is 1, probabilities are bounded by 0 and 1.
    - Different tables serve different purposes

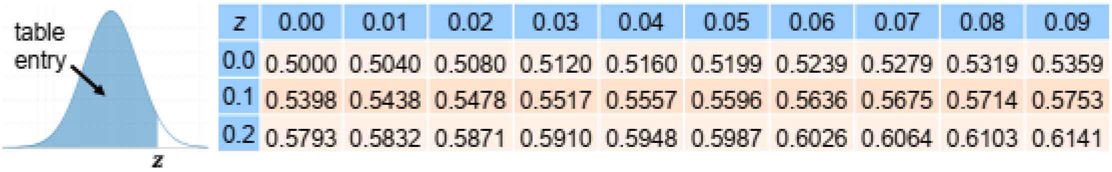


table
entry
z

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141

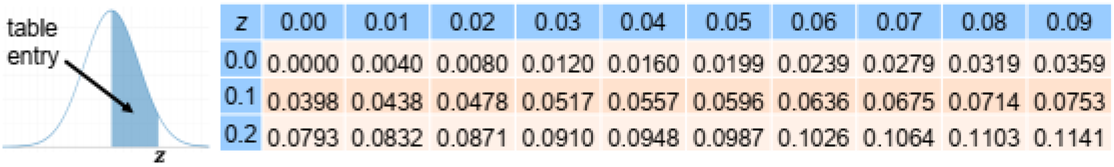
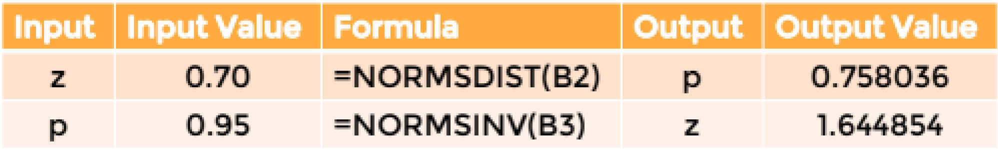


table
entry
z

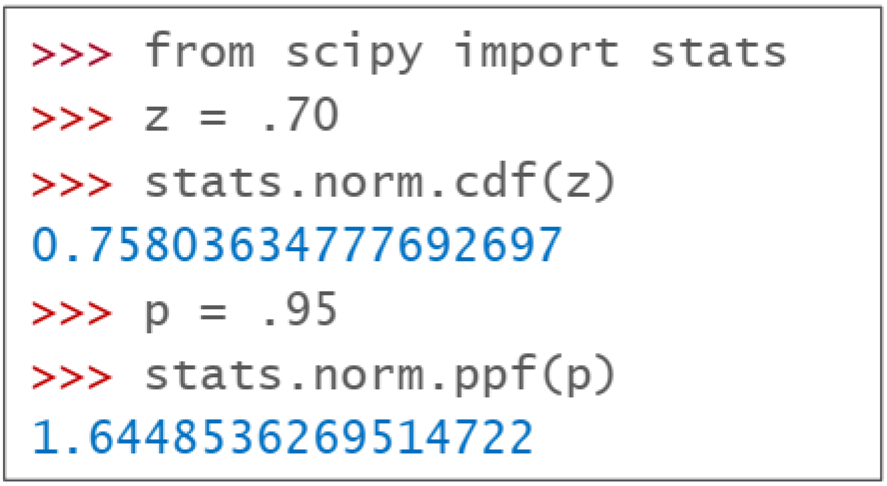
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141

- Z-Scores in MS Excel



Input	Input Value	Formula	Output	Output Value
z	0.70	=NORMSDIST(B2)	p	0.758036
p	0.95	=NORMSINV(B3)	z	1.644854

- Z-Scores in Python



```
>>> from scipy import stats
>>> z = .70
>>> stats.norm.cdf(z)
0.7580363477692697
>>> p = .95
>>> stats.norm.ppf(p)
1.6448536269514722
```

Statistics

- Statistics is the application of what we know to what we want to know.
- **Population** is every member of the group we want to study
- **Sample** is a small set of (hopefully) random members of the population.
- A **parameter** is a characteristic of a population. Often, we want to understand parameters.
- A **statistic** is a characteristic of a sample. Often, we apply statistical inferences to the sample in an attempt to describe the population.
- **Variable** is a characteristic that describes a member of the sample.
    - Variables can be discrete
        - Age
        - gender
        - salary
        - birthplace

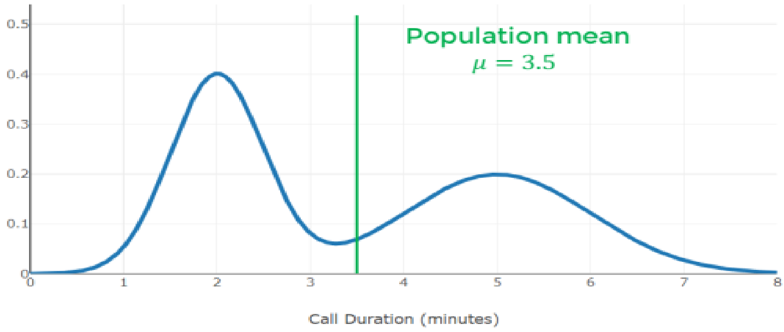
Sampling

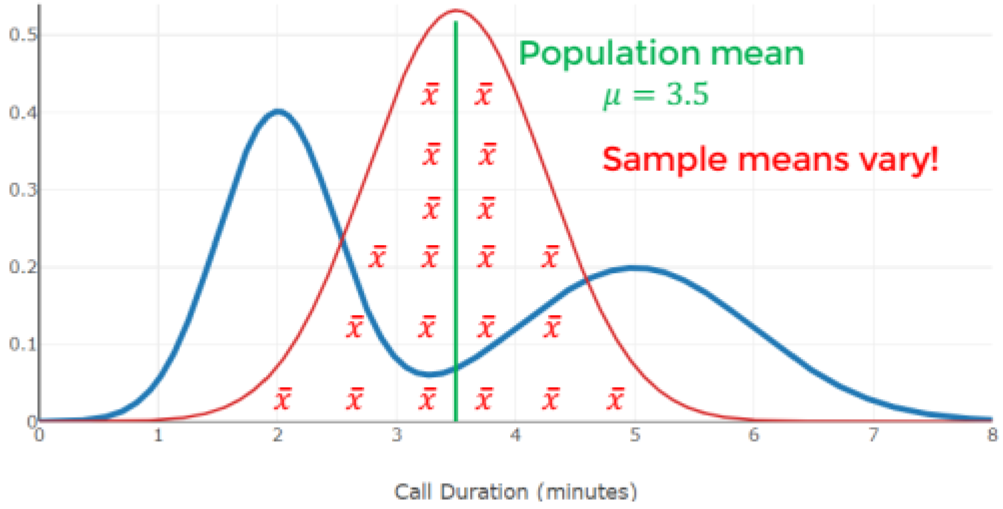
- One of the great benefits of statistical models is that a reasonably sized (>30) random sample will almost always reflect the population.
    - Challenge is to select members randomly and avoid bias.
- There are several forms of bias
    - **Selection Bias** is bias that favours those members of a population who are more inclined and able to answer polls.
    - **Under coverage Bias**: making too few observations or omitting entire segments of a population.
    - **Self-selection Bias:** people who volunteer may differ significantly from those in the population who don’t
    - **Healthy-user Bias:** the sample may come from a healthier segment of the overall population – people who walk/jog, work outside, follow healthier behaviours, etc.

    - **Survivorship Bias:** If a population improves over time, it may be due to lesser members leaving the population due to death, expulsion, relocation, etc
        - During war reinforcing those areas of the airplanes which were undamaged compared to those areas which were damaged.
- Types of Sampling
    - Random
    - Stratified Random
    - Cluster
- **Random Sampling**
    - Every member of the population has an equal chance of being selected.
    - However, since samples are usually much smaller than populations, there's a chance that entire demographics might be missed.
- **Stratified Random Sampling**
    - ensures that groups within a population are adequately represented.
    - First, divide the population into segments based on some characteristic.
        - Members cannot belong to two groups at once.
    - Next, take random samples from each group
    - The size of each sample is based on the size of the group relative to the population.
- **Clustering**
    - The idea is to break the population down into groups and sample a random selection of groups, or clusters.
        - Usually this is done to reduce costs.
        - A marketing firm sends pollsters to a handful of neighbourhoods (instead of canvassing an entire city)
        - A researcher samples fishing boats that are in port on a particular day (also known as convenience sampling)

Central Limit Theorem

- Makes sampling such a good statistical tool.
- Sample mean often varies from the population mean
- The CLT considers a large number of random sample tests.
- The CLT states that the mean values from a group of samples will be normally distributed about the population mean, even if the population itself is not normally distributed because sample means vary.

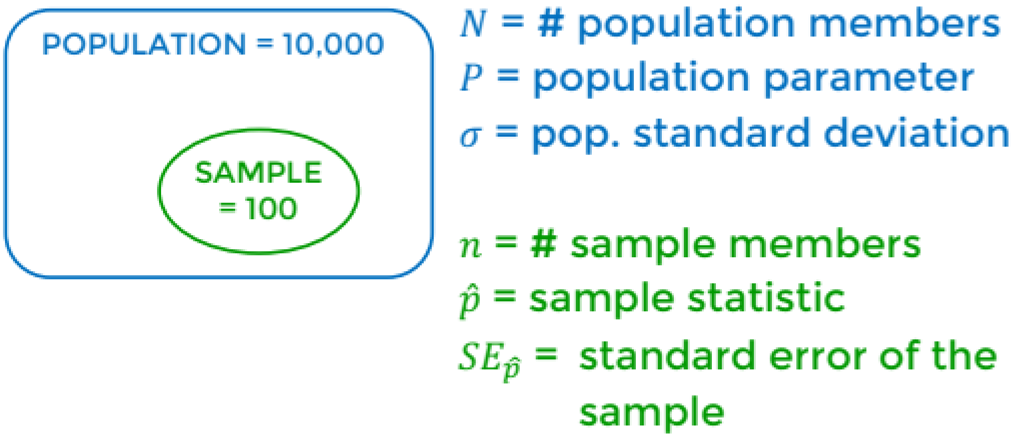




- That is, 95% of all sample means should fall within 2σ of the population mean.
- As we collect multiple samples, each mean will fall somewhere close to the population mean.

Standard Error

- It describes how far a sample mean stray from the population mean as compared to population standard deviation which describes how wide individual values stray from the population mean.



N = # population members
P = population parameter
σ = pop. standard deviation

n = # sample members
$\hat{p}$ = sample statistic
$SE_{\hat{p}}$ = standard error of the sample

- If the population standard deviation σ is known, then the sample standard error of the mean can be calculated as:

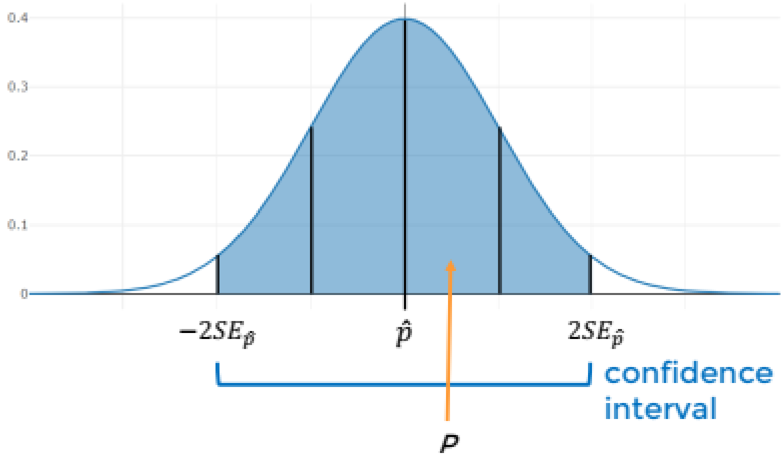
$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- An IQ Test is designed to have a mean score of 100 with a standard deviation of 15 points. If a sample of 10 scores has a mean of 104, can we assume they come from the general population?

$$n = 10 \quad \overline{x} = 104 \quad \sigma = 15$$

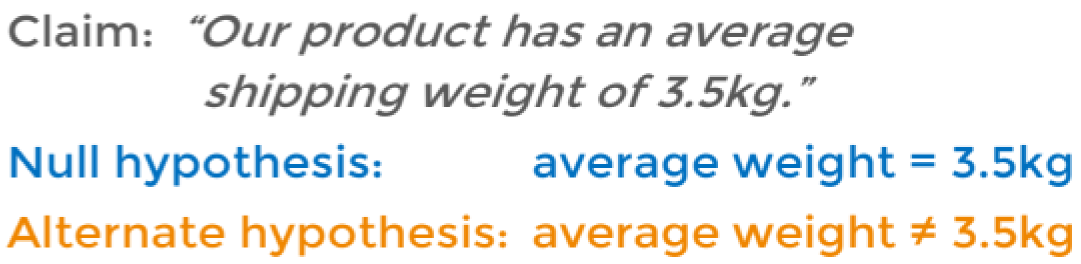
$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{15}{\sqrt{10}} = 4.743$$

  - 68% of 10-item sample means are expected to fall between 95.257 and 104.743
  - “We can say with a 95% confidence level that the population parameter lies within a confidence interval of plus-or-minus two standard errors of the sample statistic”.



Hypothesis Testing

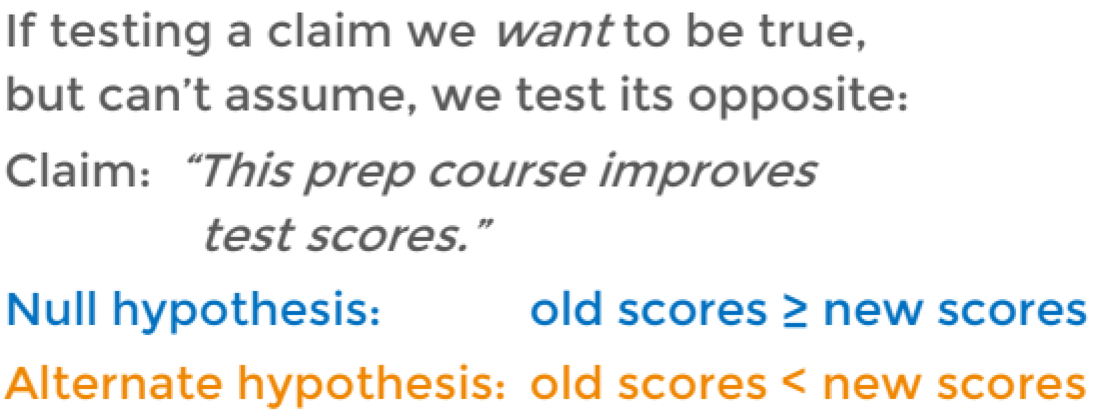
- Hypothesis Testing is the application of statistical methods to real-world questions.
- We start with an assumption, called the **null hypothesis**
- We run an experiment to test this **null Hypothesis**. Based on the results of the experiment, we either **reject or fail to reject** the null hypothesis
- If the null hypothesis is rejected, then we say the data supports another, mutually exclusive **alternate hypothesis**
- We **never “PROVE”** a **hypothesis**!
- How do we frame the question that forms our null hypothesis?
  - At the start of the experiment, the null hypothesis is assumed to be true.
  - If the data fails to support the null hypothesis, only then can we look to an alternative hypothesis
  - If testing something assumed to be true, the null hypothesis can reflect the assumption:



Claim: *"Our product has an average shipping weight of 3.5kg."*

Null hypothesis: average weight = 3.5kg

Alternate hypothesis: average weight ≠ 3.5kg



If testing a claim we *want* to be true, but can't assume, we test its opposite:

Claim: *"This prep course improves test scores."*

Null hypothesis: old scores ≥ new scores

Alternate hypothesis: old scores < new scores

- The null hypothesis should contain an equality (=, ≤ ,≥):
  - average shipping weight = 3.5kg

    $\boxed{H_0\colon \mu = 3.5}$

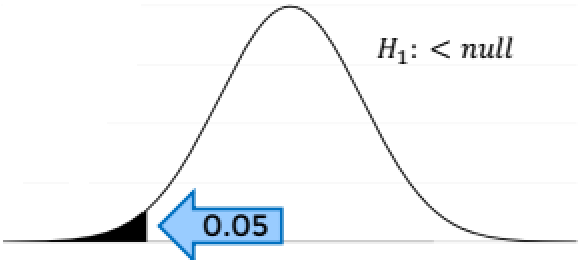
    old scores ≥ new scores

    $\boxed{H_0\colon \mu_0 \geq \mu_1}$
- The alternate hypothesis should not have an equality (≠,<,>):
  - average shipping weight ≠ 3.5kg

    $\boxed{H_1\colon \mu \neq 3.5}$

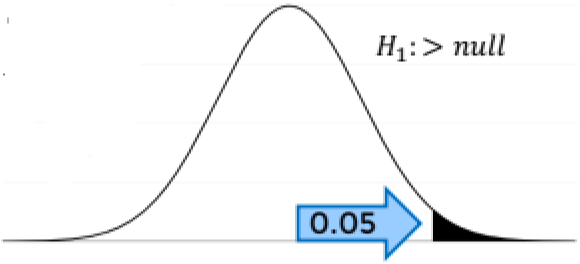
    old scores < new scores

    $\boxed{H_1\colon \mu_0 < \mu_1}$
- We reject or fail to reject the null hypothesis by running an experiment and recording the results
  - If probability of observing the result of null hypothesis is very small (<=0.05), then we reject the null hypothesis.
    - Here 0.05 is our level of significance (α = 0.05)
- The level of significance α is the area inside the tail(s) of our null hypothesis.
- If α = 0.05 and the alternate hypothesis is less than null, then the left tail of our probability curve has an area of area of 0.05.

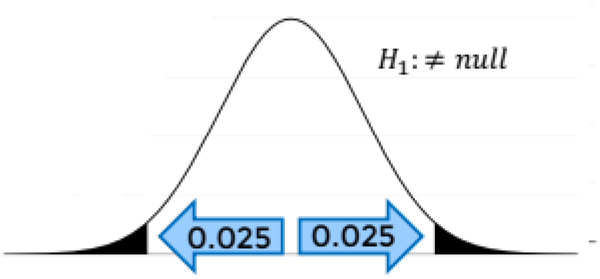


  - Notice the left hand side has negative Z-score.

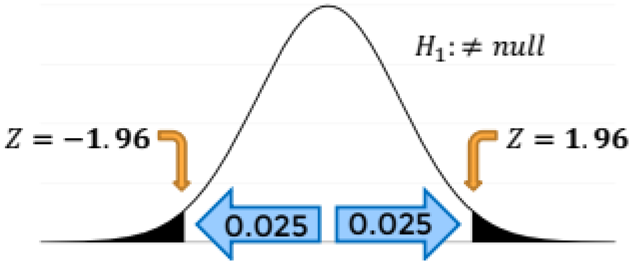
- If α = 0.05 and the alternative hypothesis is more than the null, then the right-tail of our probability curve has an area of 0.05



- If α = 0.05 and the alternative hypothesis is not equal to the null, then the two tails of our probability curve share an area of 0.05



- These areas establish our critical values or Z-scores.



Tests of Mean vs. Proportion

- Each of these two types of tests has their own test statistic to calculate.
- **Mean:** When we looking to find an average, or specific value in a population we are dealing with the means.
- When we are working with the means:

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

assumes we know the population standard deviation

- **Proportion:** Whenever we say something like "35%" or "most" we are dealing with proportions.

- When we are working with the proportions:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1 - p)}{n}}}$$

Traditional Test

- Take the level of significance α
- Use it to determine the critical values of Z-score.
- Compare the test statistic to the critical value.

P-Value Test

- Take the test statistic to calculate the Z-score.
- Use it to determine the P-value
- Compare the P-value to the level of significance α.
  - If the P value is low, the null must go!
    - Reject H_0.
  - If the P-value is high, null must fly!
    - Fail to reject H_0.

Type I and Type II Errors

- Often in medical fields (and other scientific fields) hypothesis testing is used to test against results where the "truth" is already known.
- For example, testing a new diagnostic test for cancer for patients we have already successfully diagnosed by other means.
- In this situation, we already know if the Null Hypothesis is True or False.
- In these situations where we already know the "truth", then you would know it’s possible to commit an error with your results.
- **Type I Error (False Positive):**
  - If we reject a null hypothesis that should have been supported
    - $\boldsymbol{H}_0: There\ is\ no\ fire$
    - Pull the fire alarm, only to find out there really was no fire.
- **Type II Error (False Negative):**
  - If we fail to reject a null hypothesis that should have been rejected, we’ve committed
    - $\boldsymbol{H}_0: There\ is\ no\ fire$
    - Don’t pull the fire alarm, only to find there really is a fire.

Student’s T-Distribution

- Developed by William Sealy Gossett while he was working at Guinness Brewery. Published under the pseudonym “Student” as Guinness wouldn’t let him use his name. Goal was to select the best barley from small samples, when the population standard deviation was unknown!
- Purpose: Using t-table, the student’s t-test determines if there is significant difference between two sets of data.
    - Due to variance and outliers, it is not enough just to compare mean values
    - A t-test also considers sample variances