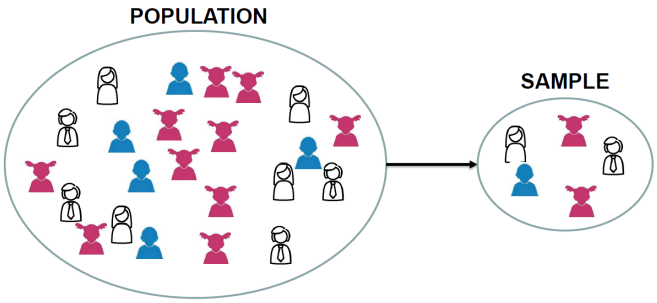
Census- complete population
Survey- Sample of population

Population and Sample



Parameter - Descriptive measurement of **population**
Statistics - Descriptive measurement of the **sample**

Greek – Population Parameter	**Roman – Sample Statistic**
Mean – μ	Mean – $\bar{x}$
Variance – σ^2	Variance – s^2
Standard Deviation - σ	Standard Deviation - s

Descriptive and Inferential Statistics

- Descriptive Statistics – Data gathered about a group to reach conclusions about the same group.
- Inferential Statistics – Data gathered from a sample and the statistics generated to reach conclusions about the population from which the sample is taken. Also known as Inductive Statistics.

Note: 1. Mostly we will be doing descriptive statistics. As most of the time data will be given to us and by understanding it we need to deliver insights.

2. Where Inferential Statistics we do on sample data to conclude about the population.

Ex- People facing diabetes in india
So in this case we will do a survey and then conclude about the population by doing inferential statistics.

Dependent and independent variable.

Let's keep it simple and straight for now
Independent variables are the parameter based on which your outcome will be predicted. That means your outcome variable depends upon the independent variables i.e parameters.

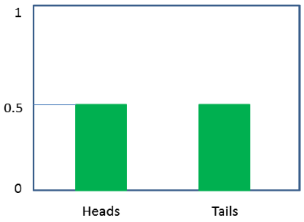
Type of data.

1. Categorical data
    a. Nominal - Anything independent ex- name, age, place.
    b. Ordinal - Region,country,state,city, area, pincode
    Ordinal follows a chain of higher to lower priority.

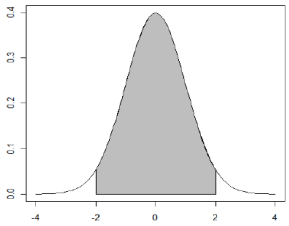
2. Numeric data

Data can be also classified as:

Discrete and continuous data.



Countable



Measurable

Example:

Time between customer arrivals at a retail outlet	Continuous
No. of customers arriving at a retail outlet during a five-minute period	Discrete
Lengths of newly designed automobiles	Continuous
Sampling 100 voters in an exit poll and determining how many voted for the winning candidate	Discrete
No. of defects in a batch of 50 items	Discrete

Describing data through stats:

Central Tendencies: Mean, Median and Mode.

Example : Calculate mean (average)

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2

$$\text{Mean, } \mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{100X10+345X1+1000X10+9833X2}{10+1+10+2} = 1348$$

f/n - frequency

Median: Arrange data in increasing order and find the mid-point $\frac{(n+1)}{2}$.

100,100,100,100,100,100,100,100,100,100,
345,1000,1000,1000,1000,1000,1000,1000,
1000,1000,1000,9833,9833

MODE - The most frequently occurring data point.
Mode plays the role of mean when it is categorical data.

Measuring Variability and Spread in the data.

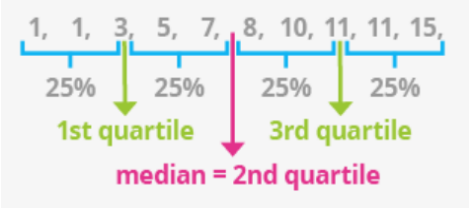
Range= Max-Min

Example - If all the three players in cricket have the same central tendencies of the score they scored in the previous last 5matches.
Now to whom will you choose in the upcoming series.
In such a case we can go for option range, from which we can infer some more information.

Detecting outliers in the data.

You have data of salary of different people, if the salary of the Tata, Birla or ambani is also present in the dataset. Then the Resultant means will not be truthful to describe the average salary of an indian.

To detect the outliers we convert our data in a range of 0-100(ascending order) i.e making it in percentage so that we can describe the data in percentile.
Then we split our data in 4 bins of 25%.
And so we get 3 boundaries - upto 25% , upto 50% and upto 75%.



We define these boundaries as quartiles which are calculated as shown below.

Lower quartile (25^{th} percentile, Q1) = $\frac{(n+1)}{4}$th

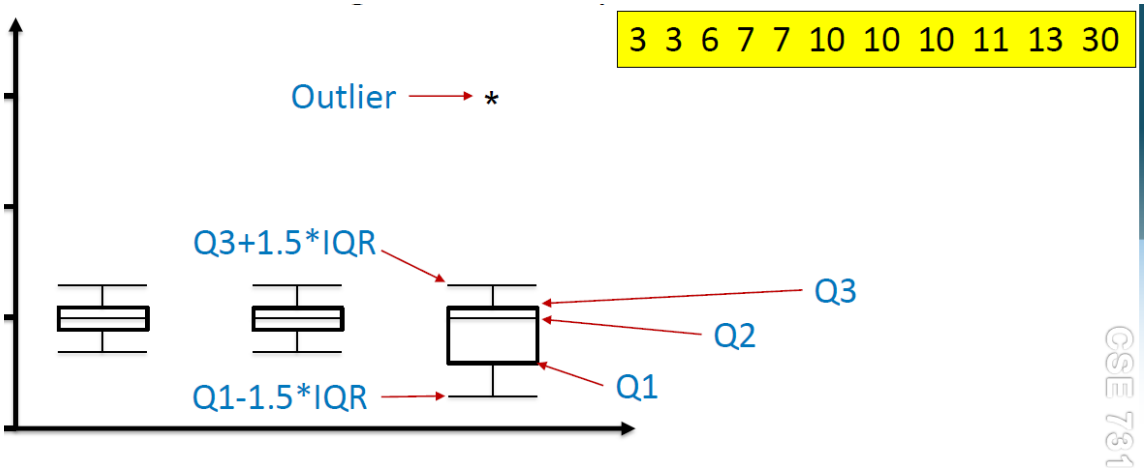
Middle quartile = Median = $\frac{2*(n+1)}{4}$th

Upper quartile (75^{th} percentile, Q3) = $\frac{3*(n+1)}{4}$th

Interquartile range, IQR = Q3-Q1 (central 50% of data)

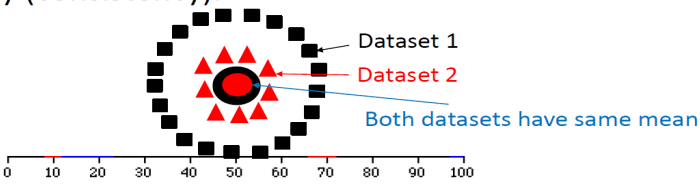
IQR represents 50% of the data.

Based on these calculations we calculate the upper and lower whisker the boundary above or below, by which we can say that the values are too high or low so i.e can be stated as outliers.



Measuring Variability and Spread

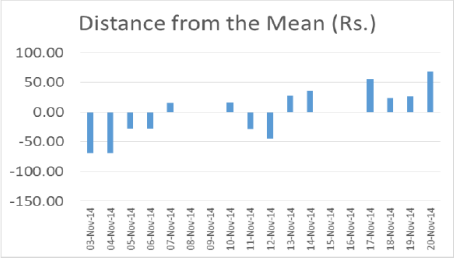
Range and IQR give the spread but still do not describe variability (consistency).

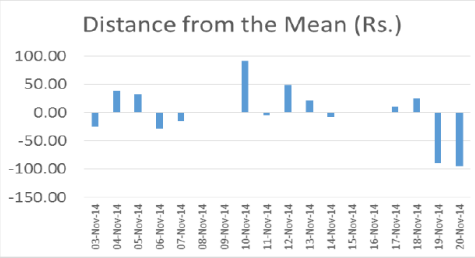


Average distance from the mean?

3 3 6 7 7 10 10 10 11 13 30

Measures of Spread – Mean Distance, Mean Absolute Deviation or Standard Deviation - Excel





- Mean Distance in both cases = 0
- Mean Absolute Deviation in both cases = 38.17
- Std Dev is 42.54 in the first case and 48.80 in the second.

So in example above:

1. The mean distance of each point when calculated for both years came to zero. Because of negative and positive trends, they got cancelled.
2. So to avoid this we calculated Mean absolute deviation where the negative signs were converted to the positive so nothing can now be cancelled. But unfortunately this time the number which we got came the same. So unable to infer the difference in the stock of 1st and 2nd year.
3. To overcome this we went for the most powerful metric which is mean square distance. In this the negative signs are also removed as we square them.
   The mean square distance is called variance. But the problem with variance is that it squares the value. And when we square the squared rupees doesn't make sense. So to avoid this we take its square root. Hence converting variance to **standard deviation.**

Variance = $\frac{\Sigma(x-\mu)^2}{n} = \frac{\Sigma x^2}{n} - \mu^2$ (Derive)

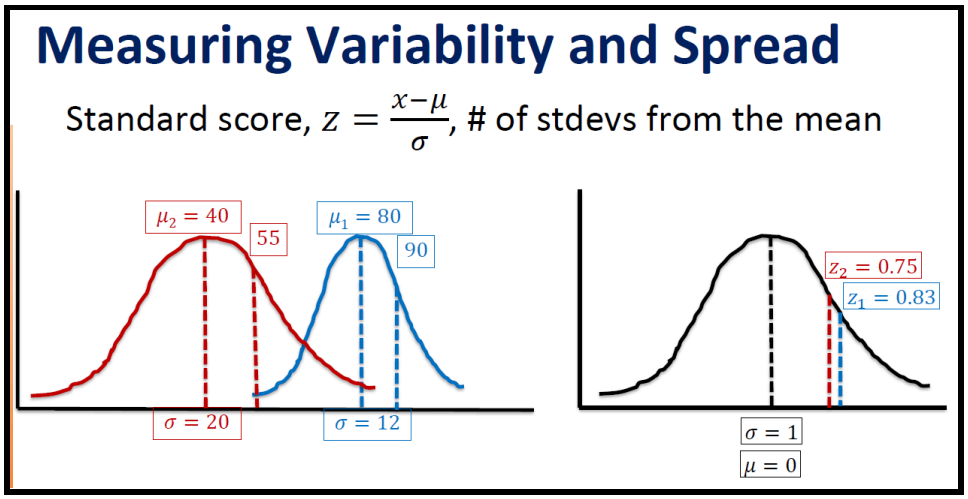
3 3 6 7 7 10 10 10 11 13 30

Units are squared, which is not intuitive.

Standard Deviation, $\sigma = \sqrt{Variance}$

Example:
Now we know the individual std deviation and mean of the players.
In the next match let's say both scored some runs.
How can we infer that from both who have scored better than their individual track records.

In such case we go for **Z-Score**



In the above example z1 is better as he/she is more deviated than his average performance.

Probability vs Statistics

- Probability – Predict the likelihood of a future event
- Statistics – Analyze the past events

- Probability – What will happen in a given ideal world?
- Statistics – How ideal is the world?

Probability is the basics of inferential statistics:

As we conclude something from the data we infer and we give the answer that how certain we are and that is in probability.

Types of probability:

1. Classical method = A priori or Theoretical - Here we can orally calculate the probability.
   Ex - what will be the probability of getting a head if I toss a coin for 1 time.
   Answer - 0.5 (known)

2. **Empirical Method** = posteriori or Frequentist.
   Here we will calculate the frequency based on which we tell the probability.
   Example- We will toss the coin 10times and count the number of times the head came.
   So probability = total counts of head / total toss of coin.

   So we will be following empirical method to deal with our inferential statistics.

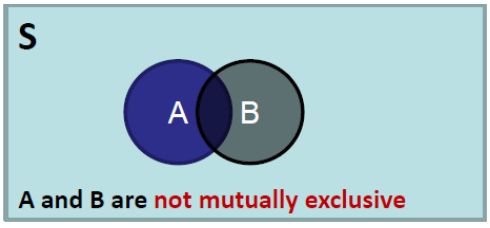
- **Sample set and Event:**
  Sample set = set of all possible outcomes, denoted by S.
  Event = A subset of the sample set.

$$Probability = \frac{Frequency\ of\ the\ Event\ of\ Interest}{Total\ Frequency}$$

Rules of probability:
P(A and B) = P(A) * P(B)

mutually exclusive - mutual means either of the answer
Either head or tails
So, P = 0.5 + 0.5 =1
P(A or B) = P(A) + P(B)

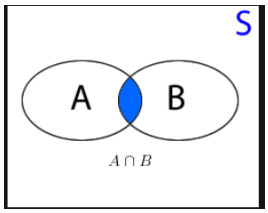
Not mutually exclusive - means both event are individual



Example:
P(A) = loan defaulter
P(B) = Are rich
So, **P(A or B) = P(A) +P(B) - P(A and B)**
As they can be both loan defaulters but rich.

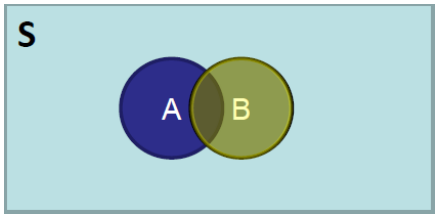
Types of probability:

1. **Joint or Intersection: P(A and B) = P(A) * P(B)**



2. Union
   P (A or B) =P(A) + P(B) - P (A and B)
3. Marginal Probability
   P(No)
   P(yes)
   Probability describing a single attribute

4. Conditional probability = Joint probability / marginal probability.
   Conditional probability - probability of A occurring given that B has occurred.



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

Similarly we can get P(B|A) and then equating 2 equation we come up with :

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

Now with this equation we can find out reverse probabilities.

Bayes' Theorem

Bayes' Theorem allows you to find reverse probabilities, and to allow **revision of original probabilities** with new information.

$$P(A|B) = \frac{P(A)*P(B|A)}{P(B|A) * P(A) + P(B \mid \text{not } A) * P(\text{not } A)}$$

Example:

Case – Spam filtering

P(Spam) = 0.50

P(Free | Spam) = 0.20

P(Free | No spam) = 0.001

P(Spam | Free) = ?

$$P(Spam|Free) = \frac{P(Spam) * P(Free|Spam)}{P(Free|Spam) * P(Spam) + P(Free|No\ spam) * P(No\ spam)}$$

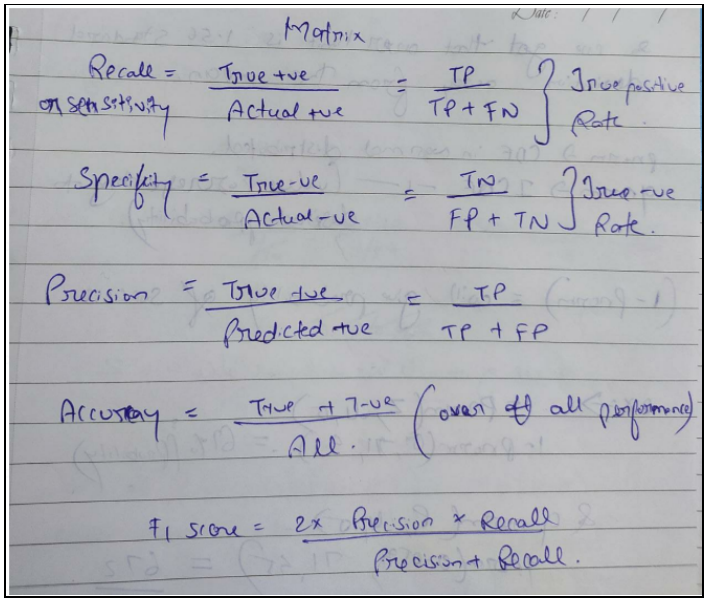
$$= \frac{0.5 * 0.2}{0.2 * 0.5 + 0.001 * 0.5} = \frac{0.1}{0.1005} = 0.995$$

Confusion matrix: **Table layout that allows visualization of the performance of an algorithm**

Confusion Matrix

Spam filtering		Predicted		Total
		Positive	Negative	
Actual	Positive	952	526	1478
	Negative	167	3025	3192
Total		1119	3551	4670

		Predicted		
		Positive	Negative	
Actual	Positive	True +ve	False –ve	Recall/Sensitivity/True Positive Rate (Minimize False –ve)
	Negative	False +ve	True –ve	Specificity/True Negative Rate (Minimize False +ve)
		Precision		Accuracy, F_1 score



Recall: its ratio of predicted positive by actual number of positives.
So it tells you how correctly you have classified True cases.

Specificity: vice versa, telling how correctly you classified False cases.

Precision: It states the quality of prediction.
Its ratio of actual number of positive by total number of cases we predicted as positive.

Accuracy:over all performance.

F1 score = weighted average of recall and precision.

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

More reliable metric, as we are taking both.

Random Variable

- A variable that can take multiple values with different probabilities.
- The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.

Points scored per game	0	1	2	3	4	5	6
Frequency, f	1	4	6	12	5	1	1

Points scored per game	0	1	2	3	4	5	6
Probability *Recall the Frequentist (empirical) approach of assigning probabilities*	$\frac{1}{30}$	$\frac{4}{30}$	$\frac{6}{30}$	$\frac{12}{30}$	$\frac{5}{30}$	$\frac{1}{30}$	$\frac{1}{30}$

Leads to Descriptive Stats

Leads to Inferential Stats

Terminology used to describe discrete and continuous distribution:

Discrete Distributions	Continuous Distributions
Probability that X can take a specific value x is $P(X = x) = p(x)$.	Probability that X is between two points a and b is $P(a \leq X \leq b) = \int_a^b f(x)dx$.
It is non-negative for all real x.	It is non-negative for all real x.
The sum of $p(x)$ over all possible values of x is 1, i.e., $\sum p(x) = 1$.	$\int_{-\infty}^{\infty} f(x)dx = 1$
Probability Mass Function	**Probability Density Function**

Once we calculate distribution, we can calculate **expectation.**
Expectation is nothing but mean. But in inferential stats we say as expectation.

EXPECTATION, $\mathrm{E}(X) = \mu = \Sigma x P(X = x)$

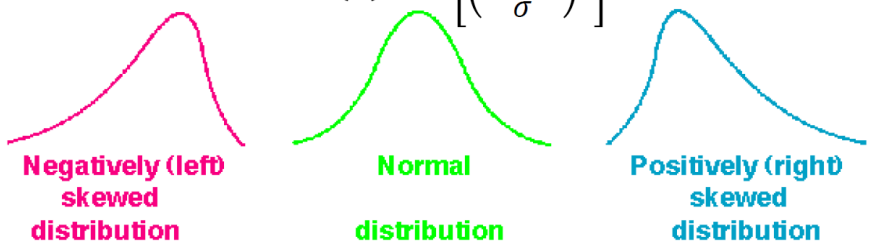
VARIANCE, $Var(X) = E(X - \mu)^2 = \Sigma(x - \mu)^2 P(X = x)$

$$\sigma = \sqrt{Var(X)}$$

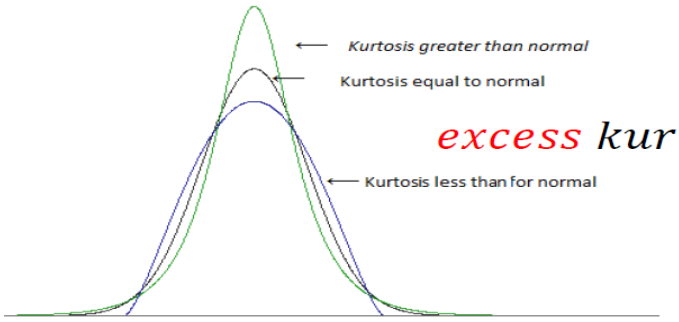
Skewness

- A measure of symmetry. Negative skew indicates mean is less than median, and positive skew means median is less than mean.

$$skew(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$$



Kurtosis:
Tells how much data is centrally located.



$$excess\ kurt(X) = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - 3$$

SOME COMMON DISTRIBUTION:

Bernoulli

There are two possibilities (loan taker or non-taker) with probability *p* of success and *1-p* of failure

- Expectation: *p*
- Variance: *p(1-p)* or *pq*, where *q=1-p*

$$Expectation, E(X) = \sum x_i P(x_i)$$
$$= 1 * p + 0 * q = p$$

$$Variance, Var = \sum (x_i - \mu)^2 P(x_i)$$
$$= (1-p)^2 * p + (0-p)^2 * (1-p)$$
$$= p(1-p)$$

We take 1 as success (P) and failure as zero (q)

Keynote:

• You run a series of independent trials.
• There can be either a success or a failure for each trial, and the probability of success is the same for each trial.

Geometric Distribution :

Number of independent and identical Bernoulli trials needed to get ONE success.
Example :- number of people I need to call for the first person to accept the loan.

Geometric Distribution
PMF*,

1. $P(X = r) = q\ r-1p$
   (r-1) failures followed by ONE success.
2. $P(X > r) = q\ r$
   Probability you will need more than r trials to get the first success.
3. CDF**, $P(X \leq r) = 1 - q\ r$
   Probability you will need r trials or less to get your first success.

$E(X) = 1/p$
$Var(X) = q / p\ 2$
* Probability Mass Function ** Cumulative Distribution Function

Keynote:

• You run a series of independent trials.
• There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
• The main thing you are interested in is how many trials are needed in order to get the first successful outcome.

Binomial Distribution:

In Binomial Distribution we are interested in the number of successes for the given number of trials.

If there are two possibilities with probability p for success and q for failure, and if we perform n trials, the probability that we see r successes is

PMF, $P(X = r) = C_r^n p^r q^{n-r}$

CDF, $P(X \leq r) = \sum_{i=0}^{r} C_i^n p^i q^{n-i}$

Where C is the combine or choice

Binomial distribution runs on the principle of Bernoulli, the only difference is we run a series of trials here.
So the equation of Expectation and Variance remains the same just n, no. of trails is added in front of it.

$E(X) = np$ $Var(X) = npq$

When to use?

- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- There are a finite number of trials, and you are interested in the number of successes or failures.

Poisson Distribution:

Let's check the difference between binomial and poisson.

Binomial: We are interested in number of successes/events (discrete) occurring randomly in fixed *number of trials* (discrete).

Poisson: We are interested in number of successes/events (discrete) occurring randomly in fixed *duration or space* (continuous).

As the factor duration is coming so we will be requiring average now 'Lambda'

$$\text{PMF, } P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!}$$

$$\text{CDF, } P(X \leq r) = e^{-\lambda}\sum_{i=0}^{r}\frac{\lambda^i}{i!}$$

Note: When number of trails are more than 50 and probability is less than 0.1
Then Binomial distribution behaves like poisson distribution. Hence results can be calculated using the poisson equation.
where , Lambda = np
Case:

The probability that no customer will visit the store in one day

$$P(X=0) = \frac{e^{-\lambda}\lambda^0}{0!} = e^{-\lambda}$$

Probability that no customer will visit in *n* days

$$e^{-n\lambda}$$

Exponential Distribution: Time between events.

Probability that a customer will visit in *n* days:

$$1 - e^{-n\lambda}$$

$$CDF = 1 - e^{-n\lambda}, n \geq 0$$
$$PDF = \lambda e^{-n\lambda}, n \geq 0$$

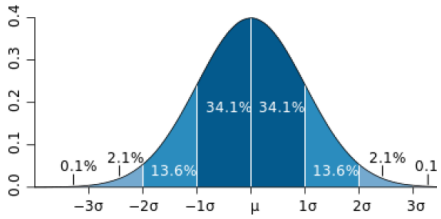
E(X) = 1/lambda and Var(X)=1/(lambda)^2

- The *number* of events in a given time period — Poisson
- The *time* until the first event
- The *time* from now until the next occurrence of the event
- The *time interval* between two successive events

— Exponential

Normal (Gaussian) Distribution

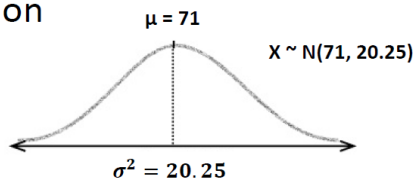
- Mean = Median = Mode
- 68-95-99.7 empirical rule
- Zero Skew and Kurtosis
- X ~ N(μ, σ^2)
- Shaded area gives the probability that X is between the corresponding values

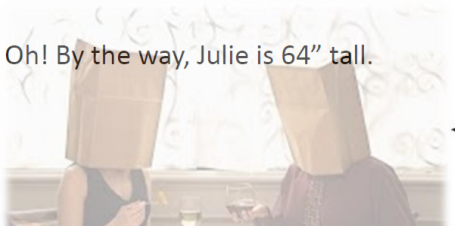


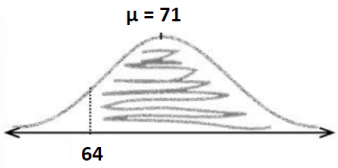
Calculating Normal Probabilities

Step 1: Determine the distribution

Julie wants to marry a person taller than her and is going on blind dates. The mean height of the 'available' guys is 71" and the variance is 20.25 inch^2 (yuck!).

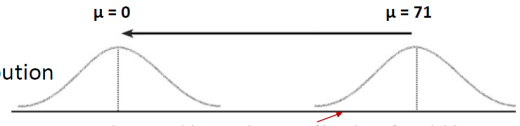




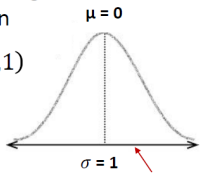


Step 2: Standardize to Z ~ N(0,1)

1. Move the mean
   This gives a new distribution
   X-71 ~ N(0,20.25)
2. Squash the width by dividing by the standard deviation
   This gives us $\frac{X-71}{4.5} \sim N(0,1)$



Random variable is **x**, the ***actual*** heights of available guys

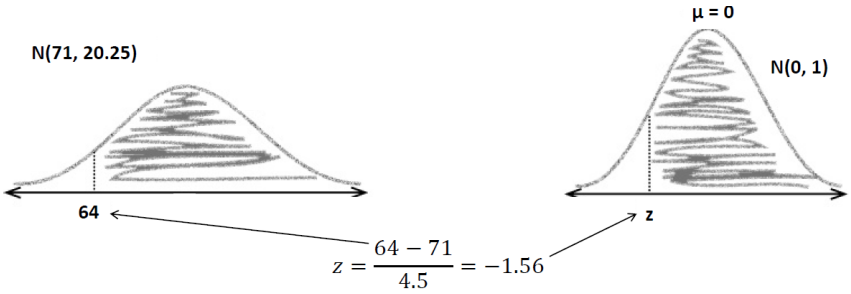


$Z = \frac{X-\mu}{\sigma}$ is called the Standard Score or the z-score.

Random variable is **z**, the the ***standardized*** heights of available guys

Note: z~N(mean,variance)

But to Calculate z= X - mean / Std deviation
So variance needs to be converted to std deviation every time.



Julie is 64" tall, i.e., she is 1.56 standard deviations shorter than the average height of the available guys.

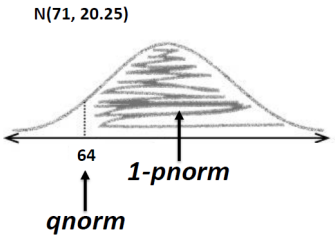
Step 3: Get the probability from R

1-pnorm(64, mean=71, sd=sqrt(20.25))
or
1-pnorm(64, 71, 4.5)

Answer: 1-0.0599 = 94.01%

qnorm(0.0599, 71, 4.5)

Answer: 64



Note: When we calculate pnorm we get area to the left , So we do 1-pnorm, 1-qnorm
pnorm gives us probability, area under the curve. p(z,u,sigma)
qnorm gives us x-axis, quartile, i.e value associated with the area. q(p,u,sigma)

Sampling Distribution of the Means

- The sampling distribution of means is what you get if you consider all possible samples of size *n* taken from the same population and form a distribution of their means.

- Each randomly selected sample is an independent observation.

Expectation and Variance for $\bar{X}$

$E(\bar{X}) = \mu$

Mean of all sample means of size *n* is the mean of the population.

$Var(\bar{X}) = \frac{\sigma^2}{n}$

Standard deviation of $\bar{X}$ tells how far away from the population mean the sample mean is likely to be. It is called the **Standard Error of the Mean** and is given by

$Standard\ Error\ of\ the\ Mean = \frac{\sigma}{\sqrt{n}}$

$\bar{X}$ ~ N(μ, σ^2/n)

Example:

- Let us assume it is a sample from infinite data
- So, if we take many such samples of large sample size (>30 as a thumb rule), the mean values, $\bar{x}$, will be hovering close to the population mean, μ, with a standard deviation, $s = \frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation and *n* is the sample size.

Margin of error- difference between actual population and sample.
Note: sample size increases margin of error decreases. So we can infer better.

Margin of error = Z * SE

Confidence Intervals

- $n = 44$
- $\bar{x} = 10.455$
- $\sigma = 7.7$

$z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$ or Margin of error = $z * \frac{\sigma}{\sqrt{n}}$

∴ Confidence Interval for the Population Mean is

Sample Mean $\pm$ Margin of Error

Note:
Confidence interval gives us the range in which the value can fall.
So we may not be 100% correct, to be sure we take + and - margin of error in our prediction, means our prediction can be deviated + and - margin of error then what we predicted.

How to decide the sample size:

1. Type1

As a rule of them, we multiply independent and dependent variables and then multiply that number with min 75 and max 100 to get the minimum and maximum sample size required.
Note: Do one hot encoding before you multiply as they may contain classes.

2. Type2 - Take 10% of the population.
3. Type3 - Take 10,000 sample sizes as it's more than enough to infer about the population.

Things to remember:

Level of confidence	Value of z
90%	1.64
95%	1.96
99%	2.58

Population Parameter	Population Distribution	Conditions	Confidence Interval
μ	Normal	You know σ^2 n is large or small $\bar{X}$ is the sample mean	$(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}})$
μ	Non-normal	You know σ^2 n is large (> 30) $\bar{X}$ is the sample mean	$(\bar{X} - z\frac{\sigma}{\sqrt{n}}, \bar{X} + z\frac{\sigma}{\sqrt{n}})$
μ	Normal or Non-normal	You don't know σ^2 n is large (> 30) $\bar{X}$ is the sample mean s^2 is the sample variance	$(\bar{X} - z\frac{s}{\sqrt{n}}, \bar{X} + z\frac{s}{\sqrt{n}})$
p	Binomial	n is large p_s is the sample proportion q_s is $1 - p_s$	$(p_s - z\sqrt{\frac{p_s q_s}{n}}, p_s + z\sqrt{\frac{p_s q_s}{n}})$

Note: Neural Network may give you good accuracy but it is a black box, you won't be able to statistically understand what's happening which variables are playing a key role to detect the target. On the other hand linear and logistic gives you statistical reports so you can infer more. Accuracy is not important but inference is for cost cutting and optimization in business organization.

INFERENTIAL STATISTICS

HYPOTHESIS TESTS

What is hypothesis :

It is a proposed explanation made on the basis of limited evidence as a starting point for further investigation.

Steps to be followed for hypothesis testing.

1. Decide the null and alternate hypothesis.
   The null hypothesis is we accept the given claim/statement.
   Alternate will be that we don't accept such.

Note: We start with a null hypothesis is true.

Example:

Dr. Unsnora prescribes SnoreCull to 15 of her patients and records whether it cured them or not after 2 weeks. She found that 11 were cured and 4 were not.

If the drug maker claimed that 90% get cured, 13.5 or 14 patients should have been cured. Is the company making false claims or is the doctor's sampling biased?

Step 1: Decide on the hypothesis

SnoreCull cures 90% of the patients within 2 weeks.

This is called Null Hypothesis and is represented by H_0.

In this case, H_0: $p = 0.9$

If Null Hypothesis is rejected based on evidence, an Alternate Hypothesis, H_1, needs to be accepted. **We always start with the assumption that Null Hypothesis is true.**

In this case, H_1: $p < 0.9$

Step 2: Choose your statistic

$X \sim B(15, 0.9)$

In this case it is a binomial distribution.

Significance level (alpha) = 1 - Confidence Interval

Step 3: Specify the Significance Level

First, we must decide on the Significance Level, α. It is a measure of how unlikely you want the results of the sample to be before you reject the null hypothesis, H_0

Step 4: Determine the critical region

If X represents the number of snorers cured, the critical region is defined as $P(X < c) < \alpha$ where $\alpha = 5\%$.

Recall that in a 95% CI, there is a 5% chance that the sample will not contain the population mean. Hence if the sample falls in the critical region, the null hypothesis that 90% snorers are cured, is rejected.

That is the reason 5% or 0.05 is called the Significance Level. In a 99% CI, 0.01 is the Significance Level.

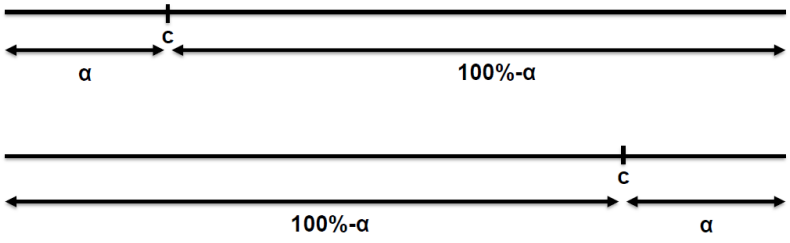
Step 5: Find the *p*-value

If value in the critical region then we fail the null hypothesis stating that the claim made is False and we accept the alternate hypothesis.
Or else if it's not in a critical region we accept the claim as null hypothesis is true.

The critical region can be single sided or two sided.

One-tailed tests

The position of the tail is dependent on H_1.
If H_1 includes a < sign, then the **lower tail** is used.
If H_1 includes a > sign, then the **upper tail** is used.

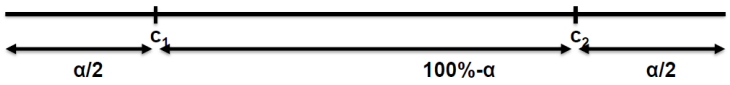


Lower tail - left sided
Upper tail - right sided

Two-tailed tests

Critical region is split over both ends. Both ends contain α/2, making a total of α.

If H_1 includes a ≠ sign, then the two-tailed test is used as we then look for a change in parameter, rather than an increase or a decrease.



Confusion matrix in hypothesis testing:
Helps to decide the false positive, true positive, true negative and true positive.
Actual (null and alternate) vs predicted (null and alternate)

Common Test Statistics for Inferential Techniques

Inferential techniques (Confidence Intervals and Hypothesis Testing) most commonly use 4 test statistics:

- z
- t
- χ^2 (Chi-squared)
- F

Closely related to Sampling Distribution of **Means**

- Closely related to Sampling Distribution of **Variances**
- Derived from Normal Distribution

t-Distribution

If the sample size is small (<30), the variance of the population is not adequately captured by the variance of the sample. Instead of z-distribution, t-distribution is used. It is also the appropriate distribution to be used when <u>population variance is **not known**</u>, irrespective of sample size.

$$t\ statistic\ (or\ t\ score), t = \frac{(\bar{x} - \mu)}{\frac{s}{\sqrt{n}}}$$

Degrees of freedom, ν: # of independent observations for a source of variation minus the number of independent parameters estimated in computing the variation.*

We will be subtracting n by 1 while calculating std deviation. As while calculating we variance there is a constraint present. **(n-1)**

$$\bar{x} - t_{(\frac{\alpha}{2},\nu)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(\frac{\alpha}{2},\nu)} \frac{s}{\sqrt{n}}$$

Where alpha is level of significance , as its two sided so alpha divided by 2, and v is the new , which is degree of freedom. In this case the degree of freedom is n-1.

t-Distribution - Example

Mean, $\bar{x}$ = 101.24 mg

Standard deviation, s = 2.48

n = 10

$\nu = 10 - 1 = 9$

At 95% level, $\alpha = 0.05$, and $\therefore, \frac{\alpha}{2} = 0.025$

R: qt(0.025,9) -> $t_{critical}$ = -2.262

Mean, $\bar{x}$ = 101.24 mg, Standard deviation, s = 2.48

n = 10, $\nu = 10 - 1 = 9$

$$\bar{x} - t_{(\frac{\alpha}{2},\nu)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{(\frac{\alpha}{2},\nu)} \frac{s}{\sqrt{n}}$$

$$101.24 - 2.262 * \frac{2.48}{\sqrt{10}} \leq \mu \leq 101.24 + 2.262 * \frac{2.48}{\sqrt{10}}$$

$$99.47 \leq \mu \leq 103.01$$

Two-Sample t-Test for Unpaired Data

$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$

Test statistic, $t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$

Assuming the two samples come from populations with the **same standard deviation** (Rule of thumb: The ratio between the higher *s* and the lower *s* is less than 2), pooled variance can be used to calculate SE. This is called the **Student's t-test with pooled variance**.

If the ratio is more than 2 then we go for Welch's t-test

2. **For paired data (both data have same properties)**

In unpaired t-test, **difference in means** is studied. In paired t-test, **mean of the differences** is studied.

We calculate the difference first then take its mean and std dev and total observation.

Chi square (X^2) test:

It is a one side test: rightly skewed

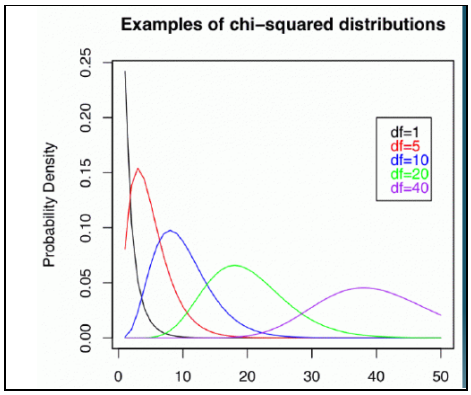
It is used to check variance:

What is expected and what you observed, so we get the difference i.e variance.

$X^2 = \sum \frac{(O-E)^2}{E}$, where O is the observed frequency and E the expected frequency.

Degree of freedom: In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary.

df= n-1 , where one value will be restricted as with help of that the sum of deviation ends up to zero.



Using df, Chi square determines the skewness and calculates accordingly.

$$Z^2 = \frac{(X-\mu)^2}{\sigma^2}$$
$$Z^2 = \chi^2_{(1)}$$

Properties of chi square:

- A X^2 random variable takes values between 0 and ∞.
- Mean of a χ^2 distribution is ν.
- Variance of a χ^2 distribution is 2ν.
- The shape of the distribution is skewed to the right.
- As ν increases, Mean gets larger and the distribution spreads wider.
- As ν increases, distribution tends to normal.

χ^2 goodness of fit works for any probability distribution

Distribution	Condition	ν
Binomial	You know p (probability of success or the proportion of successes in a population)	$\nu = n - 1$
	You don't know p and have to estimate it from observed frequencies	$\nu = n - 2$
Poisson	You know λ	$\nu = n - 1$
	You don't know λ, and have to estimate it from observed frequencies	$\nu = n - 2$
Normal	You know μ and σ^2	$\nu = n - 1$
	You don't know μ and σ^2, and have to estimate them from observed frequencies	$\nu = n - 3$

Use of chi square test:

- To test **goodness of fit**.
- To test **independence** of two variables.
- To test hypothesis about **variance** of a population.

Use for single sample:

Alternate formula for chi square, to calculate variance.

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

F Distribution: use for 2 sample

- χ^2 was useful in testing hypotheses about a single population variance.
- Sometimes we want to test hypotheses about difference in variances of two populations:
  - Is the variance of 2 stocks the same?
- *Recall* $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$. So, F is also a ratio of 2 chi-squares, each divided by its degrees of freedom, i.e.,

$$F = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

Note: whichever sample has higher values that goes to the numerator.

Application of F-Test

- Test for equality of variances.
- Test for differences of means in ANOVA.
- Test for regression models (slopes relating one continuous variable to another, e.g., Entrance exam scores and GPA)

ANOVA

The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.

Example:
u1=u2=u3, significance level (alpha) = 0.10

Group 1			Group 2			Group 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56$ $\bar{X}_2 = 5.56$ $\bar{X}_3 = 5.78$

$$\bar{\bar{X}} = \frac{134}{27} = 4.96$$

$$Total\ Sum\ of\ Squares, SST$$
$$= (2-4.96)^2 + 5*(3-4.96)^2 + 4*(4-4.96)^2 + 7*(5-4.96)^2 + 5*(6-4.96)^2$$
$$+ 4*(7-4.96)^2 + (8-4.96)^2 = \mathbf{62.96}$$

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{mn-1}$, since we can calculate one member knowing the overall mean.

$$Total\ Sum\ of\ Squares\ Within, SSW$$
$$= (2-3.56)^2 + 4*(3-3.56)^2 + 2*(4-3.56)^2 + 2*(5-3.56)^2 + (3-5.56)^2 + 2*(4-5.56)^2$$
$$+ (5-5.56)^2 + 2*(6-5.56)^2 + 2*(7-5.56)^2 + (8-5.56)^2 + 4*(5-5.78)^2 + 3*(6-5.78)^2$$
$$+ 2*(7-5.78)^2 = \mathbf{36.00}$$

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{m(n-1)}$, since we can calculate one member knowing the group mean.

$$Total\ Sum\ of\ Squares\ Between, SSB$$
$$= 9*(3.56-4.96)^2 + 9*(5.56-4.96)^2 + 9*(5.78-4.96)^2 = \mathbf{26.96}$$

When there are *m* groups, the degrees of freedom are $\boldsymbol{m-1}$.

SST = SSW + SSB

Also, for degrees of freedom, $\boldsymbol{mn-1 = m(n-1) + (m-1)}$

$$F - statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{26.96}{2}}{\frac{36}{24}} = 8.9876$$

If numerator is much bigger than the denominator, it means variation **between** means has bigger impact than variation **within**, thus rejecting the null hypothesis.

The *df* are 2 for numerator and 24 for denominator.

F_c, the critical F-statistic, therefore, is 2.53833. 8.9876 is way higher than this and hence we reject the null hypothesis.

SUMMARY						
Groups	*Count*	*Sum*	*Average*	*Variance*		
Group1	9	32	3.55556	1.027778		
Group2	9	50	5.55556	2.777778		
Group3	9	52	5.77778	0.694444		
ANOVA						
Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	26.963	2	13.4815	8.987654	0.0012	2.5383
Within Groups	36	24	1.5			
Total	62.963	26				