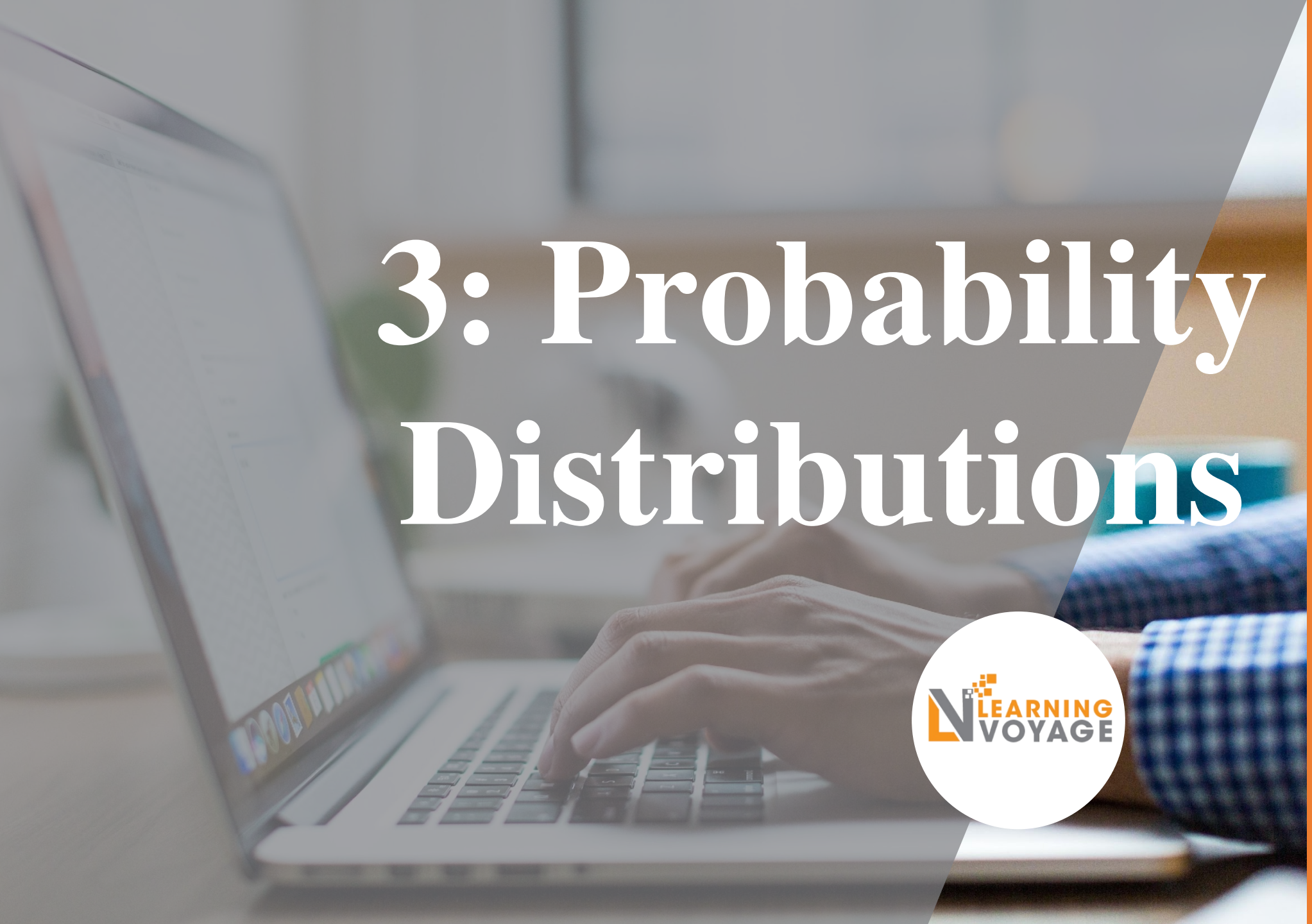
DAY 2

LEARNING VOYAGE



3: Probability Distributions

LEARNING VOYAGE

Probability Distributions

- Probability implies ‘likelihood’ or ‘chance’.
- When an event is certain to happen then the probability of occurrence of that event is 1 and when it is certain that the event cannot happen then the probability of that event is 0.

Assigning Probabilities

- **Classical method** – A prior or Theoretical Probability can be determined prior to conducting any experiment.

$$P(E) = \frac{\#\ of\ outcomes\ in\ which\ the\ event\ occurs}{total\ possible\ \#\ of\ outcomes}$$

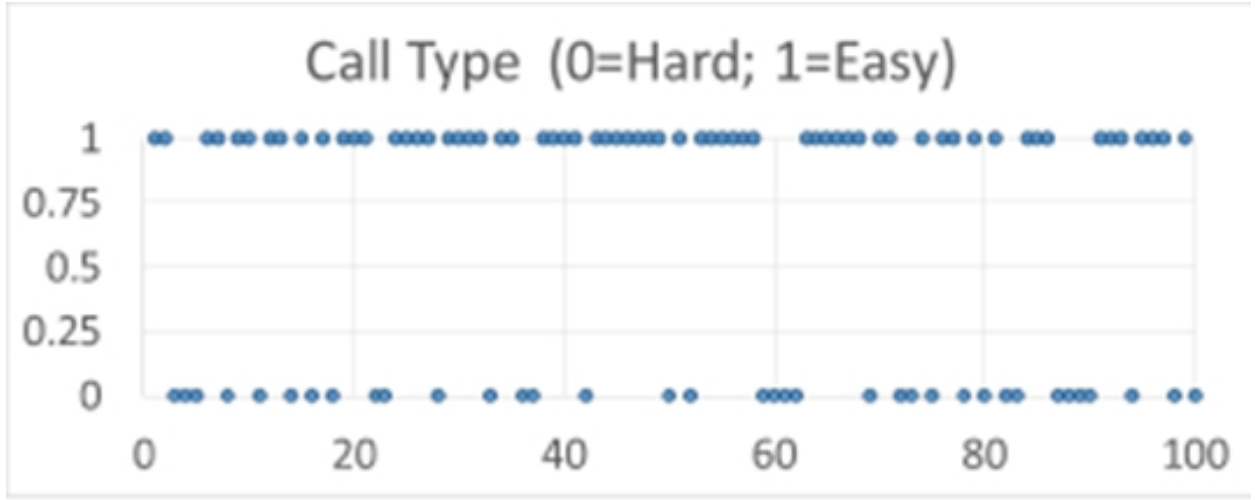
Assigning Probabilities

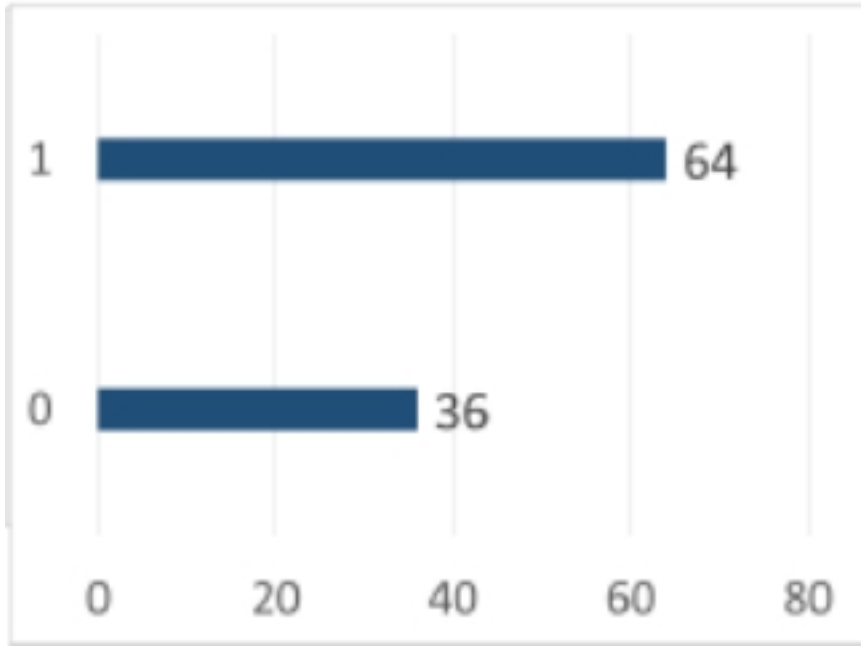
Experiment: Tossing of a fair dice

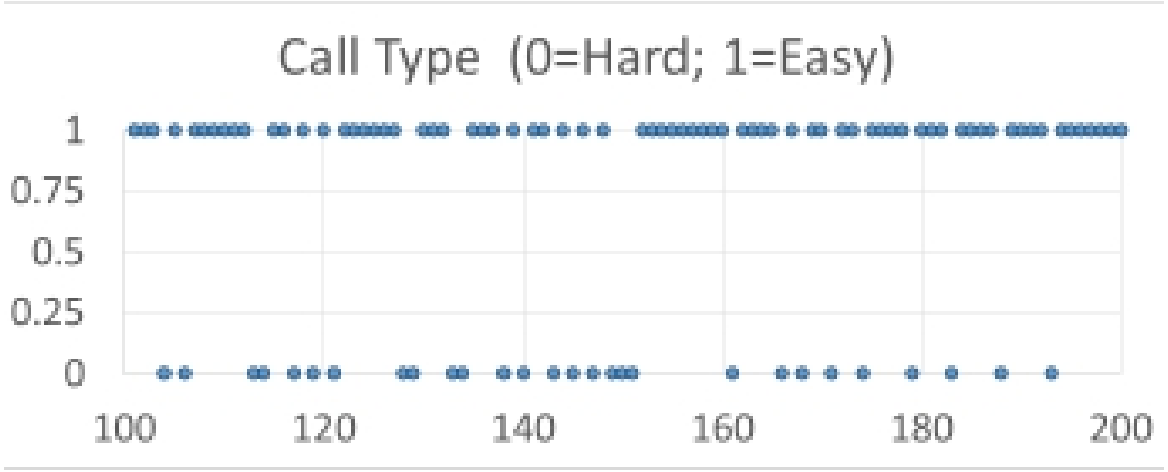
Assigning Probabilities

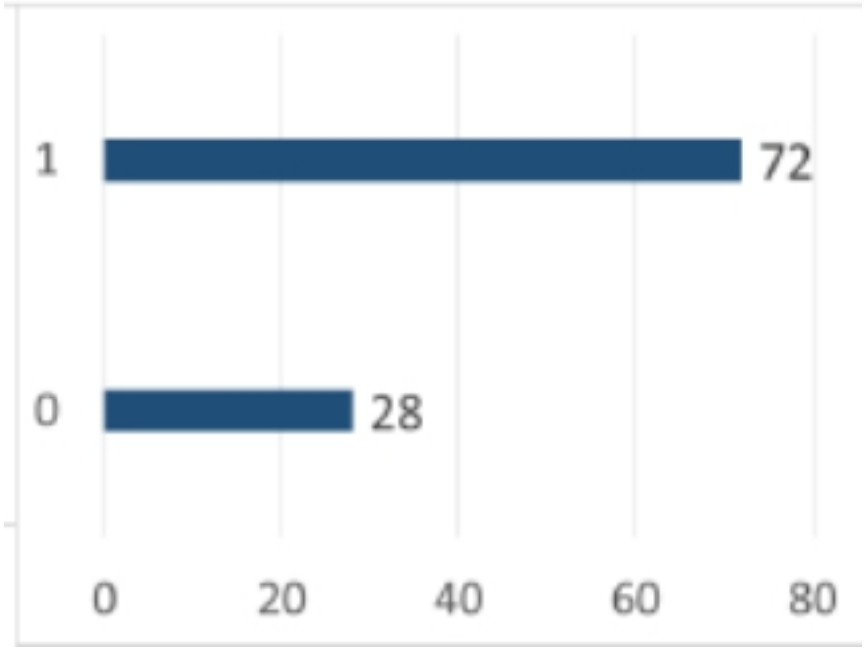
- Empirical Method – A posteriori or Frequentist Probability can be determined post conducting a thought experiment.

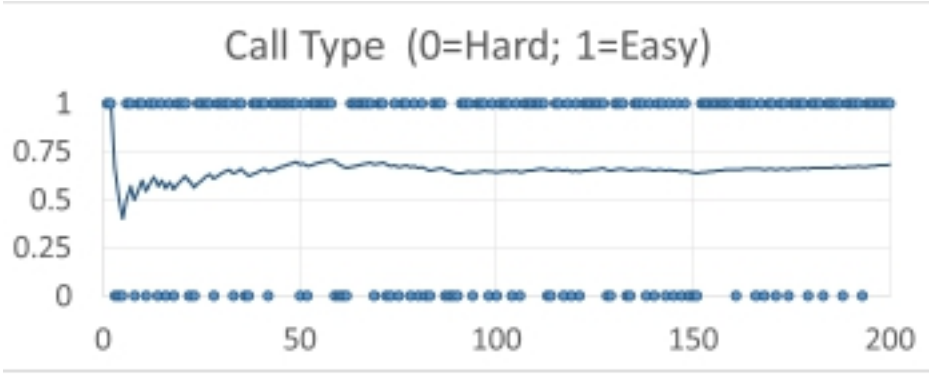
$$P(E) = \frac{\#\ of\ times\ an\ event\ occurred}{total\ \#\ of\ opportunities\ for\ the\ event\ to\ have\ occurred}$$

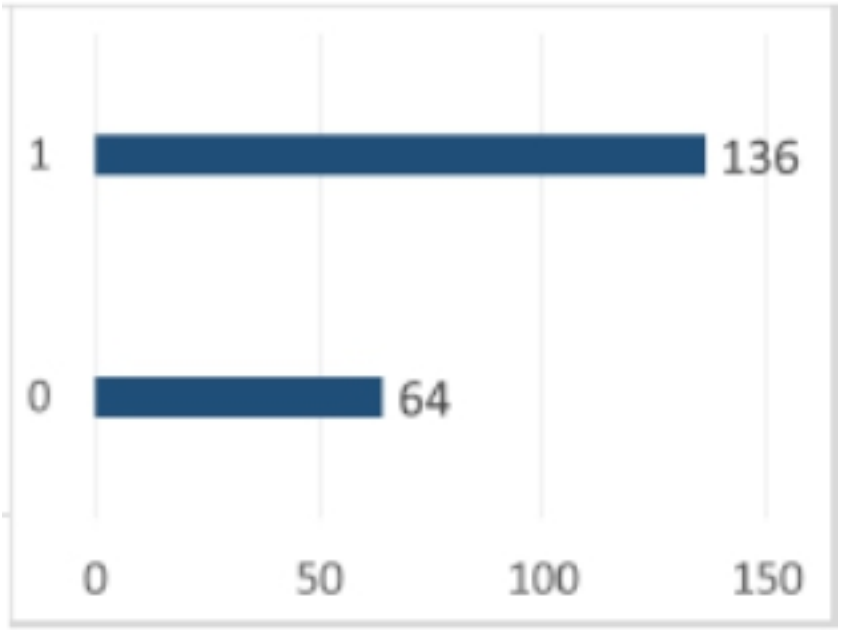

Call Type (0=Hard; 1=Easy)
1
0.75
0.5
0.25
0
0
20
40
60
80
100


1
64
0
36
0
20
40
60
80


Call Type (0=Hard; 1=Easy)
1
0.75
0.5
0.25
0
100
120
140
160
180
200


1
72
0
28
0
20
40
60
80


Call Type (0=Hard; 1=Easy)
1
0.75
0.5
0.25
0
0
50
100
150
200


1
136
0
64
0
50
100
150

P (easy) = 0.7

Probability Terminology

- Sample Space – Set of all possible outcomes, denoted S.

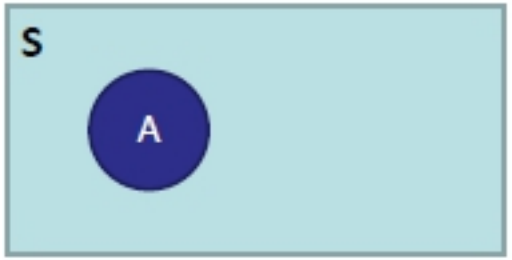
Example: After 2 coin tosses, the set of all possible outcomes are {HH, HT, TH, TT}

- Event – A subset of the samples space.

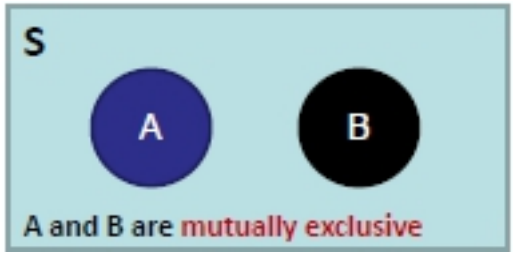
An Event of interest might be – HH

Probability – Rules

P(S) = 1



0 <= P (A) <= 1



P (A or B) = P (A) + P (B)

Mutually Exclusive

When two events (call them “A” and “B”) are Mutually Exclusive than it is impossible for them to happen together.

- If A and B are mutually exclusive

$$P(A \text{ and } B) = 0$$

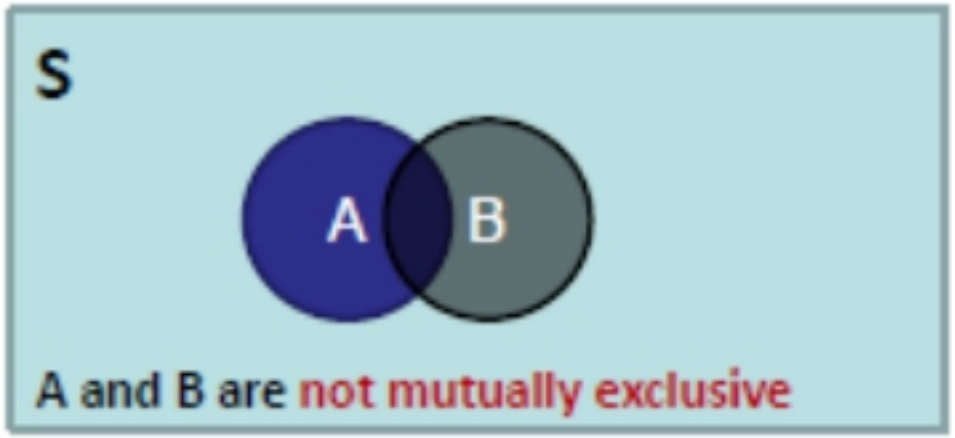
- But the probability of A or B is the sum of the individual probabilities.

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

Mutually Exclusive

- When we combine those two events:

P (King or Queen) = (1/13) + (1/13) = 2/13



P (A or B) = P (A) + P (B) – P (A and B)

Mutually Non-Exclusive Events

- Two events A and B are said to be mutually non-exclusive events if both the events A and B have at least one common outcome between them.

Probability – Types

- Contingency table summarizing 2 variables, Loan Default and Age:

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Probability – Types

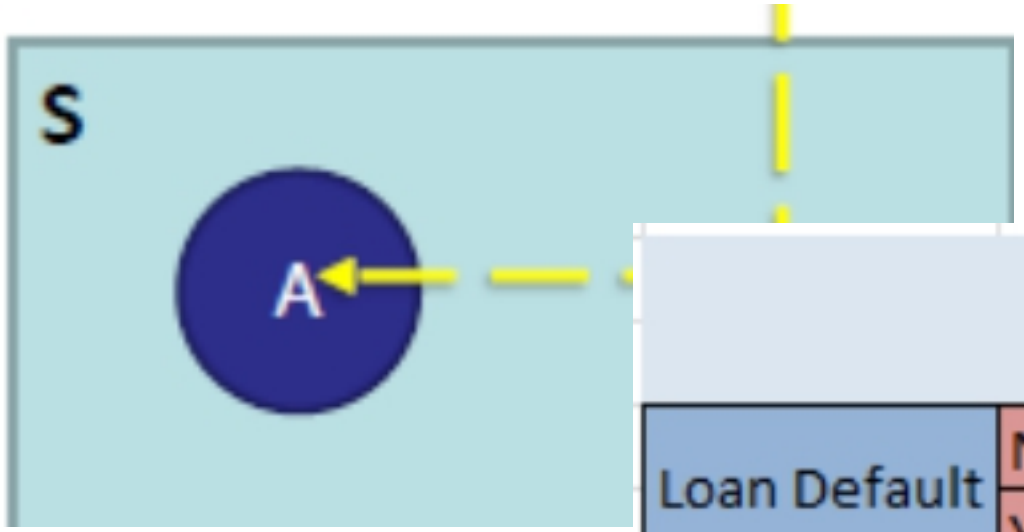
		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

- Convert it into probabilities:

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.00

Probability – Types

- Marginal Probability: Probability describing a single attribute

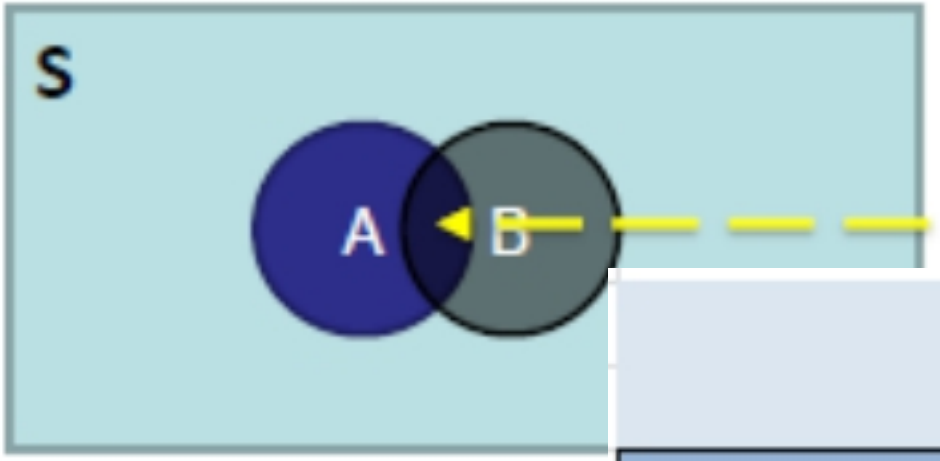


- P (No) = 0.816
- P (Old) = 0.008

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.00

Probability – Types

Joint Probability: Probability describing a combination of attributes



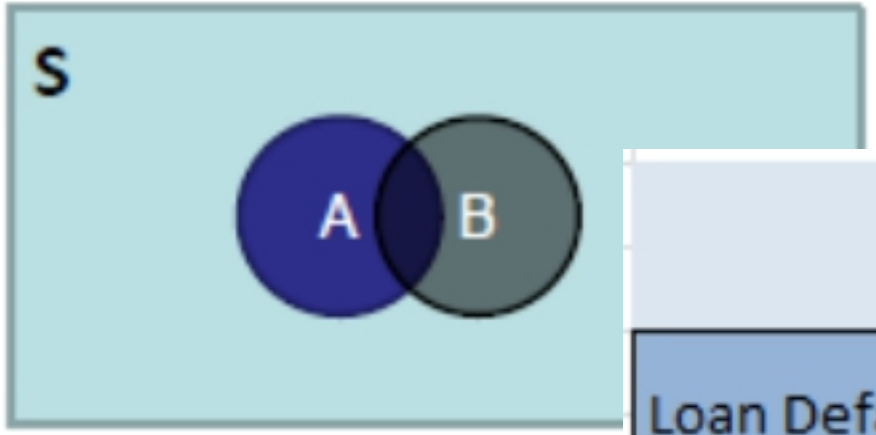
P (Yes and Young) = 0.077

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.00

Probability – Types

- Union Probability: Probability describing a new set that contains all of the elements that are in at least one of the two sets.

P (Yes or Young) = P (Yes) + P (Young) – P (Yes and Young) =
0.184 + 0.302 – 0.077 = 0.409

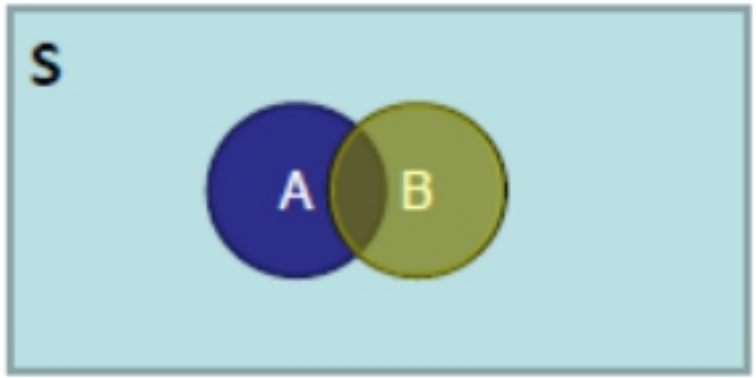


		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	0.586	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	0.690	0.008	1.00

Conditional Probability

The probability of an event (A), given that another event (B) has already occurred.

- The sample space is restricted to a single row or column. This makes the rest of the sample irrelevant.



Example:

What is the probability that a person will not default on the loan payment given he/she is middle-age?

P (No | Middle-Aged) = P (no and middleage) / P (middle age) = 0.586/0.690 = 0.85

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	0.225	**0.586**	0.005	0.816
	Yes	0.077	0.104	0.003	0.184
	Total	0.302	**0.69**	0.008	1.00

Probability – Types

- Note that this is the ratio of Joint Probability to Marginal Probability, i.e.

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

Probability – Types

- Equating, we get

P (A/B) * P (B) = P (A) * P (B/A)

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

- Now, given that the probability that someone defaults on a loan is 0.184, find the probability that an older person defaults on the loan.
- Older people make up only 0.8% of the clientele. P (Yes/Old) =?

$$P(Yes/Old) = (P(Yes) * P(Old/Yes))/P(Old)$$

P (Yes) = 8557/46687 = 0.184
P (Old) = 379/46687 = 0.008
P(old and yes) = 120 / 46687 = 0.003
P (Old/Yes) = P (Old and Yes) / P (Yes) = (120/46687) / (8557/46687) =0.014

P (Yes/Old) = P (old and yes) / p (old) = (120/46687) / (379/46687) = 0.32

The Probability that an older person defaults on the loan is 32%

		Age			
		Young	Middle-aged	Old	Total
Loan Default	No	10,503	27,368	259	38,130
	Yes	3,586	4,851	120	8,557
	Total	14,089	32,219	379	46,687

Histogram:

A series of contiguous rectangles that represent the frequency of data in the given class intervals.

How many class intervals?

- Rule of thumb: 5-15 (not too many and not too few)

The Freedman-diaconis rule:

$$No.\,of\ bins = \frac{(max - min)}{2 * IQR * n^{\frac{-1}{3}}},$$

Histogram – Excel

Passenger Traffic 2013 FINAL (Annual)

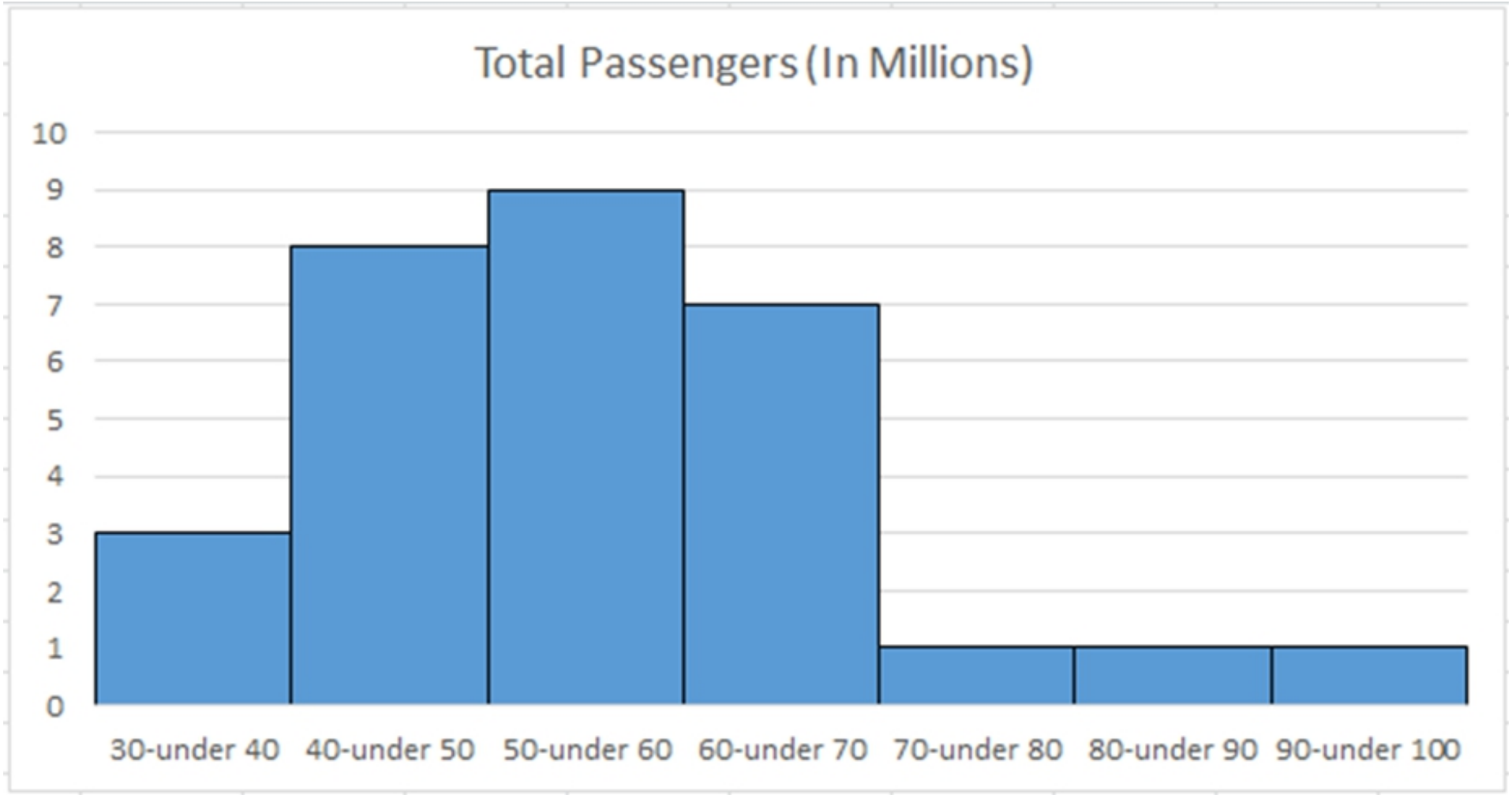
Last Update: 22 December 2014

Passenger Traffic

Total passengers enplaned and deplaned, passengers in transit counted once

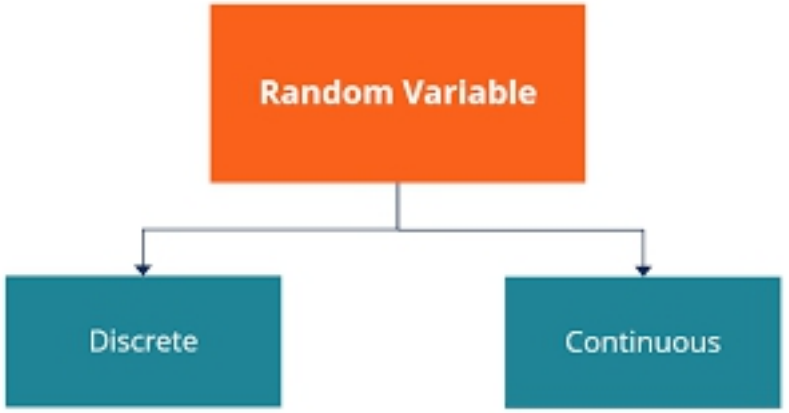
Rank	City (Airport)	Passengers 2013	% Change
1	ATLANTA GA, US (ATL)	94,431,224	-1.1
2	BEIJING, CN (PEK)	83,712,355	2.2
3	LONDON, GB (LHR)	72,368,061	3.3
4	TOKYO, JP (HND)	68,906,509	3.2
5	CHICAGO IL, US (ORD)	66,777,161	0.2
6	LOS ANGELES CA, US (LAX)	66,667,619	4.7
7	DUBAI, AE (DXB)	66,431,533	15.2
8	PARIS, FR (CDG)	62,052,917	0.7
9	DALLAS/FORT WORTH TX, US (DFW)	60,470,507	3.2
10	JAKARTA, ID (CGK)	60,137,347	4.1
11	HONG KONG, HK (HKG)	59,588,081	6.3
12	FRANKFURT, DE (FRA)	58,036,948	0.9
13	SINGAPORE, SG (SIN)	53,726,087	5
14	AMSTERDAM, NL (AMS)	52,569,200	3
15	DENVER CO, US (DEN)	52,556,359	-1.1
16	GUANGZHOU, CN (CAN)	52,450,262	8.6
17	BANGKOK, TH (BKK)	51,363,451	-3.1
18	ISTANBUL, TR (IST)	51,304,654	13.7
19	NEW YORK NY, US (JFK)	50,423,765	2.3

Histogram – Excel



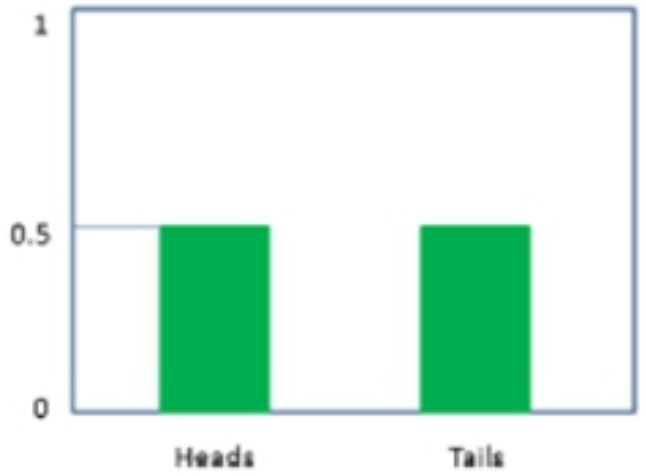
Random Variable

- A Random Variable is a set of possible values from a random experiment.



Discrete Random Variable

- The discrete random variable is a variable that may take on only a countable number of distinct values.



Countable

Probability Distributions

Types of Discrete Probability Distributions

- Bernoulli Distribution.
- Binomial Distribution.
- Poisson Distribution.

Bernoulli distribution

- A Bernoulli distribution is a discrete probability distribution for a Bernoulli trial – a random experiment that has only two outcomes (usually called “Success” or “Failure”.

Binomial Distribution

- A binomial distribution is the probability of the “success” or “failure” outcome of an experiment or survey that is repeated multiple times.
- A binomial distribution is the probability of a “success” or “failure” outcome in an experiment or survey that is repeated multiple times.

Notation: X ~ Bio(n, P)

n: number of times the experiment runs

p: probability of one specific outcome

Probability Mass Function:

$$b(x; n, P) = {}_nC_x \cdot P^x \cdot (1 - P)^{n - x}$$

Where:

- b = binomial probability.
- x = total number of “success”.
- P = probability of success on an individual trail.
- n = number of trails.

Mean and Variance of Binomial distributio

$$\mathbf{E(X) = np}$$
$$\mathbf{Var(X) = npq}$$

Criteria – Binominal distribution must meet the following three criteria:

- The number of trials is fixed.
- Each trail is independent of others.
- The probability of "success" (trail, head, fail or pass) is the same from one trail to another.

Poisson distribution

- The Poisson distribution is the discrete probability distribution of the number of events occurring in a given time period, provided that the events occur at a constant mean rate and are independent of the time since last event.

Probability Mass Function:

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

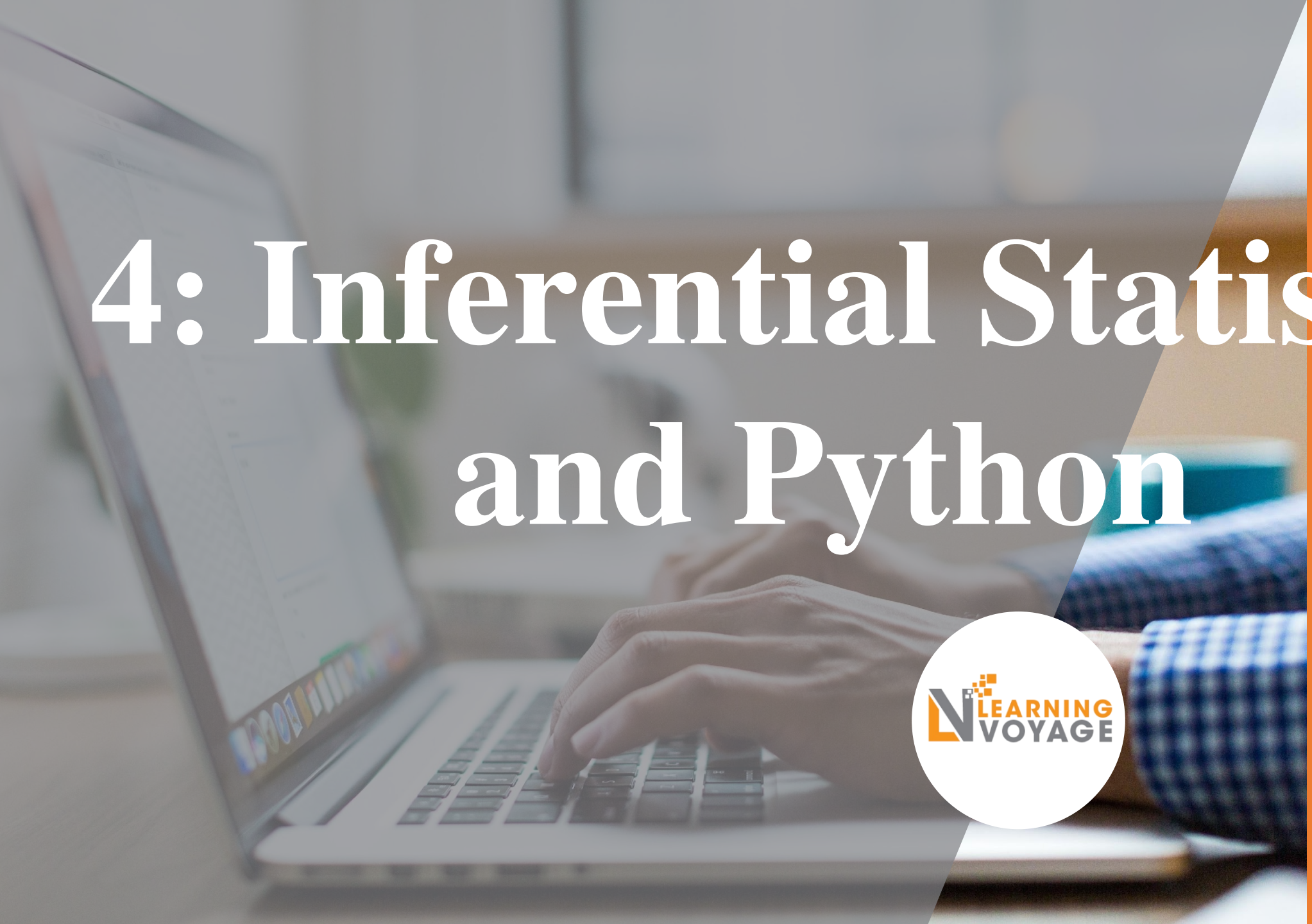
Where:

- The symbol "!" is a factorial.
- M (The expected number of occurrences) is sometimes written as λ. It is sometimes called the event rate or rate parameter.

"Complete Lab 3"

"Complete Case Study"

"Complete Programming Assignment "



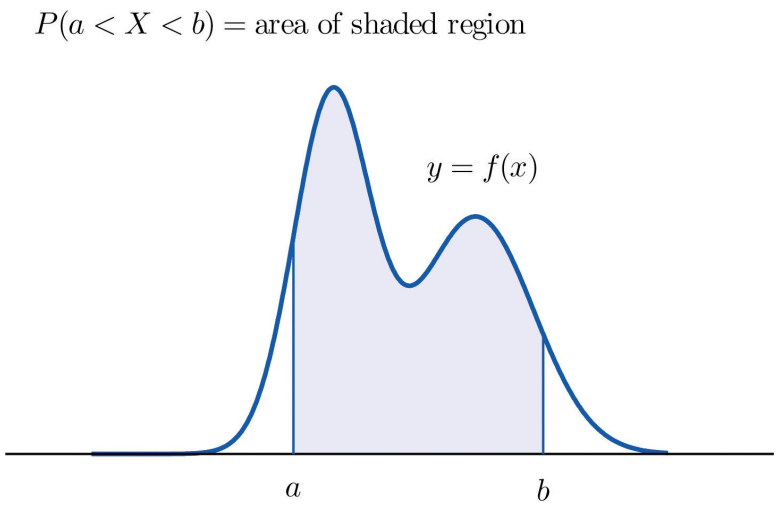
4: Inferential Statistics and Python

LEARNING VOYAGE

Continuous Random Variable

- A continuous random variable is a random variable, in which the data can take infinitely many values.

- Continuous random variables usually are the measurements.



Continuous Random Variable

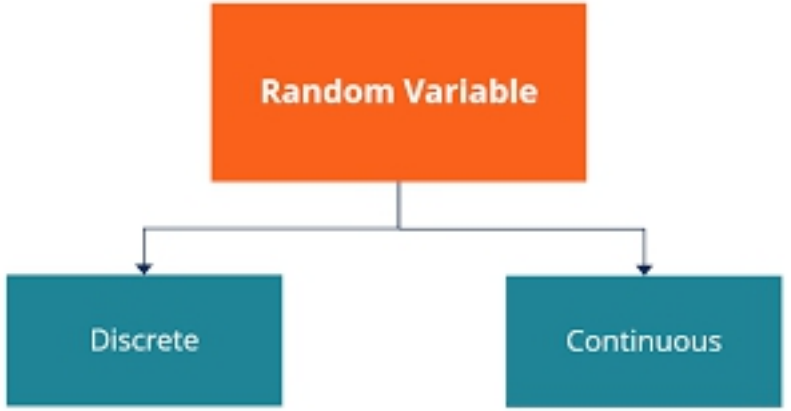
The probability function of the continuous random variable is probability density function (PDF) and represented by the area under a curve.

It is formulated as follows:

- The formula for $E(X)$ is $E(X)=\int_{-\infty}^{\infty} xf(x)dx$.
- The formula for $E(X^2)$ is $E(X^2)=\int_{-\infty}^{\infty} x^2 f(x)dx$.
- Formula for $P(a<X<b)$ is $P(a<X<b)=\int_{a}^{b} f(x)dx$.
- Formula for cumulative distribution function is $F(X)=\int_{-\infty}^{x} f(z)dz$.

Types of Continuous Random Variables

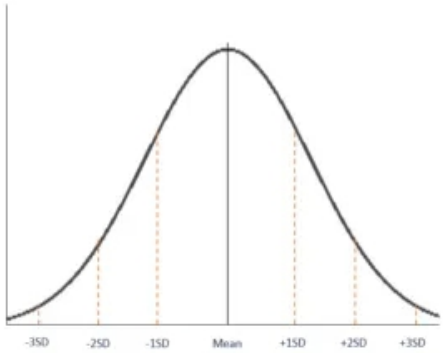
- Normal Distribution.
- Standard Normal Distribution.
- Uniform Distribution / Rectangular Distribution.



Normal Distribution.

- Normal distribution also known as the Gaussian distribution, is the probability distribution which is symmetric about the mean

- Which shows that the data close to the mean are more frequent in occurrence than the data further away from the mean

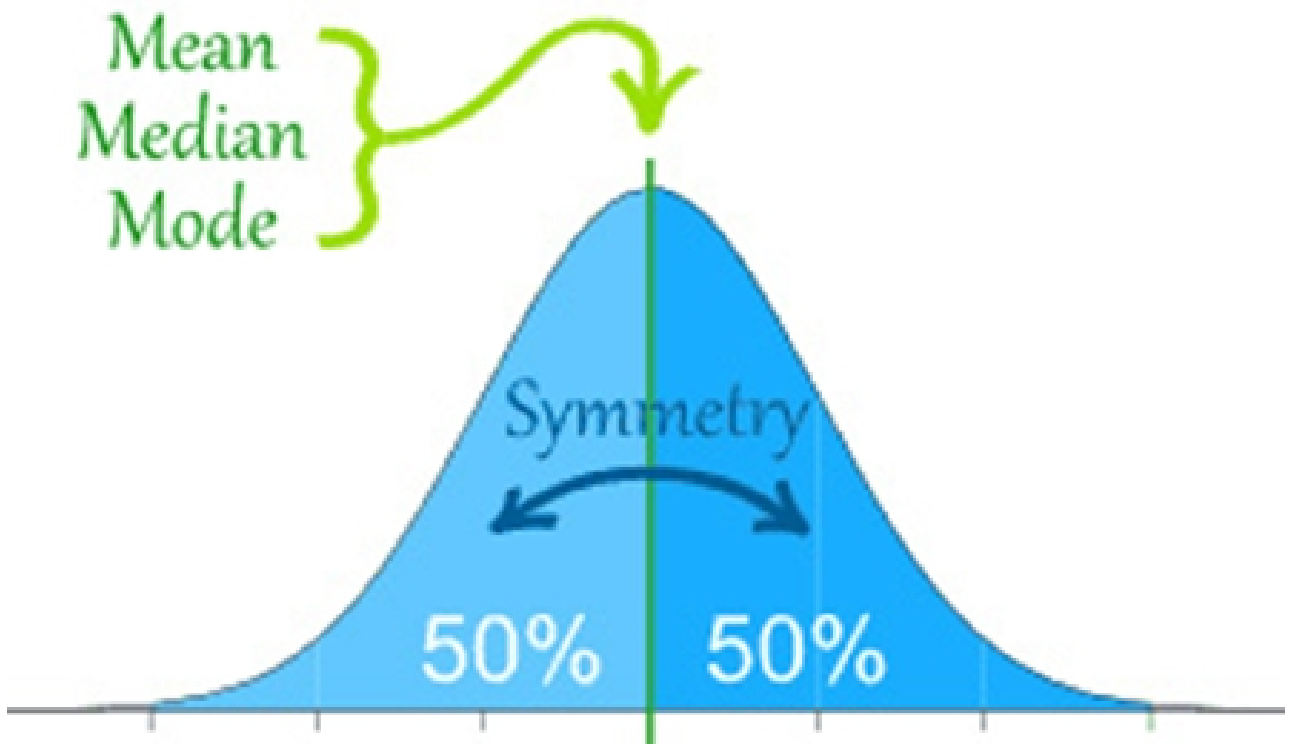


Normal Distribution.

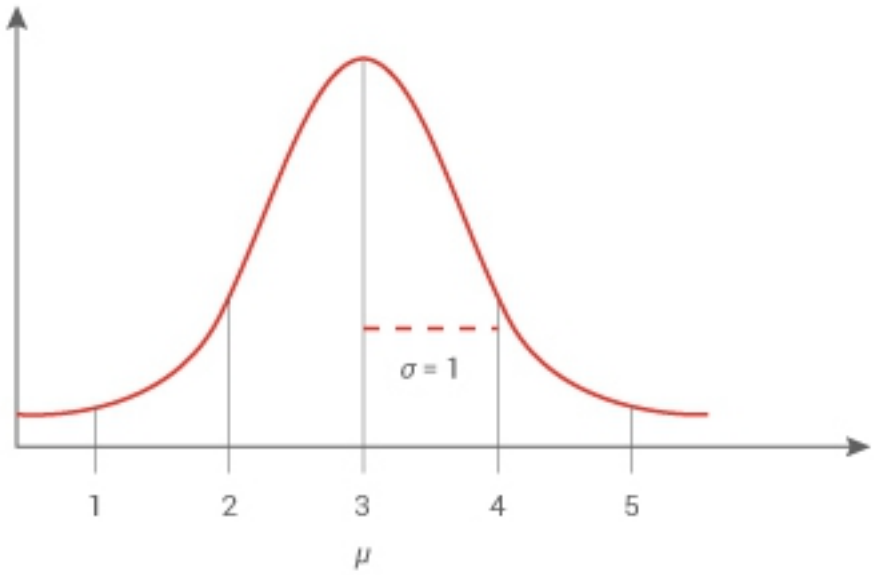
- Probability Density Function

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

Normal Distribution.

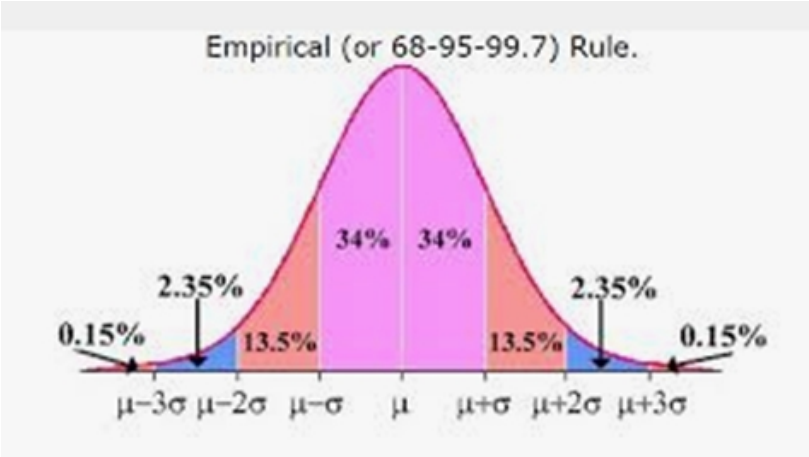


Normal Distribution.

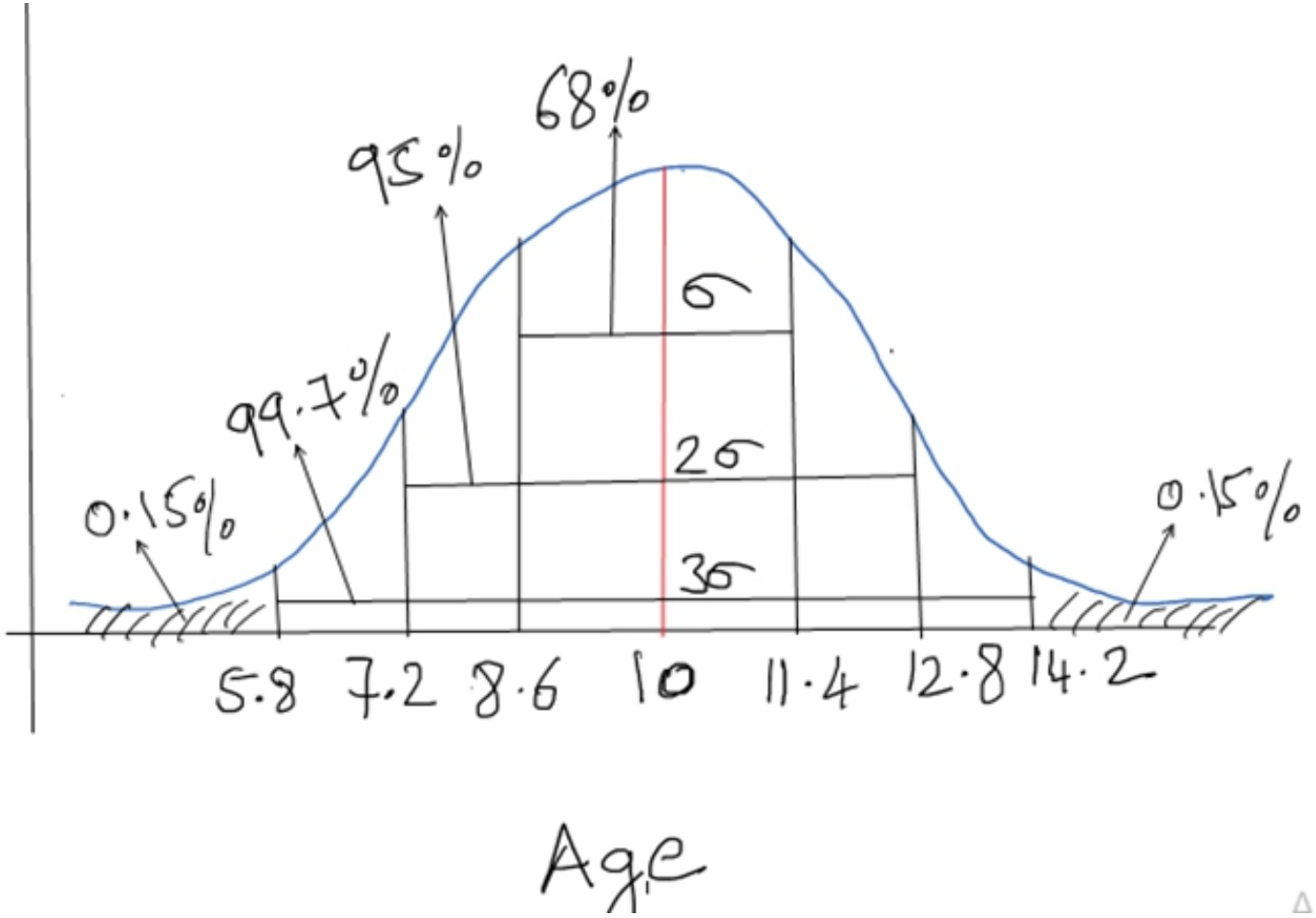


Empirical Formula

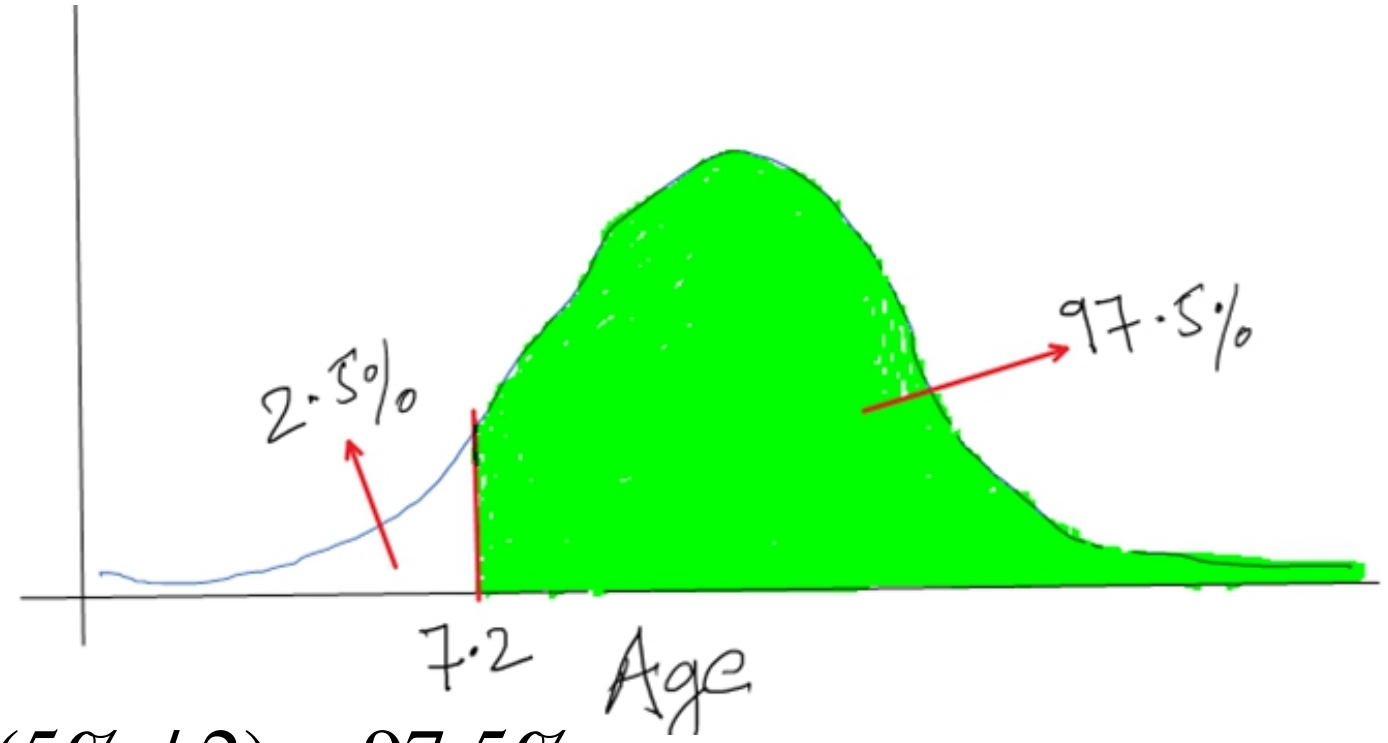
- The empirical rule states that for the Normal Distribution, nearly all observed data sets will fall within 3 standard deviations.
- The empirical rule is also known as Three-Sigma rule or 68-95-99.7 rule.



Empirical Formula



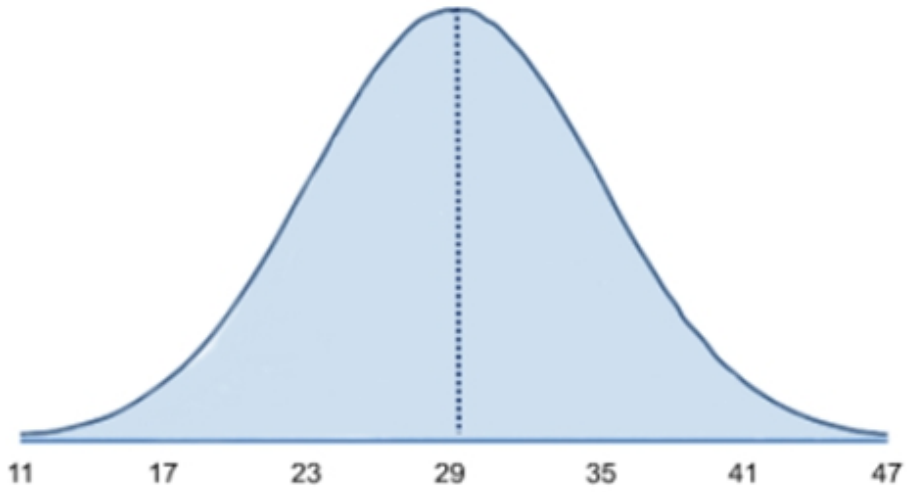
Empirical Formula



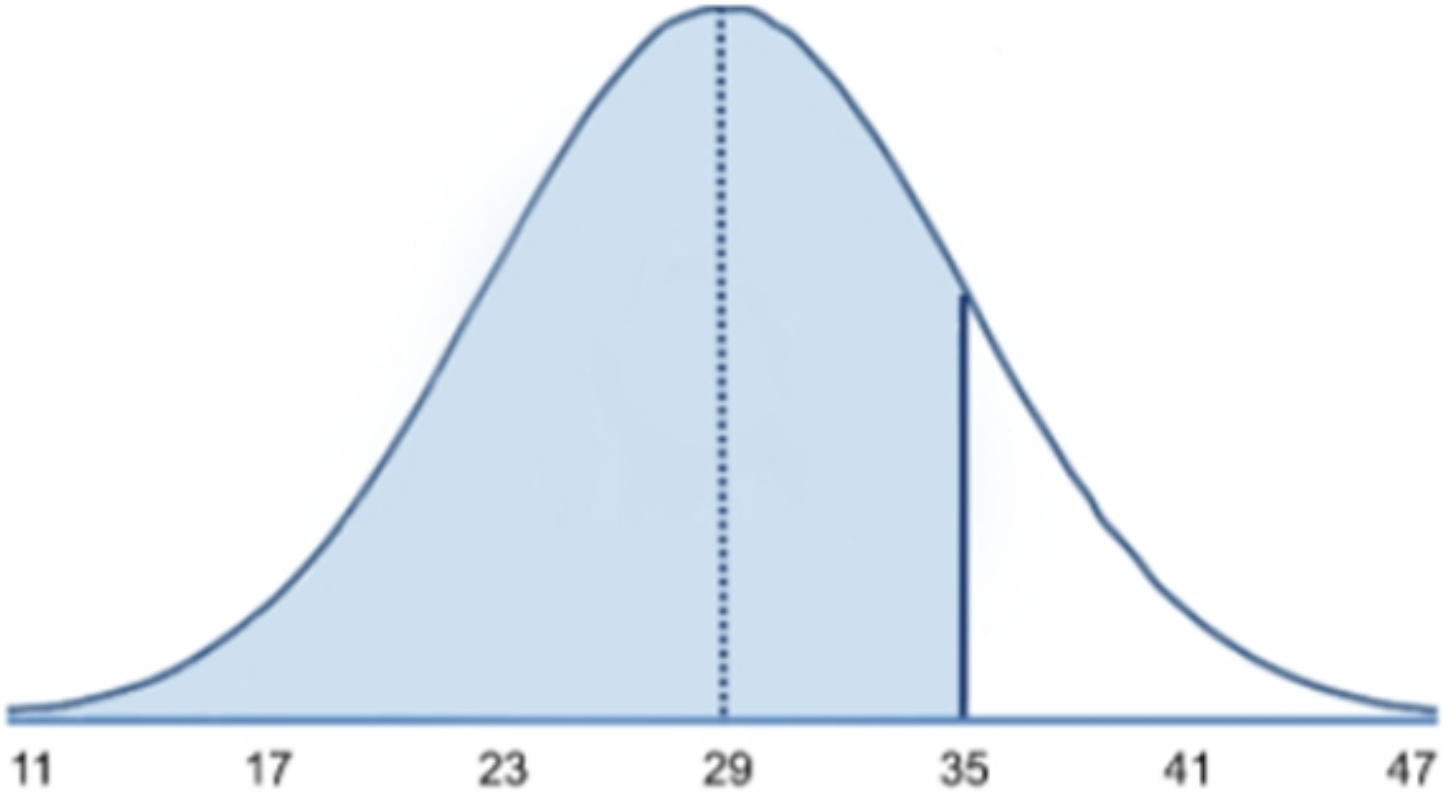
- 95% + (5% / 2) = 97.5%
- Thus, the probability of living more than 7.2 years is 97.5%.

Empirical Formula

- Example 2: Consider the weight values of a population of 15 years old males where the weight is normally distributed and has a mean value =29 and a standard deviation = 6.



Empirical Formula



Standard Normal Distribution

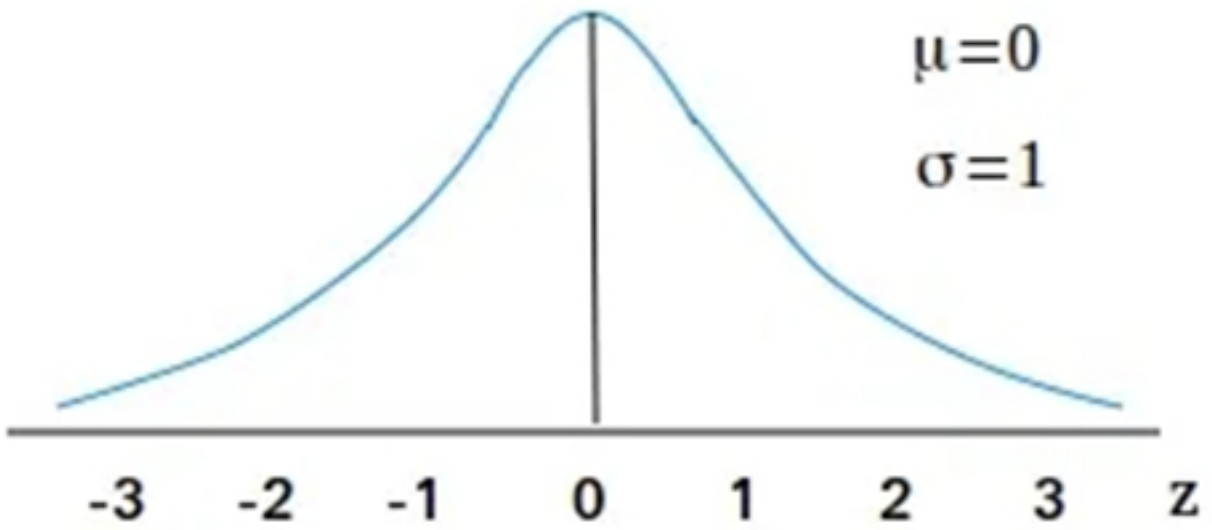
- The standard normal distribution is a special case of the normal distribution.

- It is the distribution that occurs when a normal random variable has the mean value of zero and standard deviation of one

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

Standard Normal Distribution

The standard normal distribution curve



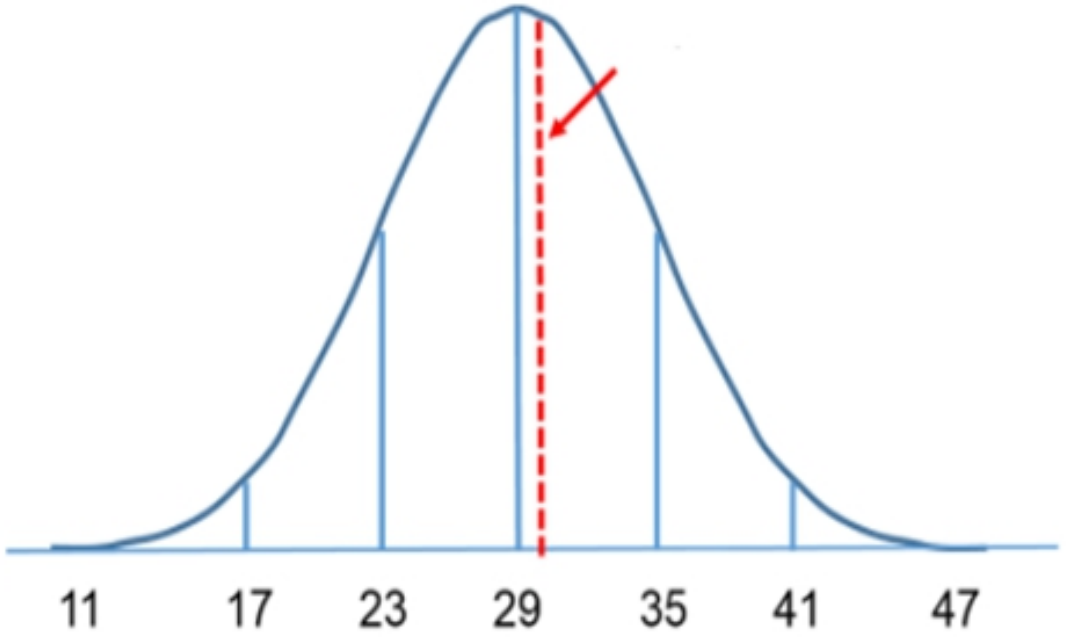
Standard Normal Distribution

Solution: To compute P (X < 30) we convert the X=30 to its corresponding Z score (this is called standardizing):

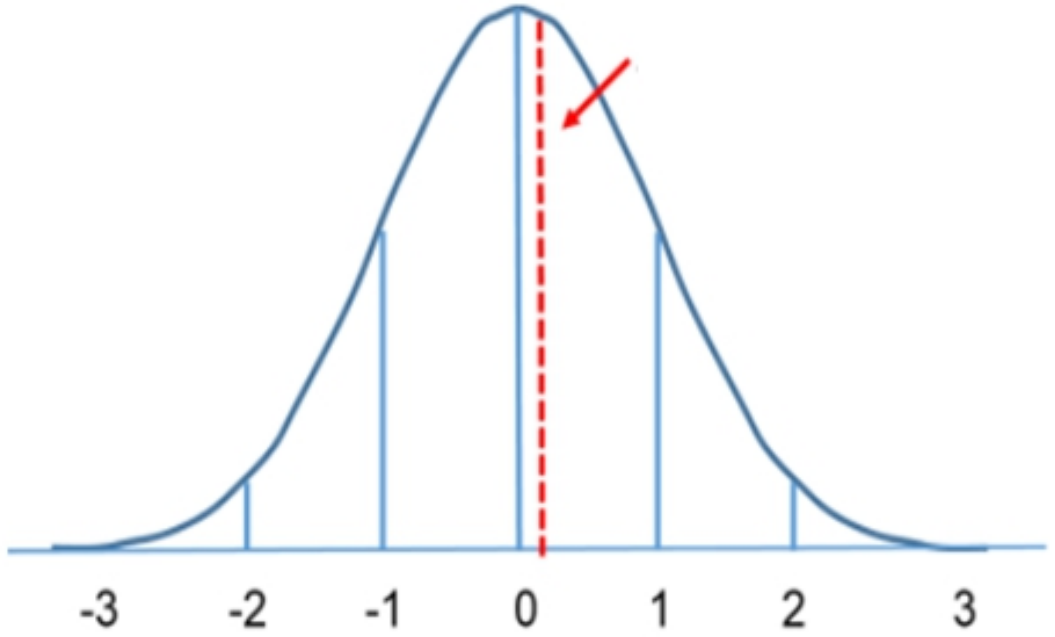
$$Z = 30 - 29 / 6 \Rightarrow 1/6$$

$$Z = 0.1667$$

Standard Normal Distribution



mean= 29; SD=6



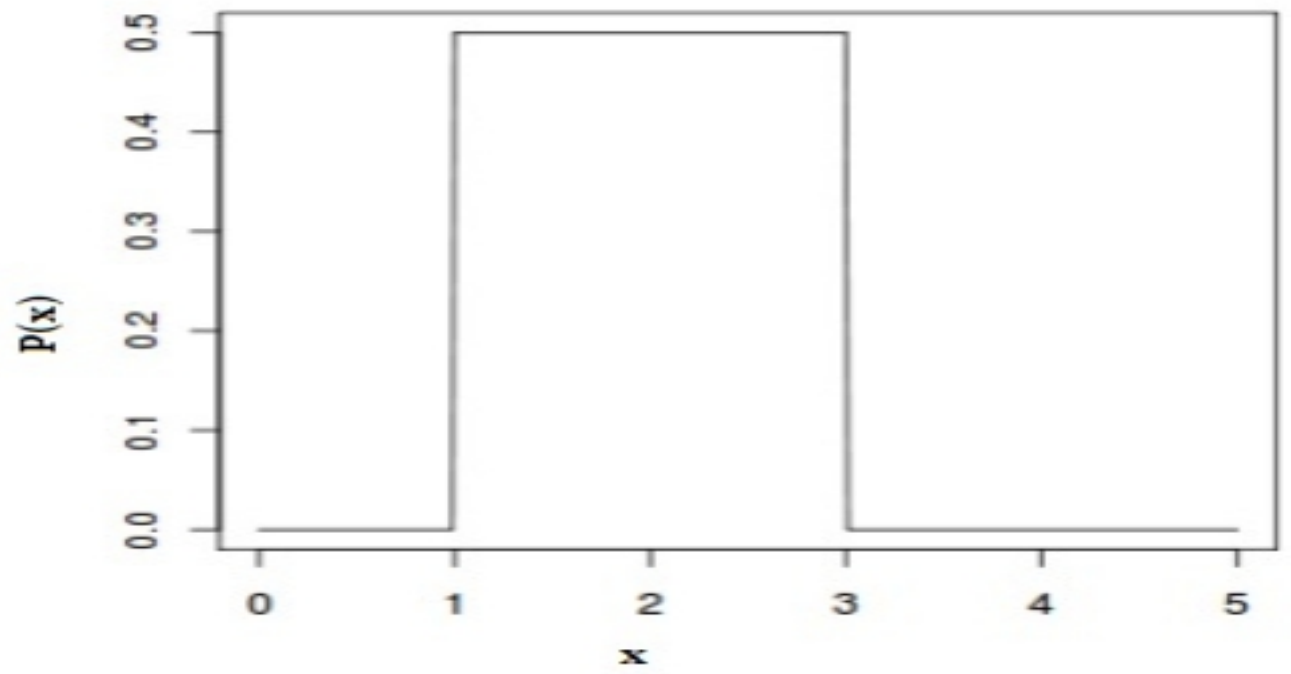
Z-Score

Standard Normal Distribution

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621

Uniform Distribution / Rectangular Distribu

- A uniform distribution, also called a rectangular distribution, is a probability distribution that has a constant probability

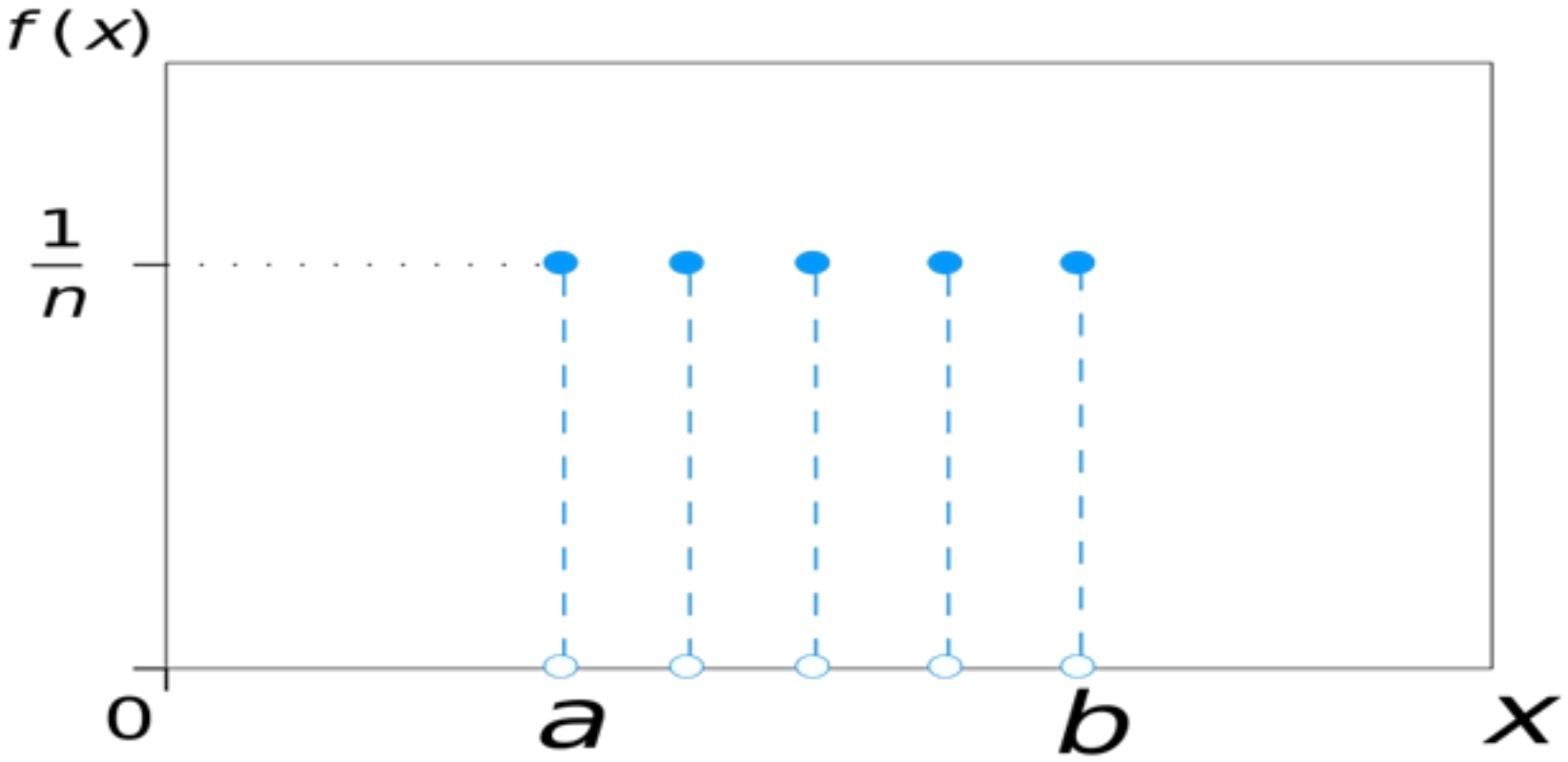


Uniform Distribution / Rectangular Distribu

- Probability density function:

$$f(x) = \frac{1}{b-a}$$

Uniform Distribution / Rectangular Distribu



Continuity Correction Factor:

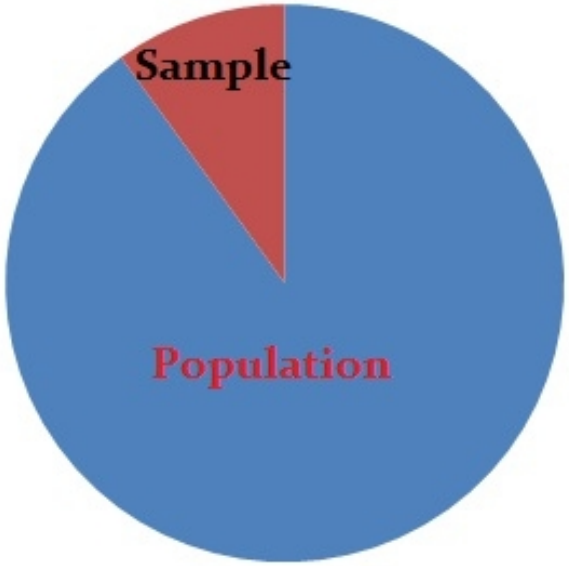
- A continuity correction factor is applied when you use a continuous probability distribution to approximate a discrete probability distribution.

- For example: when you want to use the normal distribution to approximate a binomial distribution, following conditions needs to be satisfied.

$$(n * p) \text{ and } (n * q) >= 5$$

Inferential statistics

- Inferential statistics use a random sample of data from a population to make inference about the whole population.



Central Limit Theorem

- The central limit theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger, irrespective of the shape of the population distribution.

Expectation and Standard deviation of X b

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

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

- Standard deviation tells us how far away from the population mean the sample mean is likely to be and is called the Standard Error of the Mean, and is represented as follows

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

If $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \sigma^2/n)$

Point Estimate

- A point estimate of a population parameter is a single value of a statistic.

- Example: Xbar is a point estimate of the population mean μ. Similarly, the population proportion “p” is a point estimate of the population proportion “P” .

Confidence Intervals of Mean, Proportion, & Varia

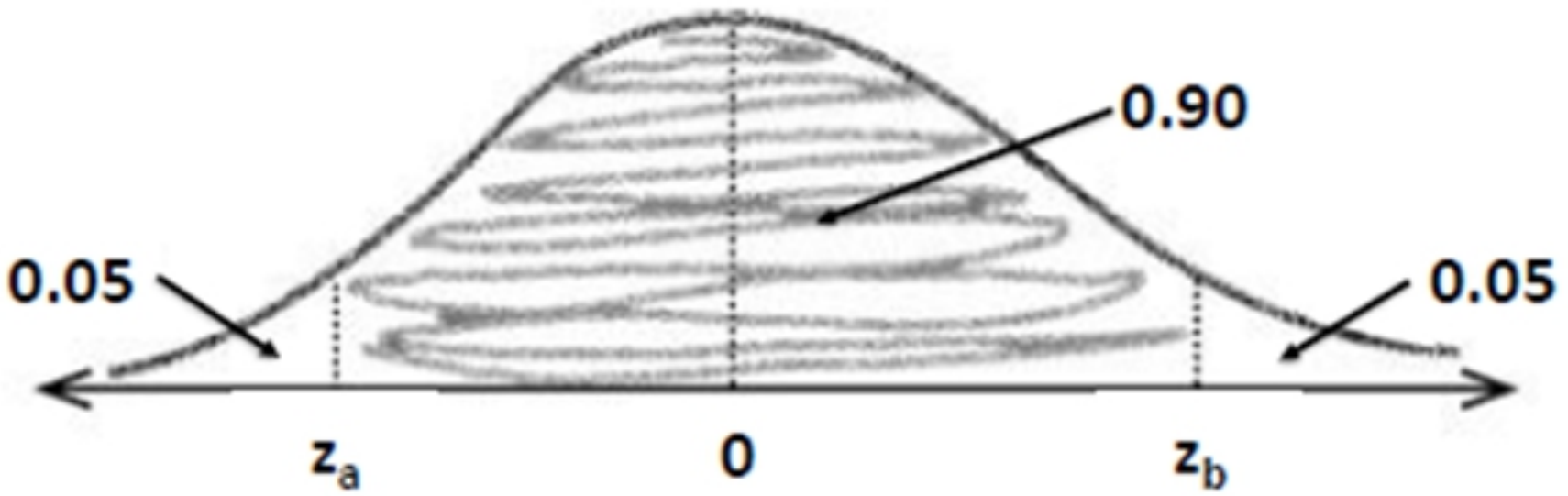
Mean:

$$CI = \bar{x} \pm z\frac{s}{\sqrt{n}}$$

Confidence Intervals of Mean, Proportion, & Varia

Sample mean ± Margin of Error

Find Za and Zb where P (Za < Z < Zb) = 0.90



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455

Proportion:

$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$$CI = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Proportion:

- Example: In a poll by FOX News conducted between November 1 and 3, 2016, a survey of 1107 randomly sampled likely voters predicted that 45% would vote for Hillary Clinton:

Fox News Poll			
	Nov	Oct 22 - 25	Oct 15 - 17
Clinton	45%	44%	45%
Trump	43%	41%	39%
November 1 - 3, 2016			
Likely Voters +/- 3% Pts			

t-Distribution:

- If the sample size is small (<30), the variance of the population is not adequately captured by the variance of the sample.
- In such cases, instead of z-distribution, t-distribution is used.
- It is also the appropriate distribution type to use when the population variance is not known.

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

t-Distribution:

- Confidence Interval to Estimate μ:

$$\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

t-Distribution:

- Data are as follows (in mg):

98.6	102.1	100.7	102	97
103.4	98.9	101.6	102.9	105.2

t-Distribution:

	Table of Critical Values for T Two Tailed Significance					
DF	0.2	0.1	0.05	0.01	0.005	0.001
2	1.89	2.92	4.30	9.92	14.09	31.60
3	1.64	2.35	3.18	5.84	7.45	12.92
4	1.53	2.13	2.78	4.60	5.60	8.61
5	1.48	2.02	2.57	4.03	4.77	6.87
6	1.44	1.94	2.45	3.71	4.32	5.96
7	1.41	1.89	2.36	3.50	4.03	5.41
8	1.40	1.86	2.31	3.36	3.83	5.04
9	1.38	1.83	2.26	3.25	3.69	4.78
10	1.37	1.81	2.23	3.17	3.58	4.59

t-Distribution:

$$\bar{x} - t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1,\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$$

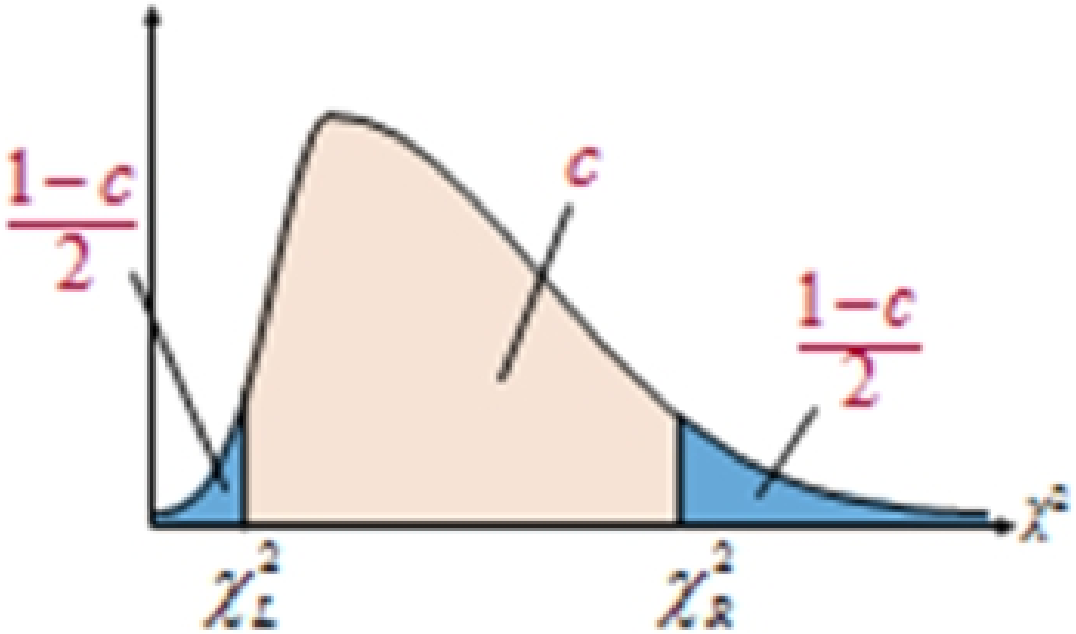
Chi Square Distribution for Variance:

- You can use the Chi Square distribution to construct the confidence interval for the variance and standard deviation.

Formula:

$$C.I. = \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$$

Chi Square Distribution for Variance:



The area between the left and right critical values is c.

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025	0.01	0.005
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336

For Mean

with known variance & n>= 30	Use Z Distn Confidence Interval	$C.I. = \bar{X} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$
with unknown variance or n<30	Use t Dstn Confidence Interval	$C.I. = \bar{X} \pm t_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$
For Variance	Use the χ^2 Distn Confidence Interval	$C.I. = \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}$
For Proportion	Use Z Distn Confidence Interval if $n \times \hat{p}$ and $n \times \hat{q}$ are each >= 5	$C.I. = \hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Test of Hypothesis

- Hypothesis testing is a statistical method for making the statistical decisions about the hypothesis using experimental data.

- Hypothesis Testing is basically makes an assumption about the population parameter and uses a given sample to test whether or not the statistical claims are likely to be true or not.

Test of Hypothesis

Normal distribution is a good approximation,

Standard Error = s / sqrt (n) = 1.0 / sqrt (40)

Standard Error = 0.158

$X \sim N(0.7, 0.158^2) = N(0.7, 0.025)$

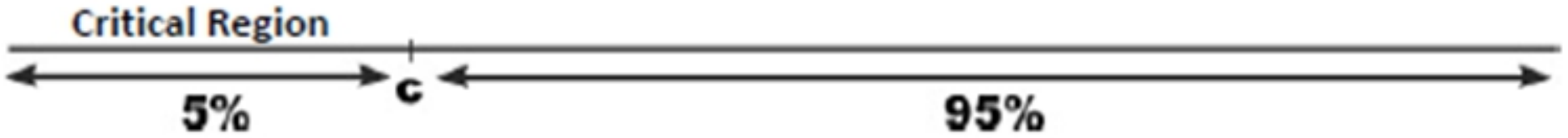
$$Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Z = (0.55 – 0.7) / 0.158

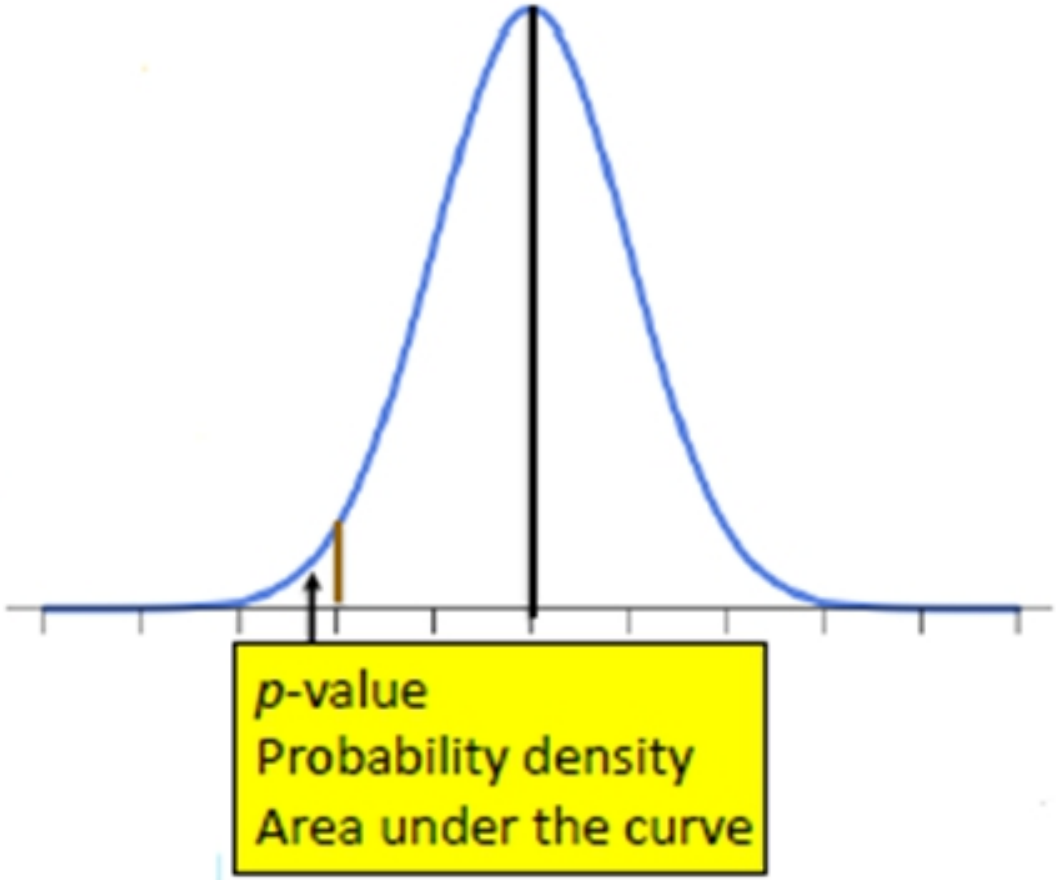
Z = 0.94

Step 4: Determine the critical region.

- If X represents the mean score of the sample, the critical region is defined as $P(X < c) < \alpha$ where $\alpha = 5\%$



Test of Hypothesis



Test of Hypothesis

- If P ($X \leq 0.55$) < 0.05 (Significance Level), it indicates that 0.55 is inside the critical region, and hence H0 can be rejected.

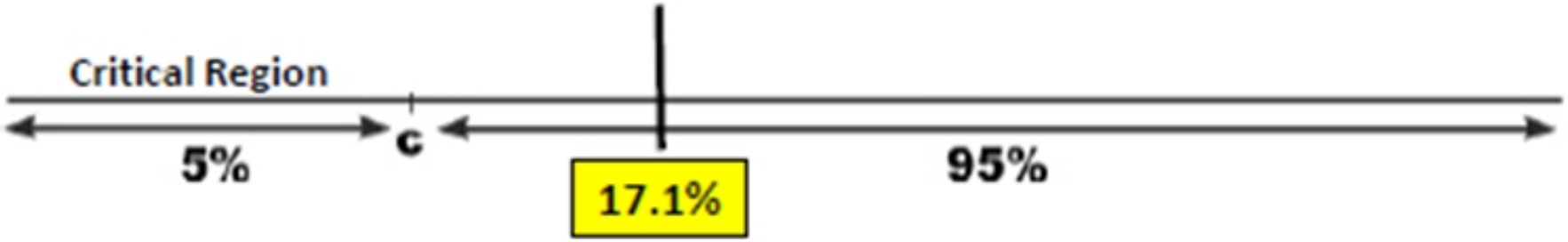
Given that Z = -0.94, P ($X \leq 0.55$) = 0.171

```
> pnorm(0.55,0.7,1/sqrt(40))
[1] 0.1713909
```

- Thus, there is a 17% probability of finding a mean score of 5.5/10 or less.

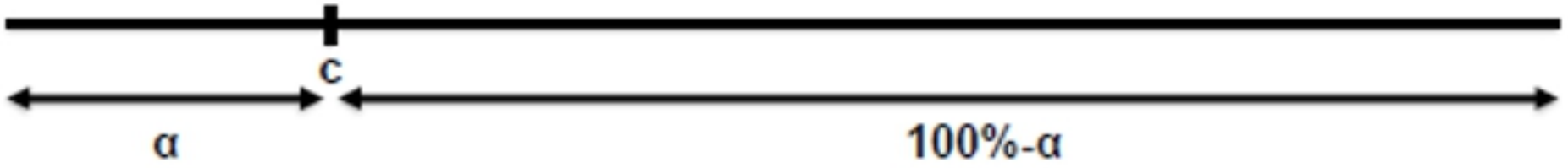
Test of Hypothesis

- Step 6: Is the sample result in the critical region?

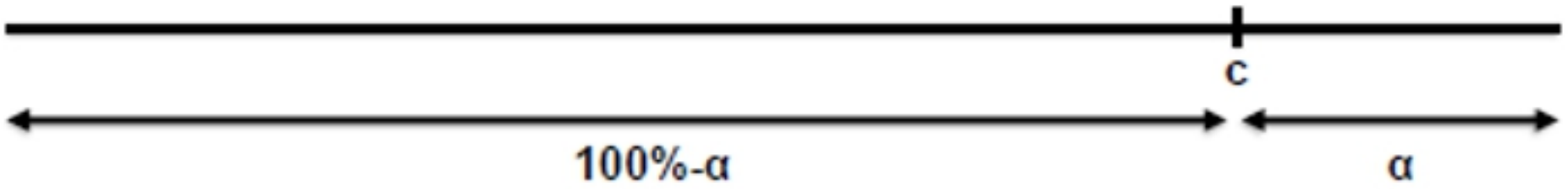


Critical Region up Close

If H_1 includes a **"<" sign (lesser than)**, then the lower tail is utilized.



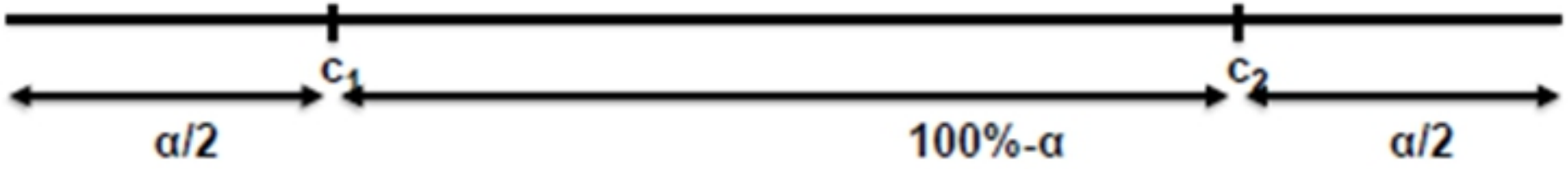
If H_1 includes a **">" sign (greater than)**, then the upper tail is utilized.



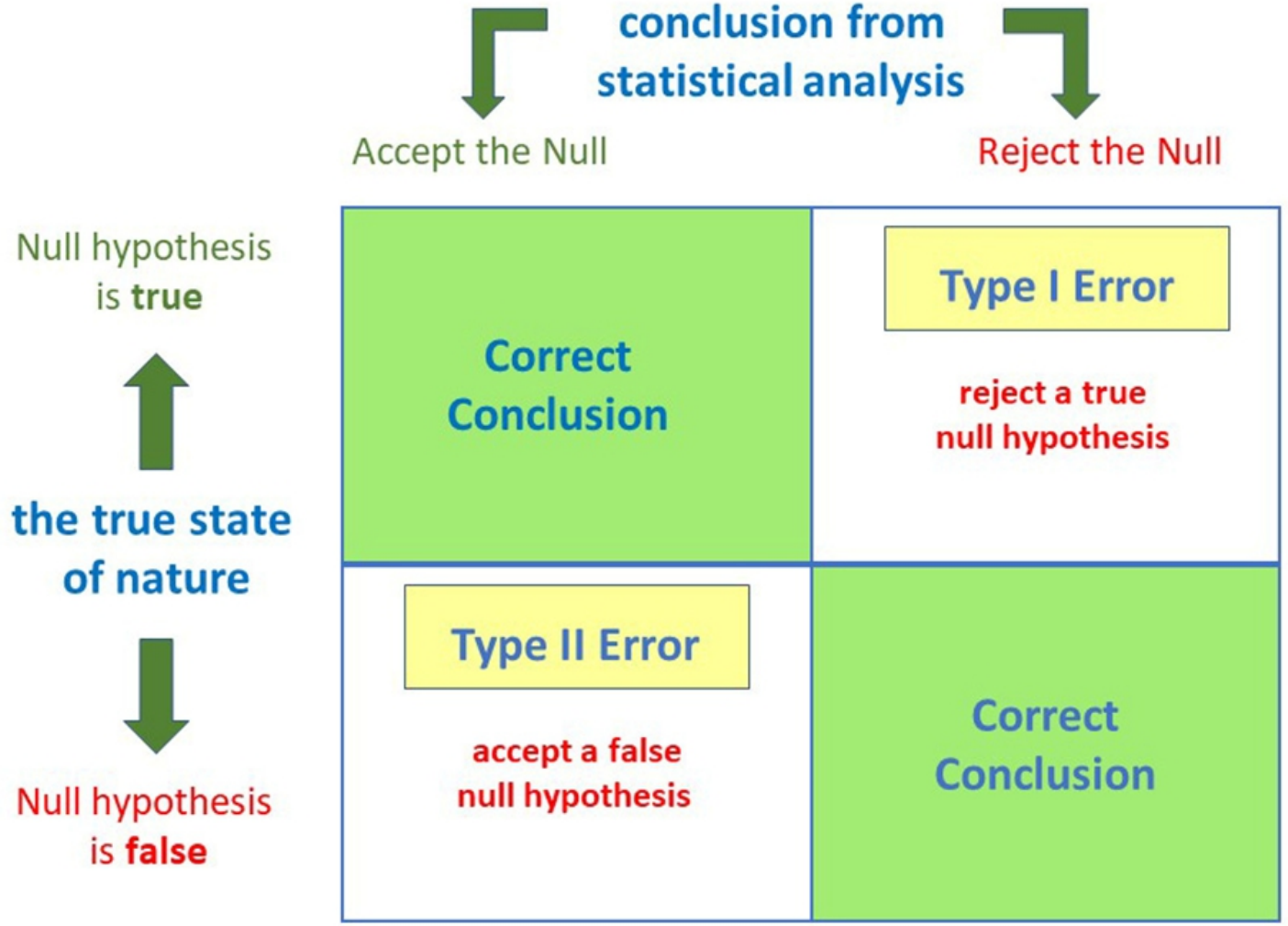
Two-tailed tests:

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

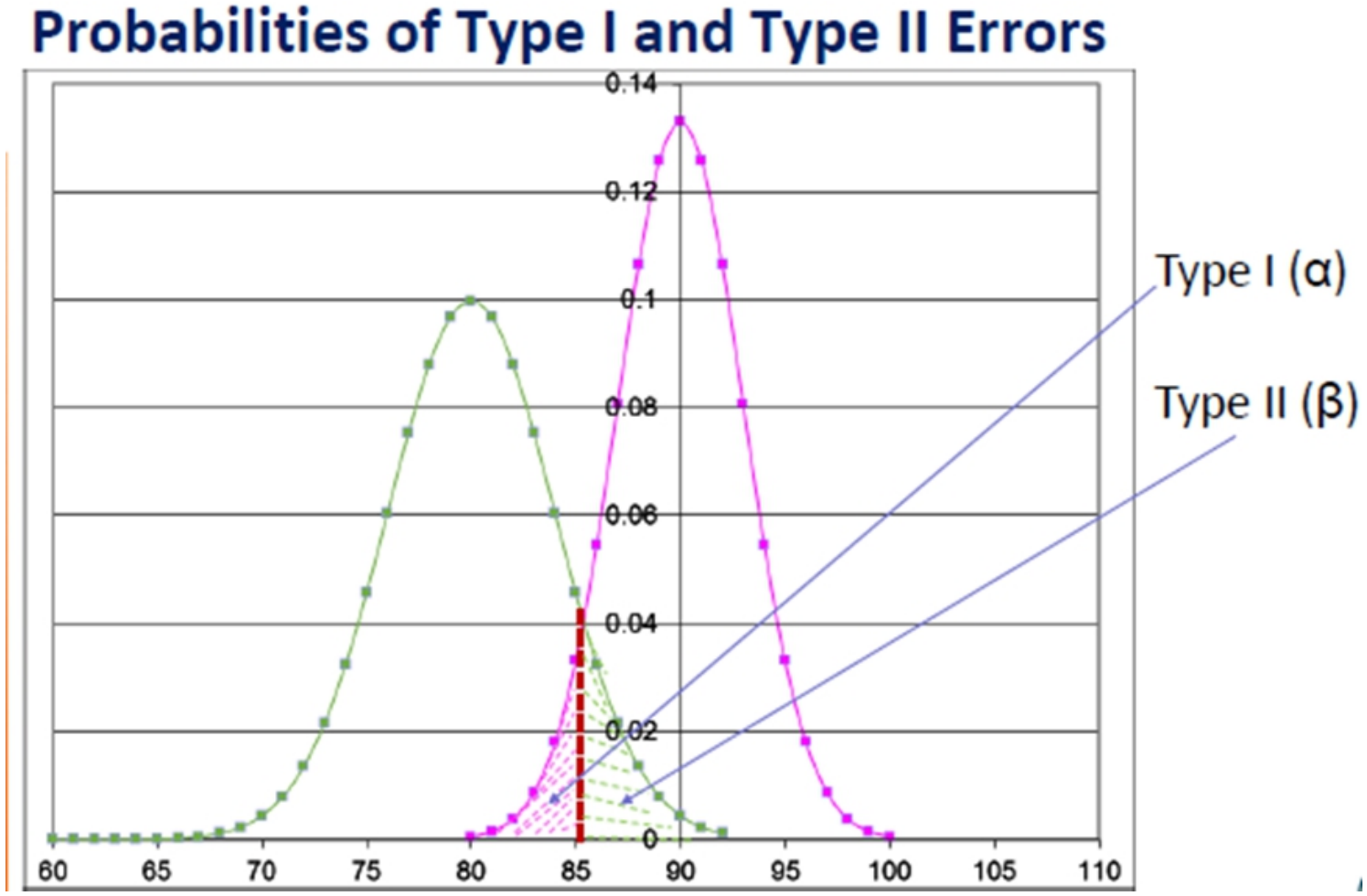
- If H1 includes a "≠" sign (Not equal to), then the two-tailed test is used, since we then look for a change in parameter, rather than an increase or a decrease.



Types of Errors

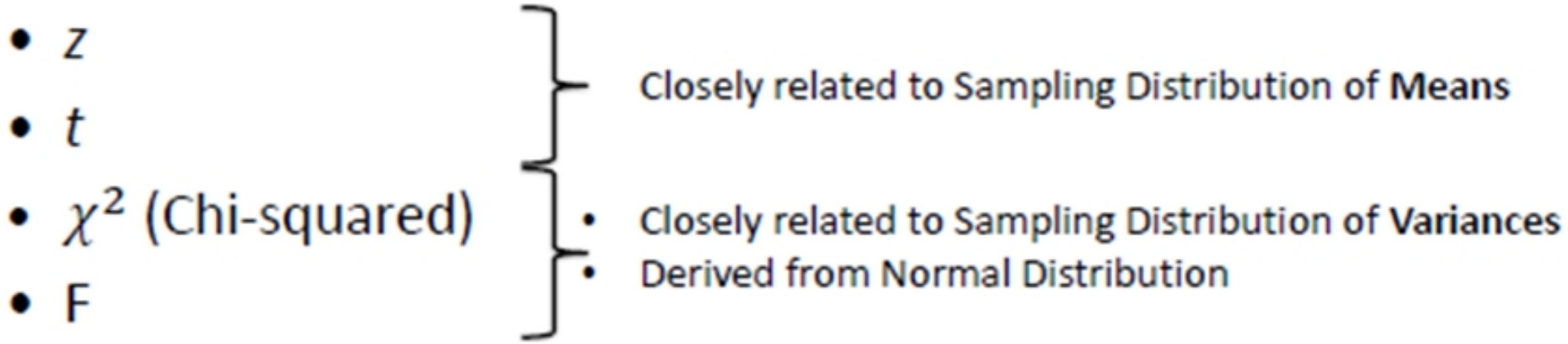


Critical Region up Close



Common Test Statistics for Inferential Techniques

- Inferential techniques (Confidence Intervals and Hypothesis Testing) most commonly use 4 test statistical methods:



Two-Sample t-test for Paired Data:

- To study if their mean values are the same – we can create a new data set from the difference of the individual data points.

$$Xnew = X1 - X2$$

- Subsequently, we can look at how far away from zero is the mean E (Xnew)

$$t = \frac{\overline{X_{new}} - 0}{SE(\overline{X_{new}})}$$

Time to Solve the puzzle (Minutes)			
Volume	After	Before	A-B
1	63	55	8
2	54	62	-8
3	79	108	-29
4	68	77	-9
5	87	83	4
6	84	78	6
7	92	79	13
8	57	94	-37
9	66	69	-3
10	53	66	-13
11	76	72	4
12	63	77	-14
Total	842	920	-78
Mean	70.17	76.67	-6.50

Two-Sample t-test for Paired Data:

Mean of the difference,

$$\bar{d} = -6.5$$

Standard Deviation of the differences,

$$s_d = 15.1$$

Critical Region up Close

Standard Error of the mean,

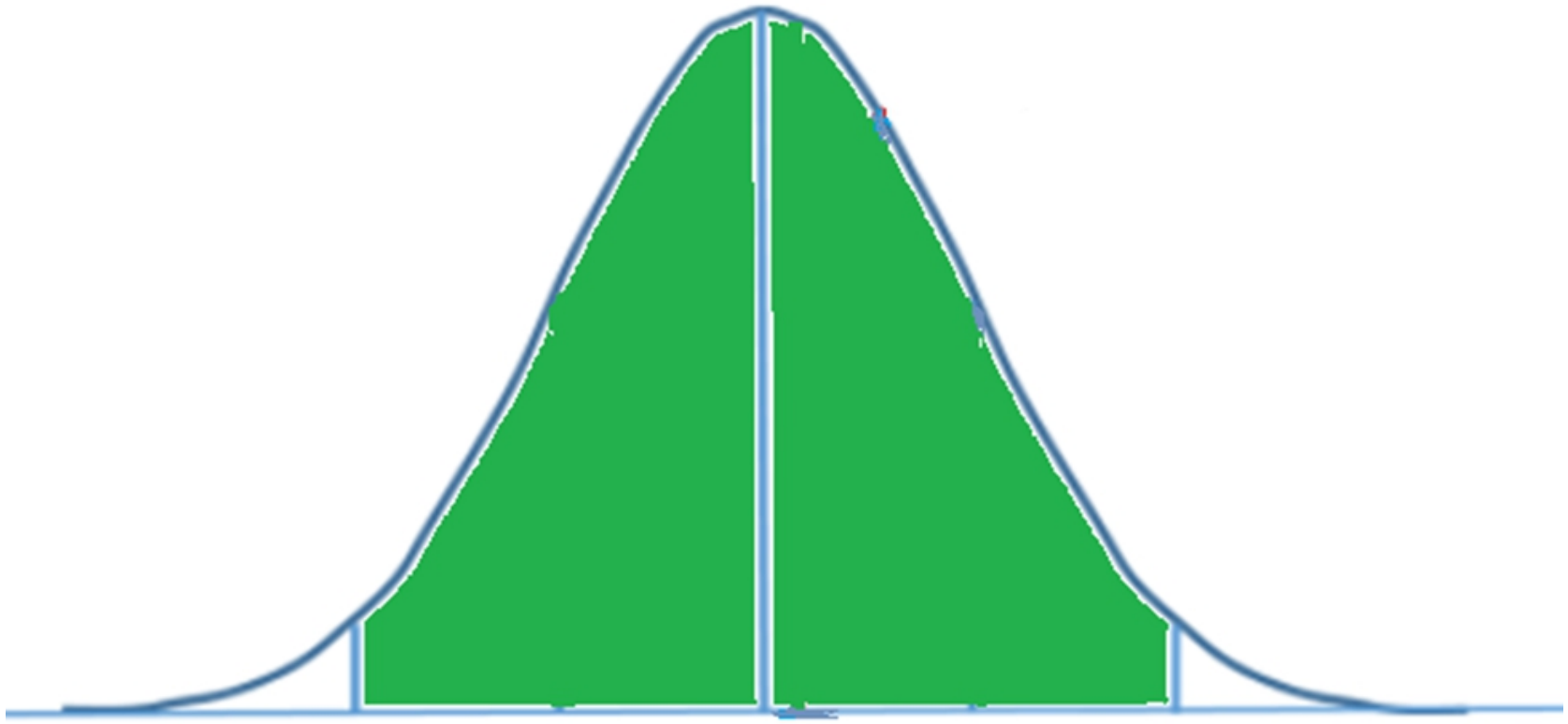
$$SE(\bar{d}) = \frac{s_d}{\sqrt{n}} = 4.37$$

$$t = \frac{\bar{d}}{SE(\bar{d})}$$

t = - 6.5 / 4.37

t = -1.487

Critical Region up Close



Two-Sample t-Test for Unpaired Data:

$$\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2$$

$$\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$$

$$Z = \frac{observed\ difference - expected\ difference}{SE\ of\ the\ difference} = \frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}}$$

Two-Sample t-Test for Unpaired Data:

- This is the test statistic for a 2-sample z-test.

H0: μ1 = μ2
H1: μ1 ≠ μ2

- t-test statistic,

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

Two-Sample t-Test for Unpaired Data:

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

$$t = \frac{(\overline{x_1} - \overline{x_2})}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

With (n_1 and $n_2 - 2$) degrees of freedom

Clearance of theophylline	
Control Group	**Treated Group**
0.81	1.15
1.06	1.28
0.43	1.00
0.54	0.95
0.68	1.06
0.56	1.15
0.45	0.72
0.88	0.79
0.73	0.67
0.43	1.21
0.46	0.92
0.43	0.67
0.37	0.76
0.73	0.82
0.93	0.82

It is a Two-tailed test.

$$s_p{}^2 = \frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{(n_1-1)+(n_2-1)}; \; t = \frac{(\overline{x_1}-\overline{x_2})}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ with } (n_1+n_2-2) \text{ df.}$$

$$S_p{}^2 = \frac{(((15-1) * (0.0408)) + ((15-1) * (0.0467))}{(15 - 1) * (15 + 1)}$$

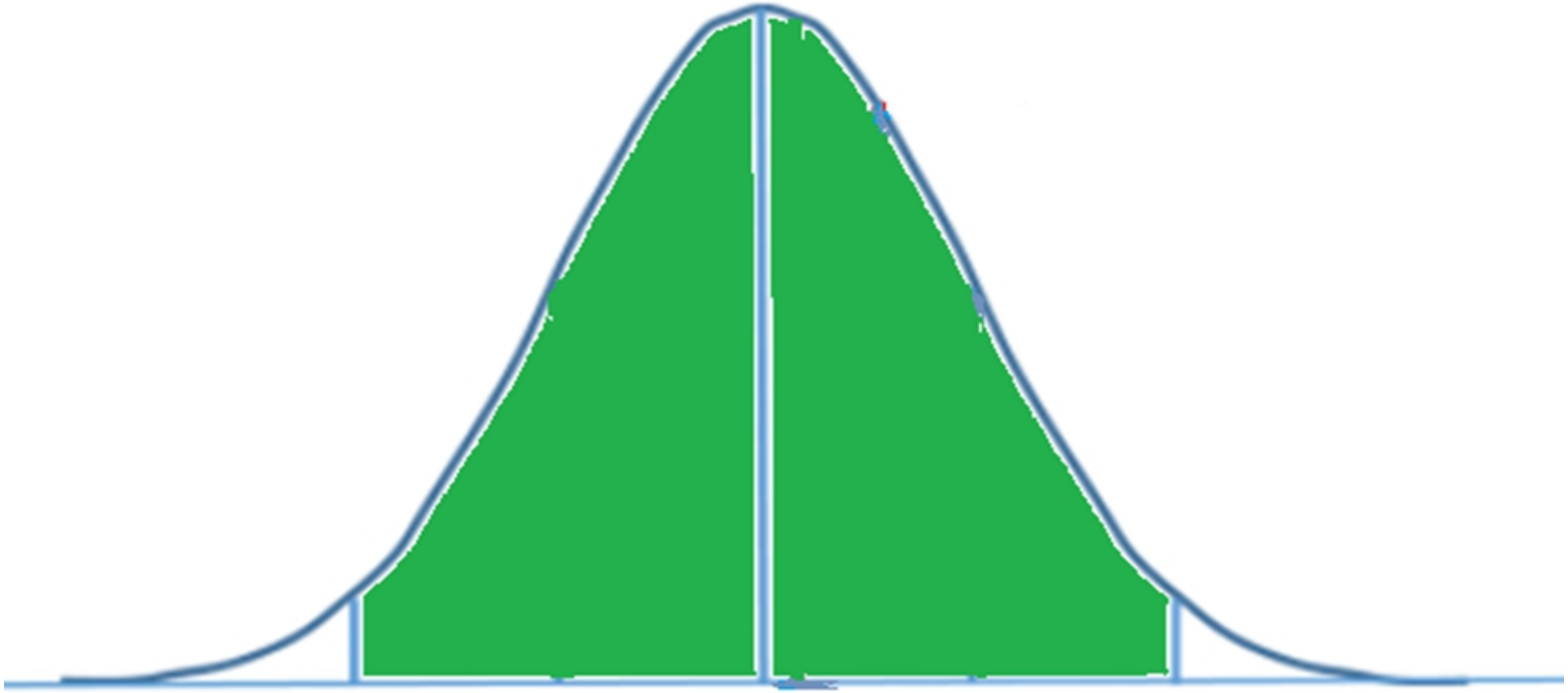
$$S_p{}^2 = 0.04375$$

$$S_p = 0.209$$

$$t = \frac{0.931-0.633}{0.209*\sqrt{\frac{1}{15}+\frac{1}{15}}}$$

t = 3.91

Two-Sample t-Test for Unpaired Data:



Confidence Intervals

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Rewriting

$$(\overline{x_1} - \overline{x_2}) - ts_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq \mu_1 - \mu_2 \leq (\overline{x_1} - \overline{x_2}) + ts_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

((0.298) – (2.048 * (0.0763)) ≤ μ1 - μ2 ≤ ((0.298) + (2.48 * (0.0763))

95% CI: (0.142, 0.454)

Welch's t-test using Welch-Satterthwaite equation to calculate the degrees of freedom

$$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2; \text{Test statistic, } t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$$

When two standard deviations are not equal, we use the above formula.

In this case, the degree of freedom is calculated as:

$$\nu = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\left[\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}\right]}$$

"Complete Lab 4"

"Complete Case Study"

Complete Assessment, Fill in the spaces

Programming Assignment for Lab 3 and Lab 4 for Homework