

Mean	$\bar{x} = \frac{\sum x}{n}$
Mean (Frequency Table)	$\bar{x} = \frac{\sum (f \cdot x)}{\sum f}$
Standard Deviation	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$
Standard Deviation (Shortcut)	$s = \sqrt{\frac{n(\sum x^2) - (\sum x)^2}{n(n - 1)}}$
Standard Deviation (Frequency Table)	$s = \sqrt{\frac{n[\sum (f \cdot x^2)] - [\sum (f \cdot x)]^2}{n(n - 1)}}$

Addition Rule (Mutually Exclusive Events)	$P(A \text{ or } B) = P(A) + P(B)$
Addition Rule (Not mutually exclusive)	$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Multiplication Rule (Independent Events)	$P(A \text{ and } B) = P(A) \cdot P(B)$
Multiplication Rule (Dependent Events)	$P(A \text{ and } B) = P(A) \cdot P(B A)$
Rule of Complements	$P(A) = 1 - P(\bar{A})$
Permutation (No element alike)	${}_nP_r = \frac{n!}{(n - r)!}$
Permutation (Some element alike)	$\frac{n!}{n_1! n_2! \cdots n_k!}$
Combination	${}_nC_r = \frac{n!}{(n - r)! r!}$

Mean of Probability Distribution	$\mu = \sum [x \cdot P(x)]$
Standard Deviation of Probability Distribution	$\sigma = \sqrt{\sum [x^2 \cdot P(x)] - \mu^2}$
Probability of Binomial Distribution	$P(x) = \frac{n!}{(n - x)! x!} \cdot p^x \cdot q^{n-x}$
Mean of Binomial Distribution	$\mu = n \cdot p$
Standard Deviation of Binomial Distribution	$\sigma = \sqrt{n \cdot p \cdot q}$

Z-Score (Standard Score)	$z = \frac{x - \mu}{\sigma} \text{ or } \frac{x - \bar{x}}{s}$
Central Limit Theorem	$\mu_{\bar{x}} = \mu$
Standard Error	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Margin of Error (Population Proportion)	$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$
Confidence Interval (Population Proportion)	$\hat{p} - E < p < \hat{p} + E$
Margin of Error (Population Mean σ Known)	$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
Margin of Error (Population Mean σ Unknown)	$E = t_{\alpha/2} \frac{s}{\sqrt{n}}$
Confidence Interval (Population Mean)	$\bar{x} - E < \mu < \bar{x} + E$
Confidence Interval (Variance)	$\frac{(n-1)s^2}{\chi_R^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_L^2}$

Sample Size Determination (Proportion p q unknown)	$n = \frac{[z_{\alpha/2}]^2 0.25}{E^2}$
Sample Size Determination (Proportion p q known)	$n = \frac{[z_{\alpha/2}]^2 \hat{p}\hat{q}}{E^2}$
Sample Size Determination (Mean)	$n = \left[\frac{z_{\alpha/2} \sigma}{E} \right]^2$

Test Statistic (Proportion)	$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$
Test Statistic (Mean σ known)	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
Test Statistic (Mean σ Unknown)	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
Test Statistic (Standard Deviation)	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$

1.1 Key Words and Definitions

Data	Any observation that has been collected and recorded.
Statistics	Collect, Analyze, Summarize, Interpret, and Draw Conclusions from Data.
Population	The complete set of elements being studied.
Sample	Some subset of a Population
Census	Collection of data from every member of a Population
Parameter	A characteristic of a Population
Statistic	A characteristic of a Sample

1.3 Categories of Data, Levels of Management

Qualitative (Catagorical)	<ul style="list-style-type: none"> Non-numbers: Gender, Race, Religion, Zip Codes Mathematical Operations are Meaningless
Quantitative	<ul style="list-style-type: none"> Numerical: Height, Weight, Wages, MPH, Temperature Math operations are meaningful
Types of Quantitative Data	
Discreet	Countable or Finite: Usually a count, # of books, dice
Continuous	Infinite number of possible values (not countable): Usually a measurement, temperature
Levels of Measurement	
Nominal	Categories. Not ordered: Religion
Ordinal	Can be ordered: Differences are meaningless: Rank in race, Colors in a spectrum
Interval	Ordered, differences are meaningful, No "Natural Zero": Temperature
Ratio	Ordered, differences are meaningful, Have a "Natural Zero": Amount of Money
Univariate vs. Bivariate	
Classification of data based on Number of Variables being studied.	
Univariate	Data that has only one variable. A survey to estimate the average weight of high school students, working with one variable (weight).
Bivariate	Data that has two variables. A survey to see if there were a relationship between the height and weight of high school students, working with two variables (height and weight).

1.5 Sampling Techniques: How to Develop Random Sample

Observation	Measures specific traits, But DOES NOT MODIFY subjects of observation
Experiment	Apply a treatment of subject of experiment and then measure the effect on the subjects.

Random: Each Member of a Population HAS AN EQUAL CHANCE of being selected in the sample.

Simple Random Sample: Each group of size 'n' from the population has an equal chance of being selected in the sample. Properties:

- The population consists of N objects.
- The sample consists of n objects.
- **All possible samples of n objects are equally likely to occur.**

Sampling Techniques

Convenience Sample	Use the subjects from population that are easy to get (Not Random).
Systematic Sample	Put a Population in some order and select every ' k^{th} ' member.
Stratified Sample	Break Population into subgroups based on a characteristic, then sample each subgroup.
Cluster Sample	Divide Population into Clusters (Regardless of Characteristics), Randomly Select a certain # of Clusters, and Then collect data from Entire Cluster.

Sampling With Replacement and Without Replacement

Suppose we use the lottery method described above to select a simple random sample. After we pick a number from the bowl, we can put the number aside or we can put it back into the bowl. If we put the number back in the bowl, it may be selected more than once; if we put it aside, it can be selected only one time.

When a population element can be selected more than one time, we are **sampling with replacement**. When a population element can be selected only one time, we are **sampling without replacement**.

Errors in Sampling

Non-sampling Error	When statistician records information incorrectly, does some mathematical calculation error.
Sampling Error	When you are sampling a population there may be inherent errors, it is the difference between sample and population, because no sample can represent the population absolutely correctly .

2.2 Organizing & Summarizing Data: Frequency Distribution & Histograms

Frequency Distribution: A List of Values with corresponding frequencies.

Class Width: Difference between two "Lower Class Limits".

Lower Class Limit: Smallest Value belonging to a Class. E.g. if all students in a group are between 18 and 21 years old, then lower class limit is 18. (In the example below: 18, 22, 26, ..., 46)

Upper Class Limit: Largest value in a Class. (In the example below: 21, 25, ..., 49)

Class Mid-point: $(\text{Upper Class Limit} + \text{Lower Class Limit}) / 2$ (In the example below: 19.5, 23.5, ..., 47.5)

Class Boundaries: Used to separate classes without gaps (In the example below: 17.5, 21.5, 25.5,..., 45.5, 49.5)

Steps with example (A class with students with max age 44, lowest age 18):

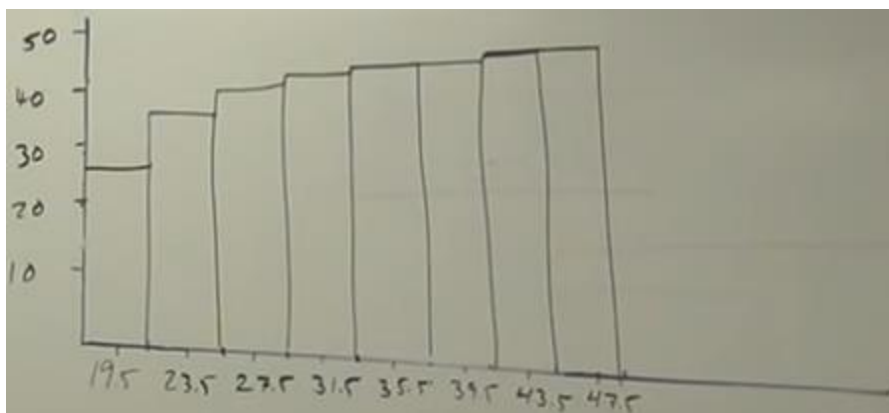
1. Determine # of classes: 8
2. Class Width = $(\text{Max value} - \text{Min value}) / \# \text{ of classes} = (44 - 18) / 8 = 26 / 8 = 3.25 = 4$ (Round Up)
3. Start with smallest value = 18
4. Create classes with class width, lower class limits first

Age	Frequency (f)	Relative Frequency (Percentage) $= (f / \Sigma f) * 100$	Cumulative Frequency (Adds Sequential classes together)
18-21	25	58.1	25
22-25	10	23.3	35
26-29	4	9.3	39
30-33	2	4.7	41
34-37	1	2.3	42
38-41	0	0	42
42-45	1	2.3	43
46-49	0	0	43
	$\Sigma f = 43$		

Normal Distribution: It goes up (rises to a peak) then comes down (falls down) symmetrically, thus the above data is not a normal distribution.

Histograms of Cumulative Frequency Distribution: Touching bar charts.

- Horizontal axis: Class Mid-points or Boundaries
- Vertical axis: Frequency



3.2 Data Characteristics: Center of Data set, Mean, Median, Mode

Five Characteristics:

1. Center
2. Variation
3. Distribution
4. Outliers
5. Changes over Time

Center: The MIDDLE of the data set.

Mean	Arithmetic Average Add all the values and divide by # of values you added. Sample Mean $\bar{X} = \Sigma X / n$ Population Mean $\mu = \Sigma X / n$	Mean = $\Sigma X / \# \text{ of values}$ Σ = Sum X = Data value n = # of Items in Sample N = # of Items in Population \bar{X} = Sample Mean μ = Population Mean
Median	Middle Value of data set. <ul style="list-style-type: none"> • MUST be in Order • If odd number of values in data set then Median is the middle number Unordered - 8, 3, 5, 11, 13, 4, 6 Bring to Order - 3, 4, 5, 6, 8, 11, 13 $M = 6$ • If even number of values in data set then Median is the average of two middle numbers 1, 3, 4, 5, 6, 7 $M = 4.5$ • Can be evaluated for decimal numbers also • Note: for calculating Media for decimal numbers always round to one decimal place more than given numbers. 	
Mode	Most common occurring data value. 5.40, 1.10, 0.42, 0.73, 0.48, 1.10 Mode = 1.10 27, 27, 27, 55. 55. 55. 88. 88. 89 Mode = 27, 55 1, 2, 4, 7, 9, 10, 12 Mode = No Mode, or Empty Set	

Mean of a Frequency Distribution:

Multiply frequency with middle value, add them together, divide by sum of frequencies.

AGE	f	X (MIDPT)	f · x
21-30	28	25.5	714
31-40	30	35.5	1065
41-50	12	45.5	546
51-60	2	55.5	111
61-70	2	65.5	131
71-80	2	75.5	151
Total	n = 76		$\Sigma f \cdot x = 2718$

$$\bar{X} = \frac{\Sigma f \cdot x}{n}$$

$$\bar{X} = \frac{2718}{76} = 35.76$$

Weighted Mean: Mean of a weighted distribution

Multiply weights of each data value with data value, divide by sum of weights.

	w	X	X · w
Hw	15%	70	10.5
T ₁	20%	90	18.0
T ₂	20%	68	13.6
T ₃			
T ₄			
T ₅			
T ₆			
T ₇			
T ₈			
T ₉			
T ₁₀			
T ₁₁			
T ₁₂			
T ₁₃			
T ₁₄			
T ₁₅			
T ₁₆			
T ₁₇			
T ₁₈			
T ₁₉			
T ₂₀			
T ₂₁			
T ₂₂			
T ₂₃			
T ₂₄			
T ₂₅			
T ₂₆			
T ₂₇			
T ₂₈			
T ₂₉			
T ₃₀			
T ₃₁			
T ₃₂			
T ₃₃			
T ₃₄			
T ₃₅			
T ₃₆			
T ₃₇			
T ₃₈			
T ₃₉			
T ₄₀			
T ₄₁			
T ₄₂			
T ₄₃			
T ₄₄			
T ₄₅			
T ₄₆			
T ₄₇			
T ₄₈			
T ₄₉			
T ₅₀			
T ₅₁			
T ₅₂			
T ₅₃			
T ₅₄			
T ₅₅			
T ₅₆			
T ₅₇			
T ₅₈			
T ₅₉			
T ₆₀			
T ₆₁			
T ₆₂			
T ₆₃			
T ₆₄			
T ₆₅			
T ₆₆			
T ₆₇			
T ₆₈			
T ₆₉			
T ₇₀			
T ₇₁			
T ₇₂			
T ₇₃			
T ₇₄			
T ₇₅			
T ₇₆			
T ₇₇			
T ₇₈			
T ₇₉			
T ₈₀			
T ₈₁			
T ₈₂			
T ₈₃			
T ₈₄			
T ₈₅			
T ₈₆			
T ₈₇			
T ₈₈			
T ₈₉			
T ₉₀			
T ₉₁			
T ₉₂			
T ₉₃			
T ₉₄			
T ₉₅			
T ₉₆			
T ₉₇			
T ₉₈			
T ₉₉			
T ₁₀₀			
T ₁₀₁			
T ₁₀₂			
T ₁₀₃			
T ₁₀₄			
T ₁₀₅			
T ₁₀₆			
T ₁₀₇			
T ₁₀₈			
T ₁₀₉			
T ₁₁₀			
T ₁₁₁			
T ₁₁₂			
T ₁₁₃			
T ₁₁₄			
T ₁₁₅			
T ₁₁₆			
T ₁₁₇			
T ₁₁₈			
T ₁₁₉			
T ₁₂₀			
T ₁₂₁			
T ₁₂₂			
T ₁₂₃			
T ₁₂₄			
T ₁₂₅			
T ₁₂₆			
T ₁₂₇			
T ₁₂₈			
T ₁₂₉			
T ₁₃₀			
T ₁₃₁			
T ₁₃₂			
T ₁₃₃			
T ₁₃₄			
T ₁₃₅			
T ₁₃₆			
T ₁₃₇			
T ₁₃₈			
T ₁₃₉			
T ₁₄₀			
T ₁₄₁			
T ₁₄₂			
T ₁₄₃			
T ₁₄₄			
T ₁₄₅			
T ₁₄₆			
T ₁₄₇			
T ₁₄₈			
T ₁₄₉			
T ₁₅₀			
T ₁₅₁			
T ₁₅₂			
T ₁₅₃			
T ₁₅₄			
T ₁₅₅			
T ₁₅₆			
T ₁₅₇			
T ₁₅₈			
T ₁₅₉			
T ₁₆₀			
T ₁₆₁			
T ₁₆₂			
T ₁₆₃			
T ₁₆₄			
T ₁₆₅			
T ₁₆₆			
T ₁₆₇			
T ₁₆₈			
T ₁₆₉			
T ₁₇₀			
T ₁₇₁			
T ₁₇₂			
T ₁₇₃			
T ₁₇₄			
T ₁₇₅			
T ₁₇₆			
T ₁₇₇			
T ₁₇₈			
T ₁₇₉			
T ₁₈₀			
T ₁₈₁			
T ₁₈₂			
T ₁₈₃			
T ₁₈₄			
T ₁₈₅			
T ₁₈₆			
T ₁₈₇			
T ₁₈₈			
T ₁₈₉			
T ₁₉₀			
T ₁₉₁			
T ₁₉₂			
T ₁₉₃			
T ₁₉₄			
T ₁₉₅			
T ₁₉₆			
T ₁₉₇			
T ₁₉₈			
T ₁₉₉			
T ₂₀₀			
T ₂₀₁			
T ₂₀₂			
T ₂₀₃			
T ₂₀₄			
T ₂₀₅			
T ₂₀₆			
T ₂₀₇			
T ₂₀₈			
T ₂₀₉			
T ₂₁₀			
T ₂₁₁			
T ₂₁₂			
T ₂₁₃			
T ₂₁₄			
T ₂₁₅			
T ₂₁₆			
T ₂₁₇			
T ₂₁₈			
T ₂₁₉			
T ₂₂₀			
T ₂₂₁			
T ₂₂₂			
T ₂₂₃			
T ₂₂₄			
T ₂₂₅			
T ₂₂₆			
T ₂₂₇			
T ₂₂₈			
T ₂₂₉			
T ₂₃₀			
T ₂₃₁			
T ₂₃₂			
T ₂₃₃			
T ₂₃₄			
T ₂₃₅			
T ₂₃₆			
T ₂₃₇			
T ₂₃₈			
T ₂₃₉			
T ₂₄₀			
T ₂₄₁			
T ₂₄₂			
T ₂₄₃			
T ₂₄₄			
T ₂₄₅			
T ₂₄₆			
T ₂₄₇			
T ₂₄₈			
T ₂₄₉			
T ₂₅₀			
T ₂₅₁			
T ₂₅₂			
T ₂₅₃			
T ₂₅₄			
T ₂₅₅			
T ₂₅₆			
T ₂₅₇			
T ₂₅₈			
T ₂₅₉			
T ₂₆₀			
T ₂₆₁			
T ₂₆₂			
T ₂₆₃			
T ₂₆₄			
T ₂₆₅			
T ₂₆₆			
T ₂₆₇			
T ₂₆₈			
T ₂₆₉			
T ₂₇₀			
T ₂₇₁			
T ₂₇₂			
T ₂₇₃			
T ₂₇₄			
T ₂₇₅			
T ₂₇₆			
T ₂₇₇			
T ₂₇₈			
T ₂₇₉			
T ₂₈₀			
T ₂₈₁			
T ₂₈₂			
T ₂₈₃			
T ₂₈₄			
T ₂₈₅			
T ₂₈₆			
T ₂₈₇			
T ₂₈₈			
T ₂₈₉			
T ₂₉₀			
T ₂₉₁			
T ₂₉₂			
T ₂₉₃			
T ₂₉₄			
T ₂₉₅			
T ₂₉₆			
T ₂₉₇			
T ₂₉₈			
T ₂₉₉			
T ₃₀₀			
T ₃₀₁			
T ₃₀₂			
T ₃₀₃			
T ₃₀₄			
T ₃₀₅			
T ₃₀₆			
T ₃₀₇			
T ₃₀₈			
T ₃₀₉			
T ₃₁₀			
T ₃₁₁			
T ₃₁₂			
T ₃₁₃			
T ₃₁₄			
T ₃₁₅			
T ₃₁₆			
T ₃₁₇			
T ₃₁₈			
T ₃₁₉			
T ₃₂₀			
T ₃₂₁			
T ₃₂₂			
T ₃₂₃			
T ₃₂₄			
T ₃₂₅			
T ₃₂₆			
T ₃₂₇			
T ₃₂₈			
T ₃₂₉			
T ₃₃₀			
T ₃₃₁			
T ₃₃₂			
T ₃₃₃			
T ₃₃₄			
T ₃₃₅			
T ₃₃₆			
T ₃₃₇			
T ₃₃₈			
T ₃₃₉			
T ₃₄₀			
T ₃₄₁			
T ₃₄₂			
T ₃₄₃			
T ₃₄₄			
T ₃₄₅			
T ₃₄₆			
T ₃₄₇			
T ₃₄₈			
T ₃₄₉			
T ₃₅₀			
T ₃₅₁			
T ₃₅₂			
T ₃₅₃			
T ₃₅₄			
T ₃₅₅			
T ₃₅₆			
T ₃₅₇			
T ₃₅₈			
T ₃₅₉			
T ₃₆₀			
T ₃₆₁			
T ₃₆₂			
T ₃₆₃			
T ₃₆₄			
T ₃₆₅			
T ₃₆₆			
T ₃₆₇			
T ₃₆₈			
T ₃₆₉			
T ₃₇₀			
T ₃₇₁			
T ₃₇₂			
T ₃₇₃			
T ₃₇₄			
T ₃₇₅			
T ₃₇₆			
T ₃₇₇			
T ₃₇₈			
T ₃₇₉			
T ₃₈₀			
T ₃₈₁			
T ₃₈₂			
T ₃₈₃			
T ₃₈₄			
T ₃₈₅			
T ₃₈₆			
T ₃₈₇			
T ₃₈₈			
T ₃₈₉			
T ₃₉₀			
T ₃₉₁			
T ₃₉₂			
T ₃₉₃			
T ₃₉₄	</		

3.3 Data Variation: Standard Variation of Data Set

How the data is spread. This is the second characteristic of data set.

Ways to Measure Variation

Range	<ul style="list-style-type: none"> • Max Value – Min Value • Easy to find • Does not consider all values
Standard Deviation	<p>Measures the average distance of data from the mean.</p> <ul style="list-style-type: none"> • Never negative • Can be zero if all elements alike • Greatly affected by Outliers • Closely grouped data will have a small Std. Deviation • Spread out data will have a large Std. Deviation • Sample Standard Deviation denoted by: s <p>Standard Deviation of a SAMPLE:</p> <div style="display: flex; align-items: center; justify-content: center;"> $S = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ or $S = \sqrt{\frac{n \sum (x^2) - (\sum x)^2}{n(n-1)}}$ </div> <p>Standard Deviation of a POPULATION:</p> $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$
Variance	<p>Measures how the data distributes itself about the mean, or how spread out data is from the mean.</p> <p>SAMPLE VARIANCE: S^2</p> <p>POP. VARIANCE: σ^2</p>

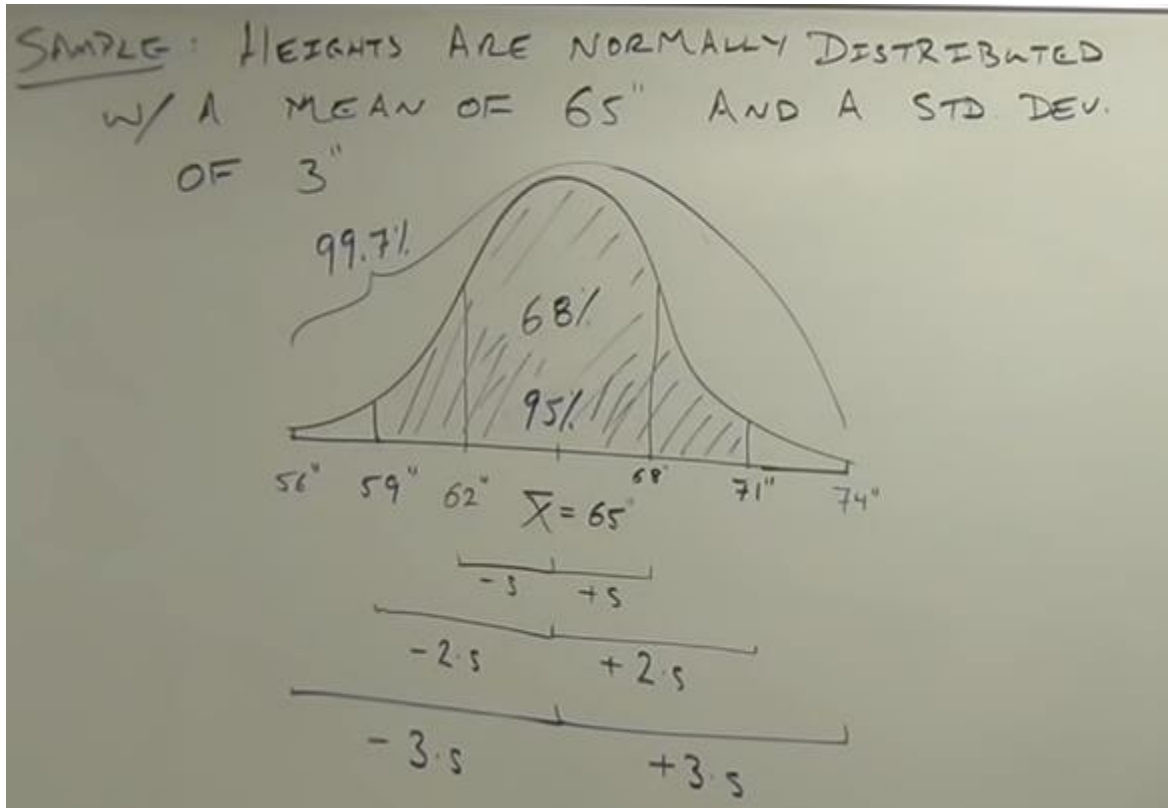
Empirical Rule

If a data set is normally distributed, we can use following rules:

- 68% of data will fall within 1 Std. Deviation of Mean
- 95% of data will fall within 2 Std. Deviation of Mean
- 99.7% of data will fall within 3 Std. Deviation of Mean

Usual Data / Unusual Data / Very Rare

- If a data value lies within 2 Std. Deviation of Mean, then it is considered 'USUAL'
- If a data value lies outside 3 Std. Deviation of Mean, then it is considered 'Very Rare'.



Co-efficient of Variation

Standard Deviations by themselves cannot be compared. E.g. in the following which one has a numerically bigger standard deviation:

	\bar{X} Mean	s Standard Deviation
Height	65"	3"
Weight	175 lbs	4 lbs

Numerically, weight has bigger standard deviation.

But which has more variation? Simply because weight has higher standard deviation does not mean that it has more variation.

For this we have two choices, either use **co-efficient of variation**, which **translates a standard deviation in comparison to its mean to a percentage** and thus says that this data varies more percentage as compared to other one. The other way is called Z-score.

$$\text{COEFFICIENT OF VARIATION} = \text{C.V.} = \left(\frac{s \text{ Standard Deviation}}{\bar{X} \text{ Mean}} \right) * 100$$

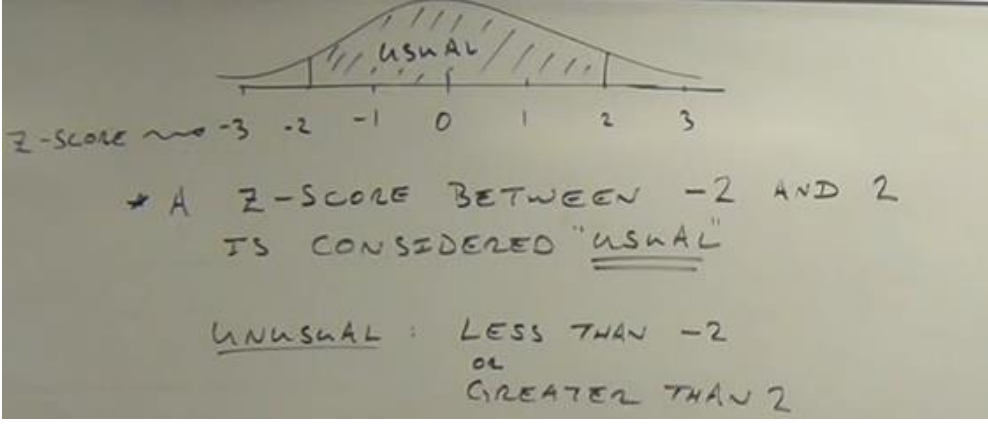
$$\text{Height C.V.} = (3 / 65) * 100 = 4.6\%$$

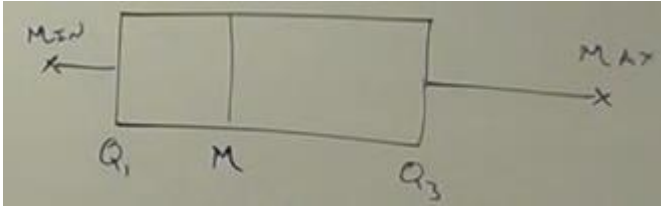
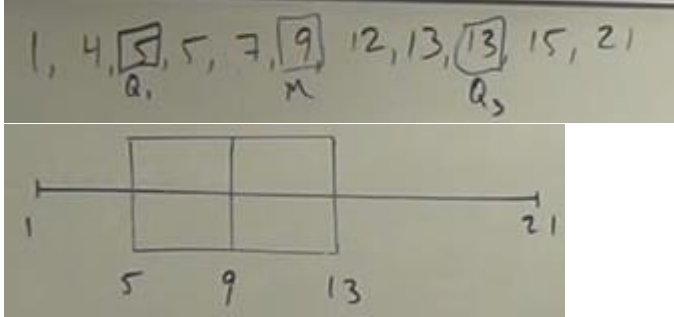
$$\text{Weight C.V.} = (4/175) * 100 = 2.3\%$$

In the example above, height is varying more than weight even though its Std. Deviation is lower.

3.4 Data Comparison: Z-Score, Percentile, Quartile

Measures of relative standing: Comparing between or within Data Sets.

Z-Score	<p>The # of Std. Dev. a Data Value (X) is away from the mean.</p> <ul style="list-style-type: none"> Allows Comparison of the variation in two different SAMPLES / POPULATIONS Z-score between -2 and 2 is considered "usual". Z-score less than -2 and more than 2 is considered "unusual". The larger the Z-score, in terms of absolute value, the rarer the data value. <div style="display: flex; justify-content: space-around;"> <div style="text-align: center;"> <p>SAMPLE</p> $Z = \frac{X - \bar{X}}{s}$ </div> <div style="text-align: center;"> <p>POPULATION</p> $Z = \frac{X - \mu}{\sigma}$ </div> </div> 
Quartiles	<p>Break data into quarters like the medians do in the middle.</p> <ul style="list-style-type: none"> 1st Quartile – Q_1 - Bottom 25% of Sorted Data 2nd Quartile – Q_2 (Median M)- Bottom 50% of Sorted Data 3rd Quartile – Q_3 - Bottom 75% of Sorted Data <p>Ex: Data – 1 3 6 10 15 21 28 36 Quartiles: 1 3 [Q1: 4.5] 6 10 [Q2:Median: 12.5] 15 21 [Q3: 24.5] 28 36</p> <p>Ex: Data – 1 3 6 10 15 21 28 36 39 Quartiles: 1 3 [Q1: 4.5] 6 10 [Q2:Median: 15] 21 28 [Q3: 32] 36 39</p>
Percentiles	<p>Separate data into 100 parts.</p> <ul style="list-style-type: none"> Therefore, there are 99 percentiles. <p>Percentile of X = (# of Values less than X / Total # of Values) * 100</p> <p>Ex: A student scored 87/100 on a test. 39 people scored lower than the student. There are 54 students in the class. Percentile of student who scored 87 = $(39/54) * 100 = 72^{\text{nd}}$ Percentile</p> <p>$P_{25} = Q_1$ $P_{50} = Q_2 = M$ $P_{75} = Q_3$</p>
Interquartile Range (IQR)	<p>Data between two quartiles.</p> <p>IQR: Represents Middle 50% of data = $Q_3 - Q_1$</p>

Box Plot	<p>Graphic representation of 5 number summary.</p> <ol style="list-style-type: none"> 1. Minimum value 2. Q_1 3. Median Q_2 4. Q_3 5. Maximum value  <p>Ex:</p> 
-----------------	---

4.2 Probability Introduction

Likelihood of an event.

Probability	<p>The probability of an event is a measure of the likelihood that the event will occur.</p> <ul style="list-style-type: none"> • The probability of event A is denoted by $P(A)$. • The probability of any event can range from 0 to 1. • Thus, if event A were very unlikely to occur, then $P(A)$ would be close to 0. And if event A were very likely to occur, then $P(A)$ would be close to 1.
Event	A collection of outcomes of a procedure.
Simple Event	A single outcome.
Sample Space	A set of all simple events that represents all possible outcomes of a statistical experiment.
Independent Event	Occurrence of one event does not affect another.
Complementary Events	<p>Mutually Exclusive Events. Denoted A and \bar{A}</p> <ul style="list-style-type: none"> • Sum of Complimentary Events = $P(A) + P(\bar{A}) = 1$ • Rolling a 5 in a Die – $P(5) = 1/6$ and $P(\bar{5}) = 5/6$

Ex:

- **Procedure:** Coin flipped 3 times.
- **Event:** 1 head, 2 tails.
- **Sample Space:** {HHH, HHT, HTH, HTT, TTT, TTH, THT, THH}

Types of Probability

Observed Probability	What did happen?	$P(A) = (\# \text{ of Times 'A' occurred}) / (\# \text{ Times Procedure Repeated})$ Ex: Flipped a coin 100 times, got 64 tails. $P(T) = 64 / 100 = .64$
Classical Probability	What could have?	$P(A) = (\# \text{ of Ways 'A' could happen}) / (\# \text{ Simple Events i.e. Outcomes})$ Ex: Probability of selecting a Heart from deck of cards. $P(\text{Heart}) = 13 / 52 = 0.25$
Subjective Probability	Educated Guess	Doctor tells patient: "There is 80% chance that you have flue, but you better get tested."

Note: The more a procedure is repeated, the closer the **Observed Probability** will be to the **Classical Probability**.

4.3 Addition Rule for Probability

The *addition rule* is a tool for finding $P(A \text{ or } B)$, which is the probability that either event A occurs or event B occurs (or they both occur) as the single outcome of a procedure.

Addition Rule: To find $P(A \text{ or } B)$, find the sum of the number of ways event A can occur and the number of ways event B can occur, adding in such a way that every outcome is counted only once. $P(A \text{ or } B)$ is equal to that sum, divided by the total number of outcomes in the sample space.

Compound Event:	An event consisting of two or more events. Example: Throw a dice and get combination of outputs.	
Disjoint Events	The addition rule is simplified when the events are disjoint. Events A and B are disjoint (or mutually exclusive) if they cannot occur at the same time. (That is, disjoint events do not overlap.)	
	<p>Not Disjoint Events</p>	<p>Ex: Not Disjoint (non-mutually exclusive) Randomly selecting someone taking a statistics course. Randomly selecting someone who is a female. (The selected person <i>can</i> be both.)</p>
	<p>Disjoint Events</p>	<p>Ex: Disjoint (Mutually Exclusive) Randomly selecting someone who is a registered Democrat. Randomly selecting someone who is a registered Republican. (The selected person <i>cannot</i> be both.)</p>

Addition Rule 1	When two events, A and B, are not disjoint (non-mutually exclusive) , the probability that A or B or both will occur is: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Addition Rule 2	When two events, A and B, are disjoint (mutually exclusive) , the probability that A or B will occur is the sum of the probability of each event. Whenever A and B are disjoint, $P(A \text{ and } B)$ becomes zero in the addition rule. $P(A \text{ or } B) = P(A) + P(B)$

If events A and B **come from the same sample space**, the **probability that event A and/or event B occur** is equal to the probability that event A occurs plus the probability that event B occurs minus the probability that both events A and B occur.

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

We have two events from the same sample space, and we want to know the probability that either event occurs, it requires to eliminate any double count.

In other words: In a single trial, probability of either A occurring or B occurring or both occurring:

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

Ex:

	Did Not Do It	Did It
Guilty	11 False Positive	72 True Positive
Not Guilty	85 True Negative	9 False Negative

$$\text{Guilty OR Did It} = 11 + 72 + 9 = 92$$

$$P(\text{Guilty OR Did It}) = 92 / 177 = .52 = 52\%$$

Ex: Addition Rule 1 (Not disjoint, non-mutually exclusive): In a math class of **30 students, 17 boys** and **13 girls**. On a unit test, **4 boys and 5 girls** made an **A grade**. If a student is chosen at random from the class, what is the probability of choosing a girl or an A student?

Probabilities: $P(\text{girl or A}) = P(\text{girl}) + P(A) - P(\text{girl and A})$

$$= \frac{13}{30} + \frac{9}{30} - \frac{5}{30}$$

$$= \frac{17}{30}$$

Ex: Addition Rule 2 (Disjoint, mutually exclusive Events): A single 6-sided die is rolled. What is the probability of rolling a 2 or a 5?

$$P(2) = \frac{1}{6}$$

$$P(5) = \frac{1}{6}$$

$$P(2 \text{ or } 5) = P(2) + P(5)$$

$$= \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

4.4 Multiplication Rule for Probability

The rule for finding $P(A \text{ and } B)$ is called the **multiplication rule** because it involves the multiplication of the probability of event A and the probability of event B (where, if necessary, the probability of event B is adjusted because of the outcome of event A).

Independent Event	Two events A and B are independent if the occurrence of one does not affect the <i>probability</i> of the occurrence of the other.
Dependent Event	If A and B are not independent, they are said to be dependent .
Conditional Probability	Probability of an event occurring subject to some earlier event having occurred . $P(B A)$ = Probability of event B occurring given that event A has already occurred, can be found by dividing the probability of events A and B both occurring by the probability of event A $= P(A \text{ and } B) / P(A)$
Multiplication Rule	Probability of event ' A ' occurring, and then, event ' B ' occurring in successive trials: $P(A \text{ and } B) = P(A) \cdot P(B A)$ $P(A \text{ and } B)$: P (event A occurs in a first trial and event B occurs in a second trial) $P(B A)$: represents the probability of event B occurring after it is assumed that event A has already occurred. (Interpret $B A$ as "event B occurring after event A has already occurred.")

Ex: Independent Event Multiplication Rule: Question 1 has 2 possible answers, question 2 is related to question 1 and has 5 possible answers, what is the probability of correct answer? $1/2 * 1/5 = 1/10$

Ex: Conditional Probability: Finding $P(B | A)$ If 1 of the 1000 test subjects is randomly selected, **find the probability** that the **subject is found guilty, given that the subject did it**. That is, find P (subject did it).

	Did It	Did Not Do It
Guilty	44 True Positive	6 False Positive
Not Guilty	90 False Positive	860 True Negative

Using the Formula for Conditional Probability:

$$P(B|A) = P(\text{Subject Did It} | \text{Subject Guilty})$$

- $P(\text{subject guilty and did it}) = 44/1000$
- $P(\text{subject guilty}) = 50/1000$
- $P(B|A) = P(\text{subject guilty and did it}) / P(\text{subject guilty}) = (44/1000) / (50/1000) = 0.88$

4.5 Probability of Complimentary Events with “At Least One”

Complementary Events	The complement of event A , denoted by \bar{A} , consists of all outcomes in which event A does <i>not</i> occur. $P(A) = 1 - P(\bar{A})$
Probability of “at least one”	Find the probability that among several trials, we get <i>at least one</i> of some specified event.
Conditional probability	Find the probability of an event occurring when we have additional information that some other event has already occurred.

Ex: Flip Coin 3 times, find probability of at least one head.

Soln.: {HHH, HHT, HTH, HTT, TTT, TTH, THT, THH}

$P(\text{At least one head}) = 1 - P(\text{No head}) = 1 - P(\text{All tails}) = 1 - P(T \text{ and } T \text{ and } T) = 1 - 1/2 * 1/2 * 1/2 = 1 - 1/8$

$P(\text{At least one head}) = 7/8$

Ex: Manufacturer supplies DVDs in lots of 50, reported defect rate is 0.5%, so the probability of an individual disk being defective is 0.005. Thus that the probability of a disk being good is 0.995. **What is the probability of getting at least one defective disk in a lot of 50?**

Soln.:

Step 1: Let A = at least 1 of the 50 disks is defective.

Step 2: Complement of A = not getting at least 1 defective disk among 50 = all 50 disks are good

Step 3: Probability of the complement = **$P(\bar{A})$** = $P(\text{all 50 disks are good})$
 $= 0.995 * 0.995 * \dots * 0.995$
 $= 0.995^{50} = 0.778$

Step 4: **$P(A) = 1 - P(\bar{A}) = 1 - 0.778 = 0.222$**

Interpretation: In a lot of 50 DVDs, there is a 0.222 probability of getting at least 1 defective DVD.

4.7 Counting, Permutation, Combination

Counting in statistics relates to number of ways the things can happen, the number of possible outcomes in a variety of different situations.

Permutations	Arrangements in which different sequences of the same items are counted separately i.e. order matters. For example, with the letters {a, b, c}, the arrangements of abc, acb, bac, bca, cab, and cba are all counted separately as six different permutations.
Combinations	Arrangements in which different sequences of the same items are not counted separately . For example, with the letters {a, b, c}, the arrangements of abc, acb, bac, bca, cab, and cba are all considered to be same combination.
Fundamental Counting Rule	Given that the first event can occur m ways and the second event can occur n ways, then

	<p>Number of ways that two events can occur = $m \times n$.</p> <p>(This rule extends easily to situations with more than two events.)</p> <p>Ex: For a two-character code consisting of a letter followed by a digit, the number of different possible codes is $26 \times 10 = 260$.</p>
Permutations Rule 1	<p>When All of the Items Are Different: Number of different <i>permutations</i> (order counts) when n different items are available, but only r of them are selected <i>without replacement</i>. (Rearrangements of the same items are counted as being different.)</p> ${}_nP_r = \frac{n!}{(n-r)!}$ <p>Ex: If the five letters {a, b, c, d, e} are available and three of them are to be selected without replacement, the number of different permutations is as follows:</p> ${}_nP_r = \frac{n!}{(n-r)!} = \frac{5!}{(5-3)!} = 60$
Permutations Rule 2	<p>When Some of the Items Are Identical: Number of different permutations (order counts) when n items are available and all n are selected without replacement, but some of the items are identical to others: n_1 are alike, n_2 are alike, . . . , and n_k are alike.</p> $\frac{n!}{n_1!n_2! \cdots n_k!}$ <p>Ex: If the 10 letters {a, a, a, a, b, b, c, c, d, e} are available and all 10 of them are to be selected without replacement, the number of different permutations is as follows:</p> $\frac{n!}{n_1!n_2! \cdots n_k!} = \frac{10!}{4!2!2!} = \frac{3,628,800}{24 \cdot 2 \cdot 2} = 37,800$
Combination Rule	<p>Number of different combinations (order does not count) when n different items are available, but only r of them are selected without replacement. (Note that rearrangements of the same items are counted as being the same.)</p> ${}_nC_r = \frac{n!}{(n-r)!r!}$ <p>Ex: If the five letters {a, b, c, d, e} are available and three of them are to be selected without replacement, the number of different combinations is as follows:</p> ${}_nC_r = \frac{n!}{(n-r)!r!} = \frac{5!}{(5-3)!3!} = \frac{120}{2 \cdot 6} = 10$

5.2 Probability Distribution Mean Standard Deviation

Probability Distribution	A TABLE, Formula, or Graph that gives the probability for each value of a Random Variable .
Random Variable	<p>A Variable X that has a value for each outcome of a procedure that is Determined by Chance.</p> <ul style="list-style-type: none"> • Discrete Random Variable: A variable that can hold collection of values that is finite or countable. (If there are infinitely many values, the number of values is countable if it is possible to count them individually, such as the number of tosses of a coin before getting tails.) • Continuous Random Variable: It has infinitely many values, and the collection of values is not countable. (That is, it is impossible to count the individual items because at least some of them are on a continuous scale.)
Expected Value	<p>The expected value of a discrete random variable X is denoted by E, and it is the mean value of the outcomes, so $E = \mu$ and E can also be found by evaluating $\sum [x \cdot P(x)]$.</p> <p>We can think of that mean as the expected value in the sense that it is the average value that we would expect to get if the trials could continue indefinitely.</p> <p>Caution An expected value need not be a whole number, even if the different possible values of x might all be whole numbers. We say that the expected number of girls in five births is 2.5, even though five specific births can never result in 2.5 girls. If we were to survey many couples with five children, we expect that the mean number of girls will be 2.5.</p>

Probability Distribution Table

Remember that **with a probability distribution**, we have a description of a **population instead of a sample**, so the values of the **mean, standard deviation, and variance are parameters** instead of statistics. The mean is the central or “average” value of the random variable. The variance and standard deviation measure the variation of the random variable. These parameters can be found with the following formulas:

Weighted Die		Calculation of MEAN and STD. DEVIATION			
X	P(X)	X.P(X)	X ²	X ² .P(X)	MEAN
1	.05	.05	1	.05	$\mu = \sum [x \cdot P(x)] = 3.4$
2	.15	.30	4	.60	
3	.35	1.05	9	3.15	
4	.30	1.20	16	4.8	STD. DEV $\sigma = \sqrt{\sum [x^2 \cdot P(x)] - \mu^2}$ $= \sqrt{(12.9 - (3.4)^2)} = \sqrt{1.34} = 1.16$
5	.10	0.50	25	2.5	
6	.05	0.30	36	1.8	

Histogram of Probability Distribution

0.35						
0.30						
0.25						
0.20						
0.15						
0.10						
0.05						

- Horizontal Axis: Values of Random Variable
- Vertical Axis: Probability

5.3 Binomial Probability Distribution

Binomial Distribution: Probability Distribution that has only TWO OUTCOMES: Success & Failure

Rules:

1. Procedure MUST be **FIXED # of Trials**
2. Trials must be **Independent** (The outcome of one trial does not affect the others)
3. Each Trial has **only Two Outcomes**: Success or Failure
4. The **probability of success remains the same in all Trials**.

Binomial Probability Formula:

- n - # of Trials
- p = P(Success) – Probability of a Successful Outcome in a Single Trial
- q = P(Failure) - Probability of a Failure
- X – Number of Successes that occur in the n Trials
- $P(X)$ – Probability of getting ' X ' Successes

$$P(x) = \frac{n!}{(n-x)!x!} \cdot p^x \cdot q^{n-x} \quad \text{for } x = 0, 1, 2, \dots, n$$

Ex: Weighted Die: The probability of Rolling a '4' is 30%. The die is rolled 10 times. Find the probability of rolling EXACTLY 8 '4's.

$$n = 10 \quad p = 0.30 \quad q = 0.70 \quad X = 8 \quad P(8) = ?$$

Soln.: $P(8) = nCx \cdot p^x \cdot q^{(n-x)} = {}_{10}C_8 \cdot (.30)^8 \cdot (.70)^{(10-8)} = 45 \cdot (.00006561) \cdot (.49)$
 $= .0014467$

Ex: Given that there is a 0.85 probability that a randomly selected adult knows what Twitter is, when five adults are randomly selected find the probability of getting exactly 3 adults who know what Twitter is.

Soln: Find $P(3)$ given that $n = 5$, $x = 3$, $p = 0.85$, and $q = 0.15$.

$$P(3) = (5! / (5 - 3)! 3!) \cdot 0.85^3 \cdot 0.15^{5-3} = 0.138$$

The probability of getting exactly three adults who know Twitter among five randomly selected adults is 0.138.

5.4 Mean, Variance, Standard Deviation of Binomial Distribution

Note: Because a binomial distribution describes a population, the **mean** and **standard deviation** are **parameters**, not statistics.

Mean	$\mu = n.p$ p = Number of Successes expected to occur from the procedure
Variance	$\sigma^2 = npq$
Standard Deviation	$\sigma = \sqrt{npq}$

Ex: A certain population is made up of 80% Mexican-Americans. Select a Jury of 12. Find Mean & Standard Deviation. Success = Selecting a Mexican-American, Failure = Selecting any other ethnicity.

Soln.: $n = 12$ $p = 0.80$ $q = 0.20$

$$\text{Mean} = \mu = n.p = 12 \cdot (0.80) = 9.6$$

$$\text{Standard Deviation} = \sigma = \sqrt{npq} = \sqrt{12 \cdot (0.80) \cdot (0.20)} = 1.39$$

Usual / Unusual Values

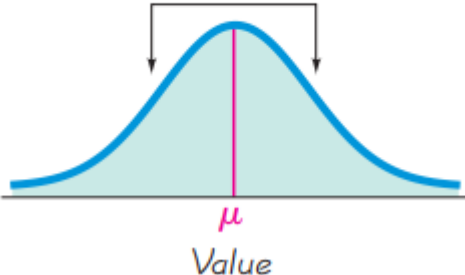
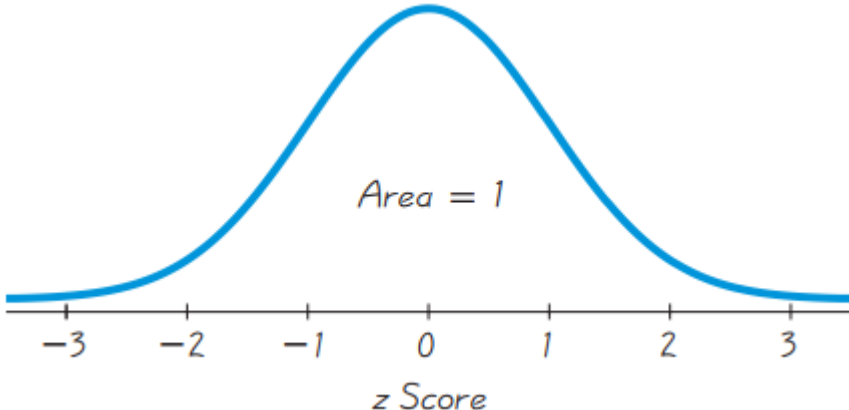
Values are considered **unusually high** or **unusually low** if they differ from the mean by more than 2 standard deviations, as described by the following:

Range Rule of Thumb

- **Maximum usual value:** $\mu + 2\sigma$
- **Minimum usual value:** $\mu - 2\sigma$

6.2 Continuous Variable and Normal Distribution

Studying a **Continuous Random Variable**.

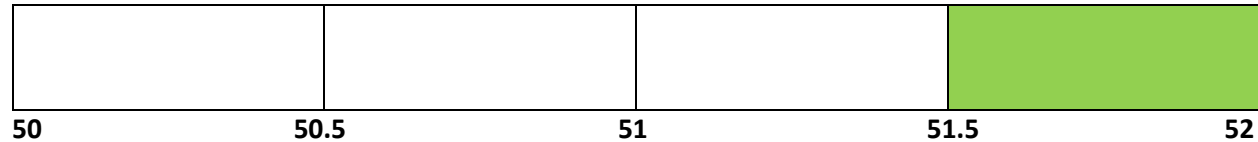
Uniform Distribution	<p>A continuous random variable has a uniform distribution if its values are spread evenly over the range of possibilities. The graph of a uniform distribution results in a rectangular shape. All values have the same probability of occurrence.</p> <p>Properties:</p> <ol style="list-style-type: none"> 1. The area under the graph of a probability distribution is equal to 1. 2. There is a correspondence between area and probability (or relative frequency), so some probabilities can be found by identifying the corresponding areas in the graph.
Normal Distribution	<p>If a continuous random variable has a distribution with a graph that is symmetric and bell-shaped, it is considered to have a normal distribution.</p> <p><i>Curve is bell-shaped and symmetric</i></p>  <p>Normal distributions occur often in real applications, and they play an important role in methods of inferential statistics.</p>
Standard Normal Distribution <p>The standard normal distribution is a normal distribution with the parameters of $\mu = 0$ and $\sigma = 1$. The total area under its density curve is equal to 1. <i>There is a correspondence between area and probability.</i></p> 	

Uniform Distribution

Ex: Class duration is 50 minutes but it always ends between 50 and 52 minute. Find the probability that the class will end between 51.5 and 52 minute.

Soln.:

.5



Probability = Total Area Under the curve must equal 1 = $2 * .5 = 1$

Area between 51.5 and 52 = $.5 * .5 = .25 = 25\%$

Convert Normal Distribution to Standard Normal Distribution

Steps:

1. Find Z-Score
2. Draw a Pic
3. Find Area [Use Table or Calculator] [Note: Table always gives Area to the left of Z-Score]

Ex.: Testing Thermometer: A thermometer that has mean 0 and standard deviation 1 has its readings normally distributed. Find Probability that a thermometer will have a reading of less than 1.58.

Soln.: $Z = (x - \mu) / \sigma = (1.58 - 0) / 1 = 1.58$

From z Scores Table Cumulative Area from the LEFT for $Z(1.58) = .9429$

Probability = 94.29%

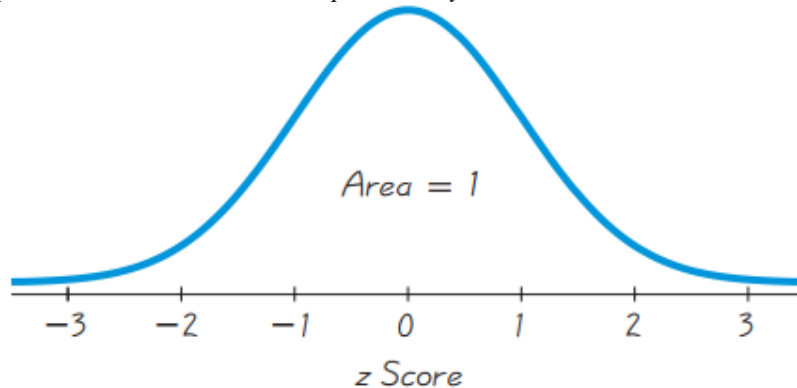
6.3 Standard Normal Distribution, Z-Score, Standard Score

Standard Normal Distribution

The **standard normal distribution** is a **normal distribution** with parameters of $\mu = 0$ and $\sigma = 1$.

The **total area under its density curve is equal to 1**.

There is a correspondence between area and probability.



Z-Score	Distance of the data point from mean, along the horizontal scale of the standard normal distribution $Z = (x - \mu) / \sigma$
Area	Region under the curve, it is the probability of the given z-score .
Data Point	We can find data value from z-score. $x = \sigma \cdot z + \mu$
Critical Values	For a normal distribution, a critical value is a z score on the borderline separating the z scores that are likely to occur from those that are unlikely. Common critical values are $z = -1.96$ and $z = 1.96$

Convert Normal Distribution to Standard Normal Distribution

When working with a normal distribution that is nonstandard (with a mean different from 0 and/or a standard deviation different from 1), transform value x to a z score, then proceed with methods.

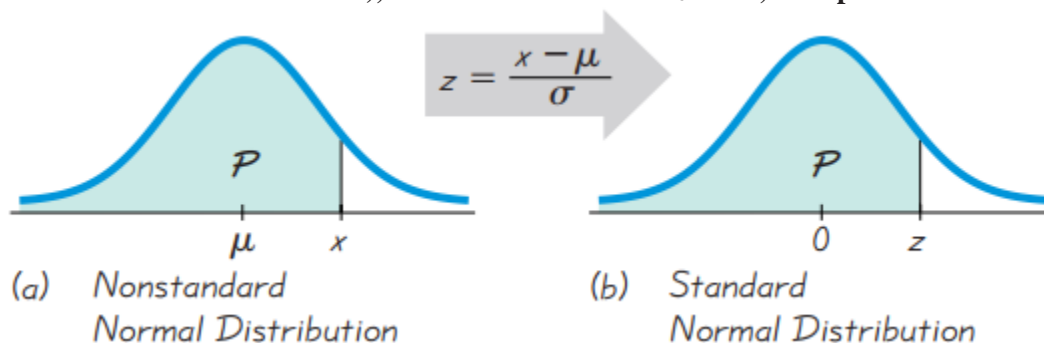


Figure illustrates the **conversion** from a **nonstandard** to a **standard normal distribution**. The area in any normal distribution bounded by some score x (figure on left) is the same as the area bounded by the equivalent z score in the standard normal distribution (figure on right). This shows that when working with a nonstandard normal distribution, we can use Table the same way it as normally used, provided that x value is first converted to z scores.

Procedure for Finding Areas (Probability) with a Nonstandard Normal Distribution

1. Find Z-Score - For each relevant value x for which the area (probability) is to be found, use formula $Z = (x - \mu) / \sigma$ to convert that value to the equivalent z score.
2. Draw a Pic (Curve of Standard Normal Distribution), label the mean and any specific x values, and then shade the region representing the desired probability.
1. Find Area of the shaded region [Use Table or Calculator] [Note: Table always gives Area to the left of Z-Score] This Area gives the Probability.

Ex: A population of men having a **mean weight of 172 lbs** and a **standard deviation of 29 lbs**, find the probability that a random selected person will have a weight of **less than 174 lbs**.

Soln.: $\mu = 172$ $\sigma = 29$ $x = 174$

Step 1. $z = (174 - 172) / 29 = 0.07$

Step 2. Draw the picture of standard normal distribution

Step 3 Find Area for calculated z -score from Table = .5279

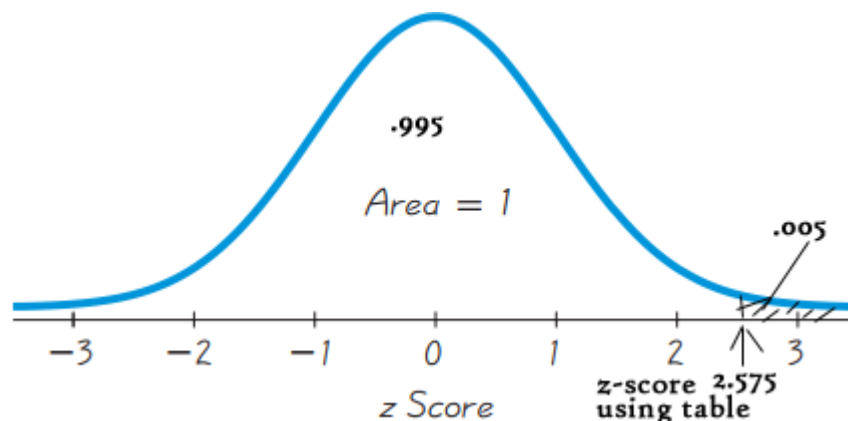
So, 52.79% are below 174 lbs.

Procedure for finding a z-Score from a Known Area

1. Draw a bell-shaped curve and identify the region under the curve that corresponds to the given probability. If that region is not a cumulative region **from the left**, work instead with a known region that is a cumulative region from the left.
2. Use Table to **find the z score**. With Table, **use the cumulative area from the left**, **locate the closest probability in the body of the table**, and **identify the corresponding z score**.

Ex. Weight Mean $\mu = 172$ lbs. $\sigma = 29$ lbs. What weight separates the **lightest 99.5%** from the heaviest .5% ?

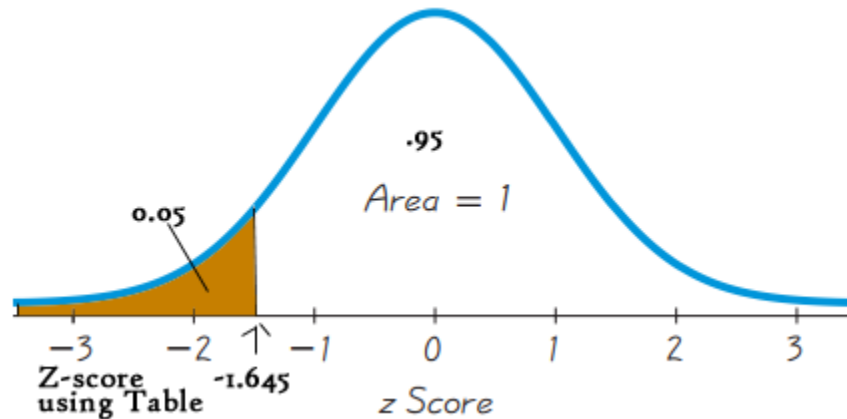
Soln: Draw the picture: From table find the Z-score for the given Area of .995. From table this is 2.575.



$$x = 172 + 2.575 (29) = 246.675 \text{ lbs.}$$

Ex.: Grip reach for women is Normally Distributed. **Mean is 27.0 inches** and **Standard Deviation is 1.3 inches**. **Find the grip reach** that represents the **longest 95% women**.

Soln. Draw the picture: From table find the Z-score for the given Area of .05. From table this is -1.645.



$$x = \mu + \sigma \cdot z = 27 - 1.3 (1.645) = 24.86$$

6.4 Sampling Distributions

- Statistics is all about USING SAMPLES to **approximate** POPULATIONS.
- We use SAMPLE STATISTIC to **estimate** POPULATION PARAMETERS.

DEFINITION The **sampling distribution of a statistic** (such as a sample mean or sample proportion) is the distribution of all values of the statistic when all possible samples of the same size n are taken from the same population. (The sampling distribution of a statistic is typically represented as a probability distribution in the format of a table, probability histogram, or formula.)

Sampling Distribution

What is Sampling Distribution?

How many Samples of 'n' are possible out of Population of size 'N'? If N is 27 and n is 5 and "Order does not matter", then combination:

$${}_N C_n = {}_{27} C_5 = 80730$$

Assume that you took all possible samples of size 'n' out of population e.g. all 80730 samples of 5 possible out of 27, and found out for Each Sample what the Mean was i.e. one statistic from one Sample, and you did it 80730 times, i.e. you have 80730 Samples and **organize all that statistic into a table, then you can call it Sampling Distribution.**

What Sampling Distribution Does?

Takes all the different Samples of a given size of population, finds out the key statistic (such as Mean, Variance, etc.) you are looking for from each of those Samples, and organizes puts them in a table.

Thus in above example if we took Mean of each Sample then in the end we will have 80730 different values of mean. If Sample Size or Population Size changes, then Sampling Distribution changes.

Sample Distribution of Proportion: Distribution of all possible sample Proportions.

We have a population that has 3 items in it, numbers [1, 2, 5]. The proportion of odd numbers in this population is two third (2/3). We took Samples of Size two (2) with replacement and listed out all Samples, the probability of any one of that Sample is one ninth (1/9), and the proportion of odd numbers in each Sample is also listed out.

Population: 1, 2, 5 Proportion of odd numbers = $P = 2/3$

Sampling Distribution of every possible sample of a certain size (2) out of a population of given size (3).

\hat{p} – Sample Proportion – In the table below it represents the proportion of odd numbers in the sample.

SAMPLE	\hat{p}	Probability	$\hat{p} \cdot P(x)$	Notes
1, 1	1.0	1/9	1/9	<ul style="list-style-type: none"> Sample Proportion will be a good estimator (targets) of Population Proportion. Sample Mean will be a good estimator of Population Mean Sample Variance will be a good estimator of Population Variance Sample Standard Deviation is NOT A GOOD estimator of Population Standard Deviation. It systematically underestimates it.
1, 2	0.5	1/9	1/18	
1, 5	1.0	1/9	1/9	
2, 1	0.5	1/9	1/18	
2, 2	0	1/9	0	
2, 5	0.5	1/9	1/18	
5, 1	1.0	1/9	1/9	
5, 2	0.5	1/9	1/18	
5, 5	1.0	1/9	1/9	
		Total	6/9 = $2/3$	This is same as proportion of odd numbers in Population

Sample Distribution of Mean: Distribution of all possible sample means.

SAMPLE	\bar{X}	Probability	$\bar{X} \cdot P(x)$	Notes
1, 1	1.0	1/9	1/9 = 2/18	Mean of Population = (1 + 2 + 5)/3 = $8/3$
1, 2	1.5	1/9	3/18	
1, 5	3	1/9	3/9 = 6/18	
2, 1	1.5	1/9	3/18	
2, 2	2	1/9	2/9 = 4/18	
2, 5	3.5	1/9	3.5/9 = 7/18	
5, 1	3	1/9	3/9 = 6/18	
5, 2	3.5	1/9	3.5/9 = 7/18	
5, 5	5	1/9	5/9 = 10/18	
		Total	48/18 = $8/3$	This is same as mean of numbers in Population

Does each Sample targets/estimates Population? NO. It's their average that does that.

The above applies to Sample Variance also.

6.5 Central Limit Theorem, Using Z-Scores, Standard Scores

Sampling distribution of sample means tends to be a **normal distribution as the sample size increases**.

We use the sampling distribution of sample means to use a normal distribution for some very meaningful applications by applying the central limit *theorem*.

No matter what the population looks like, whether it is uniform distribution, skewed to the left, skewed to the right, normal, U-shaped, not matter what shape population is, if you take Samples of **Size > 30**, if you take Sample Statistic such as Mean and organize those then their **distribution will be Normal**.

Central Limit Theorem	For all samples of the same size n with n > 30 , the sampling distribution of \bar{x} can be approximated by a normal distribution with mean μ and standard deviation σ / \sqrt{n} .						
n > 30	Population is normally distributed or not , Sample Means will have Normal Distribution. <table> <tr> <td>Mean of all values of \bar{x}:</td><td>$\mu_{\bar{x}} = \mu$</td></tr> <tr> <td>Standard deviation of all values of \bar{x}</td><td>$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$</td></tr> <tr> <td>z score conversion of \bar{x}</td><td>$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$</td></tr> </table>	Mean of all values of \bar{x} :	$\mu_{\bar{x}} = \mu$	Standard deviation of all values of \bar{x}	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$	z score conversion of \bar{x}	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
Mean of all values of \bar{x} :	$\mu_{\bar{x}} = \mu$						
Standard deviation of all values of \bar{x}	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$						
z score conversion of \bar{x}	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$						
n ≤ 30	The above formulas will work then Sample MUST come from Normally Distributed Population .						
n ≤ 30	We know nothing about Population distribution . CANNOT APPLY THESE FORMULAS .						

Z-Scores: Individual vs. Group

Individual Data Value	Group of Samples
$z = \frac{x - \mu}{\sigma}$	$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$

Ex: Population of men: **Mean 172 lbs.** **Std. Deviation 10 lbs.**

- Find probability that **a randomly selected man** will weigh **> 175 lbs.**
- Find probability that **group of 20 men** will have average weight **> 175 lbs.**

Soln.:

- $Z = (175 - 172) / 10 = 3 / 10 = 0.30$
A-left = Find area for this Z-score from Table to the left of this Z-score
Probability = 1 - A-left
- $Z = (175 - 172) / (10/\sqrt{20}) = 3 / (10 / 4.472) = 3 / 2.236 = 1.34$
A-left = Find area for this Z-score from Table to the left of this Z-score = 0.9099
Probability = 1 - A-left = 1 - 0.9099 = 0.0901

Rare Event Rule for Inferential Statistics

If, under a given assumption, the probability of a particular observed event is extremely small (such as less than 0.05), we conclude that the assumption is probably not correct.

Ex: Body temperature average (Population Mean) = 98.6 Standard Deviation = 0.62 Sample $n = 106$.
What is the probability that the average temperature of sample (Sample Mean) will be 98.2% or Lower?

Soln.: $m = 106$ $\mu = 98.6$ $\sigma = 0.62$ $\bar{x} = 98.2$

$$Z = (98.2 - 98.6) / (0.62 / \sqrt{106}) = -0.4 / .06022 = -6.64$$

Z is Negative 6.64

By looking in the Table the Area (Probability) left of this Z-score is .0001

By Calculator this Area (Probability) is 0.000000000016

Thus the probability is indicating a very rare event.

Ex: Cans of regular Coke are labeled to indicate that they contain **12 oz.** Data Set lists measured amounts for a sample of Coke cans. The sample statistics are $n = 36$ and $\bar{x} = 12.19$ oz. Assuming that the Coke cans are filled so that $\mu = 12$ oz (as labeled) and the population standard deviation is $\sigma = 0.11$ oz (based on the sample results), **find the probability that a sample of 36 cans will have a mean of 12.19 oz or greater.** Do these results suggest that the Coke cans are filled with an amount greater than 12.00 oz?

Soln.: Requirement check: We can use the normal distribution if the original population is normally distributed or $n > 30$. The sample size of $n = 36$ is greater than 30, so **we can approximate the sampling distribution of \bar{x} with a normal distribution.**

Parameters of this normal distribution:

$$\mu_{\bar{x}} = \mu = 12$$

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} = 0.11 / \sqrt{36} = 0.018333$$

$$Z = (12.19 - 12.00) / 0.018333 = 10.36$$

In Table, $z = 10.36$ is off the chart, but for values of z above 3.49, we use 0.9999 for the cumulative left area. We therefore conclude that the area on right is 0.0001.

Interpretation: The result shows that if the mean amount in Coke cans is really 12.00 oz, then there is an **extremely small probability of getting a sample mean of 12.19 oz or greater when 36 cans are randomly selected.**

Because we did obtain such a sample mean, there are two possible explanations: (1) Either the **population mean really is 12.00 oz** and **the sample represents a chance event that is extremely rare**; or (2) the population mean is actually greater than 12.00 oz. Because the probability is so close to 0, it seems more reasonable to conclude that the population mean is greater than 12.00 oz. It appears that Coke cans are being filled with more than 12.00 oz. However, the sample mean of 12.19 oz suggests that the mean amount of overfill is very small. It appears that the Coca-Cola company has found a way to ensure that very few cans have less than 12 oz while not wasting very much of their product.

7.2 Confidence Interval for Population Proportion

Major activities of Inferential Statistics are:

1. **Use Sample Data to estimate values of Population Parameters** (such as Population Proportion and Population Mean).
2. **Test Hypothesis (Claims) made about Population Parameters.**

Confidence Interval

Let's say if I see 6 blonde people in a classroom, based on this observation can I say that $1/6^{\text{th}}$ of population, students in college, have blonde hair? Does it work? Not so well.

We don't know how good that point estimate for that sample is? How good that $1/6^{\text{th}}$ is? May be it's true, may be not. OK, so with one value of our Sample we don't know how accurate it is, there is definitely some error included in our calculation because of Sample Variability, every sample is not a perfect reflection of Population, therefore that point estimate from that Sample isn't going to be perfect and represent your population.

Therefore, instead of a Single Statistic we do a Range. We are going to create an Interval, so that we are confident that our Population Proportion falls in that Range (Interval).

So, instead of using one Sample Statistic to represent Entire Population, we create an interval (like a range).

So, that we have a certain level of Confidence that our actual Population Proportion falls in that range.

This is called **Confidence Interval.**

Point Estimate

A single value from a Sample used to approximate a Population Parameter. The sample proportion \hat{p} is the **best point estimate** of the population proportion p .

- p = Population Proportion of Success
- \hat{p} = Sample Proportion of Success = x (Count of Success) / n (Count of Trials)
- \hat{q} = Sample Proportion of Failure = $1 - \hat{p}$
- \hat{p} (Sample Proportion) is a Point Estimate for p (Population Proportion)

Is Point Estimate a great estimate? We don't know.

What is needed?

1. Random Sample
2. Conditions for Binomial
 - Fixed # of Trials
 - Trials are Independent
 - Two Outcomes: Success / Failure
 - $np \geq 5$ $nq \geq 5$

Why use Confidence Interval? Touch Therapist in 280 Trials, chooses correct Hand 123 Times.

$$\hat{p} = 123 / 280 = 0.44$$

Is this a good estimate of Population Proportion? Don't know how good/accurate this point estimate is. This is why we use Confidence Interval.

Critical Value

A **critical value** is the number (**Z-score**) on the borderline **separating** sample statistics that are **likely to occur** from those that are **unlikely**, in other words it separates likely region from unlikely region.

Example: In a Standard Normal Distribution with a mean of 0 and a standard deviation of 1, find the **test score that separates** the **bottom 2.5%** and the score that separates the **top 2.5%**.

Soln.: Using Table we find the z score located to the left, we search the body of the table for an **area of 0.025**, the result is **$z = -1.96$** . To find the z score located to the right, we search the body of Table for an **area of 0.975**, note that the Table always gives cumulative areas from the left, the result is **$z = 1.96$** . The values of $z = -1.96$ and $z = 1.96$ separate the bottom 2.5% and the top 2.5%.

**Why is it called a Critical Value?**

These two values **$z = -1.96$** and **$z = 1.96$** are called **Critical Values**. The values below $z = -1.96$ are unlikely, because only 2.5% of the population have scores below -1.96, and the values above $z = 1.96$ are unlikely because only 2.5% of the population have scores above 1.96.

The expression **Z_α** denotes the Z-score with an **area of α** to its right. The number **$Z_{\alpha/2}$** is a critical value that is a Z-score with the property that it separates an area of $\alpha/2$ in the right tail of the standard normal distribution.

Confidence Interval	A confidence interval (or interval estimate) is a range (or an interval) of values used to estimate the true value of a population parameter . A confidence interval is sometimes abbreviated as CI.		
Confidence Level	The confidence level is the probability $1 - \alpha$ (such as 0.95, or 95%) that the confidence interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. (The confidence level is also called the degree of confidence , or the confidence coefficient .) Tells us how Confident/Sure we are that the actual value of our Population Parameter will fall within the given Range/ Confidence Interval. $= 1 - \alpha$ $\alpha = \text{Complement of Confidence Level}$		
Most Common Confidence Levels	Corresponding Values of α	Critical Value $Z_{\alpha/2}$	
90% (or .90) confidence level	$\alpha = 0.10$	1.645	
95% (or .95) confidence level	$\alpha = .05$	1.96	
99% (or .99) confidence level	$\alpha = .01$	2.575	

Example Statement: The 95% Confidence Interval for P is $0.381 < P < 0.497$

Interpretation: I don't know what 'P' actually is, But I am 95% sure that it falls in that range.

Margin of Error

What is Margin of Error?

Depending upon how confident you want to be, you got to have certain amount of error.

- The more confident you want to be the larger your error is going to be.
- The more confident you want to be the larger your range got to be.
- The less confident you want to be, the smaller the range is going to be.

What Margin of Error does is it lets the Critical Value determine the maximum distance for how confident you are, between point estimate and population parameter.

Margin of Error: E = The maximum difference between \hat{p} and p .

$$E = Z_{\alpha/2} \sqrt{(\hat{p} \cdot \hat{q}) / n}$$

Steps:

1. Find \hat{p} \hat{q} n
2. Use Confidence Level to Find $Z_{\alpha/2}$ (Critical Value)
3. Find $E = Z_{\alpha/2} \sqrt{(\hat{p} \cdot \hat{q}) / n}$
4. Confidence Interval = $\hat{p} - E < p < \hat{p} + E$

Example: Touch Therapist, 280 Trials, 123 Correct Identification. Construct 95% Confidence Interval of p .

Solution:

Step 1: Find \hat{p} \hat{q} n

- $n = 280$
- Proportion of Success = $\hat{p} = x/n = 123/280 = 0.4393$
- $\hat{q} = 1 - \hat{p} = 1 - 0.4393 = 0.5607$

Step 2: Find Critical Level using Confidence Level

- Confidence Level = 95% Hence $\alpha = .05$
- From Table $Z_{\alpha/2}$ [For 95%] = 1.96

Step 3: $E = Z_{\alpha/2} \sqrt{(\hat{p} \cdot \hat{q}) / n} = 1.96 \sqrt{(.4393)(.5607)/280} = 0.0581$

Step 4: Confidence Interval $p = (.4393 - .0581) < p < (.4393 + .0581)$

$$= .3812 < p < .4974$$

I don't know exactly what p is, I don't know if \hat{p} is close to p , but I am 95% sure that the maximum difference between p and .4393 is .0581, and that p will fall in this range.

Note: As Sample Size goes up, E decreases, Confidence Interval decreases.

Finding Required Sample Size of a Given E

$E = Z_{\alpha/2} \sqrt{(\hat{p} \cdot \hat{q}) / n}$	$n = (Z_{\alpha/2})^2 \hat{p} \cdot \hat{q} / E^2$
---	--

Example: Determine % people who use E-mail (Confidence Level 95%). What does Sample Size need to be for E = 4%. Case 1. Consider that 16.9% people used Email in 1997 Case 2. Assume we know nothing about previous Email usage.

Solution:

- **Case 1.**

$$n = ? \quad \hat{p} = .169 \quad \hat{q} = 1 - \hat{p} = 1 - .169 = 0.831$$

$$n = (1.96)^2 (.169) (.831) / (.04)^2 = 337.19 = 338$$

For 4% Margin of Error at Confidence Level 95%, and considering that 16.9% people used Email, the Sample Size needs to be 338.

- **Case 2**

$$n = (1.96)^2 (.25) / (.04)^2 = 600.25 = 601$$

For 4% Margin of Error at Confidence Level 95%, and considering no previous knowledge about Email usage, the Sample Size needs to be 601.

Example: In a study of 1300 Randomly Selected Medical Law Suits, 900 were dropped.

- Find **Best Point Estimate** for Population Proportion - $\hat{p} = 900 / 1300 = .6923$

Is this the percentage of all law suits dropped in America? No idea. But this is the Best Estimate I have for the population. This is the only one we have.

- Construct a **99% Confidence Interval**.

$$\hat{p} = .6923 \quad \hat{q} = 1 - .6923 = .3077 \quad n = 1300$$

$$Z_{\alpha/2} = 2.575 \text{ (for 99\%)}$$

$$E = 2.575 \sqrt{(.6923)(.3075) / 1300} = .033$$

$$\text{Confidence Interval} = (.6923 - .033) < p < (.6923 + .033)$$

Given a Confidence Interval Find \hat{p} and E

$$\text{Confidence Interval} = \hat{p} - E < p < \hat{p} + E$$

$$\hat{p} = ((\hat{p} - E) + (\hat{p} + E)) / 2$$

$$E = ((\hat{p} + E) - (\hat{p} - E)) / 2$$

Example: At 95% Confidence Level the Confidence Interval is $0.58 < p < 0.81$.

$$\hat{p} = (.81 + .58) / 2 = .695 \quad E = (.81 - .58) / 2 = .115$$

7.3 Confidence Interval Estimating Population Mean

What is required?

1. Random Sample
2. Population **Standard Deviation is KNOWN**
3. $n > 30$ OR Population Normally Distributed

Point Estimate for Sample Mean	\bar{X}
Margin of Error	$E = Z_{\alpha/2} (\sigma / \sqrt{n})$
Confidence Interval	$(\bar{X} - E) < \mu < (\bar{X} + E)$

Example: Random Sample of 40 students, Average resting heart rate = 76.3, Population Standard Deviation = 12.5. Construct a 99% Confidence Interval for the average resting heart rate of population.

Given 99% Confidence Level $Z_{\alpha/2} = 2.575$

$\sigma = 12.5$ $n = 40$ $\bar{X} = 76.3$

$E = Z_{\alpha/2} (\sigma / \sqrt{n}) = (2.575) (12.5 / \sqrt{40}) = 5.08$

Confidence Interval = $(76.3 - 5.08) < \mu < 76.3 + 5.08 = 71.22 < \mu < 81.38$

I am 99% Confident that Average Resting Heart rate of population will be between this.

Finding Required Sample Size for a given Margin of Error

$E = Z_{\alpha/2} (\sigma / \sqrt{n})$	$n = [Z_{\alpha/2} \cdot \sigma / E]^2$
--	---

7.4 Confidence Interval Estimating Population Mean

IMPORTANT: Standard Deviation is NOT KNOWN

Since we don't know σ we cannot use a Z-score. We have to use **T-score** (Known as such because it was developed in response to a student who posted in a journal by pseudo name T).

Required:

1. Random Sample
2. $n > 30$ or Sample from a Normally Distributed Population

Formulas:

$$Z = (\bar{X} - \mu) / (\sigma / \sqrt{n})$$

$$T = (\bar{X} - \mu) / (s / \sqrt{n})$$

Critical Values are given by $T_{\alpha/2}$

Degree of Freedom = $n - 1$ (Sample Size minus one)

Example: 23 Samples from a Normally Distributed Population. Find C.V. (Critical Value $T_{\alpha/2}$) for Confidence Level 95%.

$n = 23$ $\alpha = .05$ D.F. = 22 **From Table $T_{\alpha/2} = 2.074$ (For D.F. 22 and $\alpha = .05$)**_{sdf}

Note: For this 95% confidence level $Z_{\alpha/2}$ value is 1.96

- Most of the time T-scores are used for smaller samples < 30 .
- Instead of Sample Size we use Degree of Freedom
- Thus T-distribution for same level of confidence gives wider spread, for every sample size it is different.
- BUT as sample size goes up T-score becomes value wise same as Z-score.

Margin of Error $E = T_{\alpha/2} (s/\sqrt{n})$

Confidence Interval $= (\bar{X} - E) < \mu < (\bar{X} + E)$

Example: Construct a **Confidence Interval 95%** for the **average age** of people denied promotion in a **Random Sample** of **23 people**. The **average age** was **47.0** with **Standard Deviation of 7.2**. Assume this sample comes from a Population that is Normally Distributed.

$n = 23$ $DF = 22$ $\bar{X} = 47.0$ $\alpha = .05$ $s = 7.2$

$T_{\alpha/2} = 2.074$

$E = T_{\alpha/2} (s/\sqrt{n}) = 2.074 (7.2 / \sqrt{23}) = 3.1$

Confidence Interval $= (\bar{X} - E) < \mu < (\bar{X} + E) = 47.0 - 3.1 < \mu < 47.0 + 3.1$
 $= 43.9 < \mu < 50.1$

Interpretation: We are 95% sure that people being denied promotion are between 43 and 51.

7.5 Confidence Interval Estimating Variance & Std. Deviation

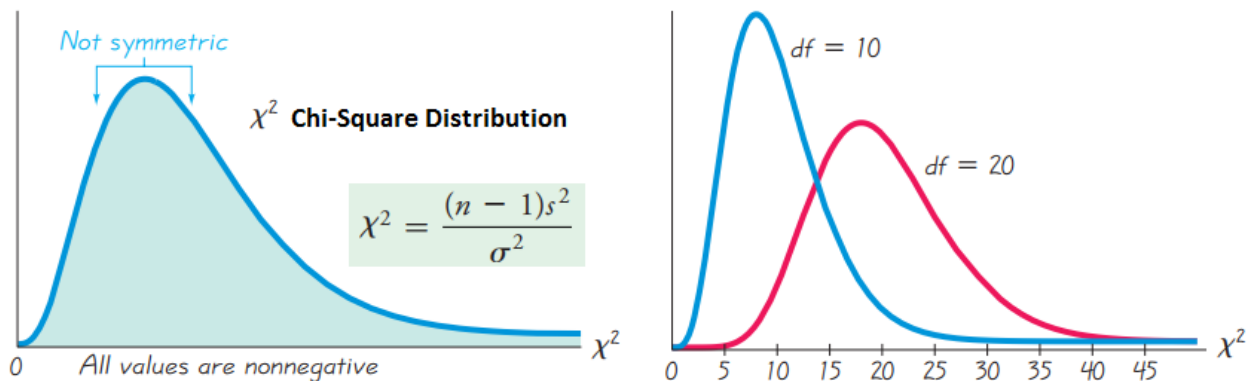
Using a sample standard deviation s (or a sample variance s^2) to **estimate** the value of the corresponding **population standard deviation σ** (or population variance σ^2). The methods require that we use a **chi-square distribution**.

Point Estimate:	The sample variance s^2 is the best point estimate (or single value estimate) of the population variance σ^2 . The sample standard deviation s is commonly used as a point estimate of σ (even though it is a biased estimator).
Confidence Interval	<i>Confidence Interval</i> estimate of a population standard deviation (or population variance) is constructed using χ^2 (chi-square) distribution .
Chi-Square Distribution	In a normally distributed population with variance σ^2 , if we randomly select independent samples of size n and, for each sample, compute the sample variance s^2 , the sample statistic $\chi^2 = (n - 1)s^2/\sigma^2$ has a sampling distribution called the chi-square distribution .
Finding Critical Values of χ^2	Right-tailed critical value is denoted χ^2_R and left-tailed critical value is denoted by χ^2_L . These critical values are found by using Table and they require a value for the number of <i>degrees of freedom</i> . Note: <ul style="list-style-type: none"> • Z-score Table gives C.V.(Critical Value) for area in Left • T-score Table gives C.V.(Critical Value) for area in Tails

	<ul style="list-style-type: none"> χ^2-Table gives C.V.(Critical Value) for area in Right
Degrees of Freedom	In general, the number of degrees of freedom (or df) for a collection of sample data is the number of sample values that can vary after certain restrictions have been imposed on all data values . Number of degrees of freedom is sample size minus 1. degrees of freedom: $df = n - 1$

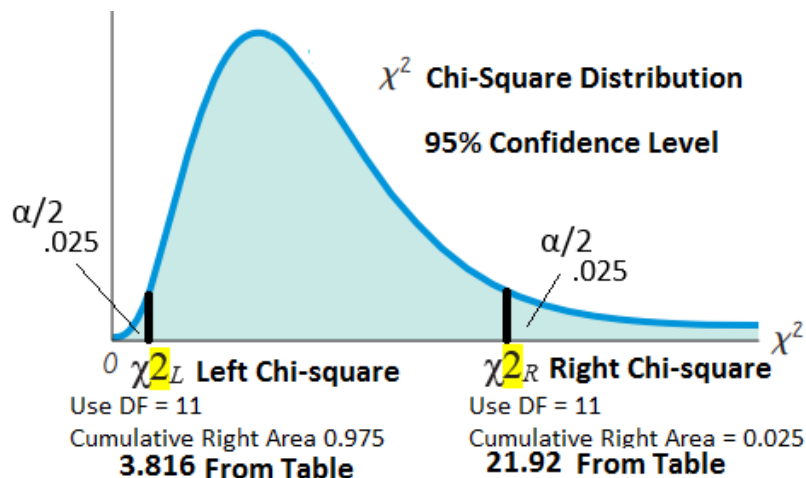
Properties of the Chi-Square Distribution

1. The chi-square distribution is **not symmetric**, unlike the normal and Student t distributions. (As the number of degrees of freedom increases, the distribution becomes more symmetric)
2. The values of chi-square can be zero or positive, but they **cannot be negative**.
3. The chi-square distribution is **different for each number of degrees of freedom**.



Example

$n = 12$ Confidence Level = 95% Find C.V. (Critical Value) for $\alpha = .05$



Confidence Interval:

$$\frac{(n-1)s^2}{\chi^2_L} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_R}$$

Degrees of Freedom	0.995	0.99	0.975
1	—	—	0.001
2	0.010	0.020	0.051
3	0.072	0.115	0.216
4	0.207	0.297	0.484
5	0.412	0.554	0.831
6	0.676	0.872	1.237
7	0.989	1.239	1.690
8	1.344	1.646	2.180
9	1.735	2.088	2.700
10	2.156	2.558	3.247
11	2.603	3.053	3.816

Degrees of Freedom	Area to the Right of the Critical Value							
	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025
1	—	—	0.001	0.004	0.016	2.706	3.841	5.024
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143
5	0.412	0.554	0.831	1.145	1.610	9.236	11.071	12.833
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920

Example

Sample of 10 appliances, Standard Deviation of 0.15 Construct a Confidence Interval (95%) for Variance.

- Given: Degree of Freedom = $10 - 1 = 9$ $s = 0.15$ $\alpha = 0.5$ $\alpha/2 = .025$
- From Table: $\chi_L^2 = 2.7$ $\chi_R^2 = 19.023$

Confidence Interval Calculation: $(9 (.15)^2 / 19.023) < \sigma^2 < (9 (.15)^2) / 2.7 = .103 < \sigma^2 < .274$

Formula Used: $\frac{(n-1)s^2}{\chi_L^2} < \sigma^2 < \frac{(n-1)s^2}{\chi_R^2}$

8.2 Hypothesis Testing

Testing whether or not a **CLAIM** is valid. In statistics, a **hypothesis is a claim** or statement **about a property of a population**. A **hypothesis test** (or **test of significance**) is a procedure for testing a claim about a property of a population. (Testing Population Parameter claims such as p , μ , σ)

Examples of claims:

- Most people get their jobs through networking. $P > .50$
- The average payload of trucks on Highway 99 is 18,000 lbs $\mu = 18000$

Rare Event Rule: "If the Probability of an assumption occurring is "Very Small" THEN the ASSUMPTION is probably Incorrect."

Parts of a Hypothesis Test: NULL Hypothesis Vs Alternative Hypothesis

Null Hypothesis: H_0 – States that the Population Parameter (Mean μ , Proportion p) is **EQUAL** to some value.

Example: $\mu = 15$ $p = .5$

Alternative Hypothesis: H_1 – States that the Population Parameter (Mean μ , Proportion p) has a value different than H_0 .

Example: $<, >, \neq$

$p < .53$ $p > .53$ $p \neq .53$

Note: If you want to support a claim you MUST state it as H_1 (Not H_0)

Example: Suppose you want to prove that your fertility drug works, they you will state:

$H_0: p = .50$ $H_1: p > .50$

How to Test Hypothesis:

- Start by assuming that H_0 is True.
- Then, use evidence to reach a CONCLUSION:
 - ✦ **REJECT H_0** : I have enough evidence to prove H_0 is WRONG.
 - ✦ **FAIL TO REJECT H_0** : I don't have enough evidence to prove H_0 wrong. Note: CANNOT ACCEPT H_0 . Thus if you fail to reject H_0 then you have not proven anything.

How to Identify H_0 and H_1 :

- State the Original Claim Symbolically.
- State the Opposite of the Original Claim as well
- **Note:** The Original Claim could be **H_0 or H_1** Depends on where the EQUALITY is.

Examples:

1. Mean of Fluid is **At Least** 12 oz. in each can.
 CLAIM: $\mu \geq 12$ -----> H_0 $\mu = 12$
 OPPOSITE: $\mu < 12$ -----> H_1 $\mu < 12$
2. Proportion of Male CEO's is **Greater Than** 0.5. (**Most** CEOs are male).
 CLAIM: $p > .50$ -----> H_1 $p > .50$
 OPPOSITE: $p \leq .50$ -----> H_0 $p = .50$
3. Mean weight of babies is **At Most** 8.9 lbs.
 CLAIM: $\mu \leq 8.9$ -----> H_0 $\mu = 8.9$
 OPPOSITE: $\mu > 12$ -----> H_1 $\mu < 8.9$
4. Mean IQ Score is 100.
 CLAIM: $\mu = 100$ -----> H_0 $\mu = 100$
 OPPOSITE: $\mu \neq 100$ -----> H_1 $\mu \neq 8.9$

Test Statistic

A **test statistic** is a **random variable** that is **calculated from sample data** and used in a **hypothesis test**. Test statistics is used to determine **whether to reject the null hypothesis**. It **compares given data with what is expected** under the null **hypothesis**.

Proportion p	$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$
Mean μ (σ Known)	$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$
Mean μ (σ Unknown)	$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$
Standard Deviation or Variance	$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$

Example: A survey of **706 companies** found that **61% of CEOs** were male. Claim MOST CEOs are male.

CLAIM: $p > .50$ -----> **H_1** $p > .50$

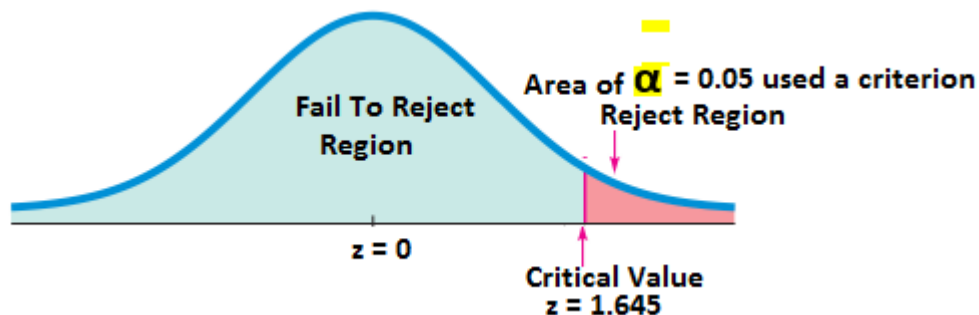
OPPOSITE: $p \leq .50$ -----> **H_0** $p = .50$

Test Statistic: $n = 706$ $p = .50$ $q = .50$ Population Proportion $\hat{p} = .61$

$$Z = (.61 - .50) / \sqrt{(.50)(.50) / 706} = .11 / .0188 = \mathbf{5.85}$$

How to Make a Decision

- **Significance Level:** α Common Values of $\alpha = .10$.05 .01
- **Critical Values:** Separates Rejection Region from the Fail to Reject Region.
- **Rejection Region:** If our Test Statistic falls into this region then Reject **H_0**



Interpretation: Since calculated Test Statistic Z is 5.85, **H_0** is Rejected, claim proven correct.

Only Two Decisions Possible:


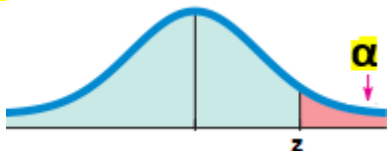
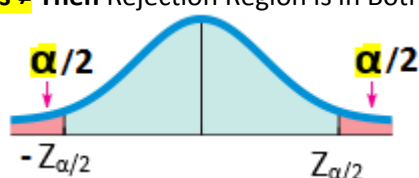
- **Reject H_0** – Consequently Accept H_1
- **Fail To Reject H_0** – You know nothing
 - a. **Reject H_0** If P-value $\leq \alpha$
 - b. **Fail to Reject H_0** If P-value $> \alpha$

Three Type of Test

The test statistic alone usually does not give us enough information to make a decision about the claim being tested. For that decision, we can use either the **P-value** approach or the **critical value** approach. Both approaches require that we first determine whether our hypothesis test is left-tailed, right-tailed, or two-tailed.

Critical region (or **rejection region**) corresponds to the **values of the test statistic** that **cause us to reject the null hypothesis**. Depending on the claim being tested, the critical region could be in the left tail, or it could be in the right tail, or it could be in the two extreme tails.

Type of test is determined by H_1

Left-Tail	If H_1 has $<$ Then Rejection Region is in Left Tail. 
Right-Tail	If H_1 has $>$ Then Rejection Region is in Right Tail 
Two-Tail	If H_1 has \neq Then Rejection Region is in Both Tails 

Two Methods:

2. **Traditional Method (Critical Value Method):** Reject H_0 if the Test Statistic falls in the reject region
3. **P-Value Method**

P-value Method

P-value is the Probability associated with Test Statistic. It is the **probability of getting a Test Statistic value that is at least as extreme as the one obtained using sample data, assuming that the null hypothesis is true.**

A p -value is a probability associated with your critical value. The critical value depends on the probability you are allowing for a Type I error. It measures the chance of getting results at least as strong as yours if the claim (H_0) were true.

Finding P-value

To find the P-value, first find the area beyond the test statistic, is the Critical Region in:

1. Left tail: P-value = area to the left of the test statistic
2. Right tail: P-value = area to the right of the test statistic
3. Two tails: P-value = twice the area in the tail beyond the test statistic

Example: A right-tailed test (**H_1 has $>$**) has a Test Statistic $z = 1.60$ Now Test Statistic $z = 1.60$ has an area of 0.0548 to its right. So, this right tailed test with test statistic $z = 1.60$ will have a P-value of 0.0548.

Example: Suppose you are testing a claim that the percentage of all women with varicose veins is 25%, and your sample of 100 women had 20% with varicose veins. Then the sample proportion $p=0.20$. The standard error for your sample percentage is the square root of $p(1-p)/n$ which equals 0.04 or 4%. You find the test statistic by taking the proportion in the sample with varicose veins, 0.20, subtracting the claimed proportion of all women with varicose veins, 0.25, and then dividing the result by the standard error, 0.04. These calculations give you a test statistic (standard score) of -0.05 divided by $0.04 = -1.25$. This tells you that your sample results and the population claim in H_0 are 1.25 standard errors apart; in particular, your sample results are 1.25 standard errors below the claim.

When testing $H_0: p = 0.25$ versus $H_a: p < 0.25$, you find that the p -value of -1.25 by finding the probability that Z is less than -1.25. When you look this number up on the above Z -table, you find a probability of 0.1056 of Z being less than this value.

Steps:

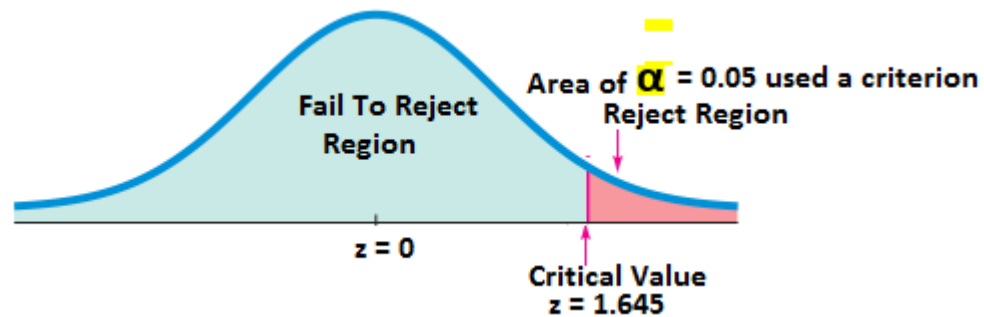
1. Identify the Claim (Hypothesis) to be tested, put it in Symbolic Form.
2. Give the Symbolic Form that **MUST** be TRUE when the ORIGINAL CLAIM is FALSE.
3. Of the two symbolic expressions let Null Hypothesis **H_0** be the one that contains EQUALITY sign. The other becomes Alternative Hypothesis.
4. Select the Significance Level based on the seriousness of Type 1 Error. Make α small if the consequences of rejecting a true are severe. The values of 0.05 and 0.01 are very common.
5. Identify the statistic that is relevant to this test and determine its sampling distribution (normal, t , chi-square).

6. Find the test statistic and find the P-value, draw the graph and show the Test Statistic and P-value.
7. Reject if the P-value is less than or equal to the significance level α . Fail to Reject if P-value is greater than α .
8. Address the Original Claim by restating the decision in simple non-technical terms.

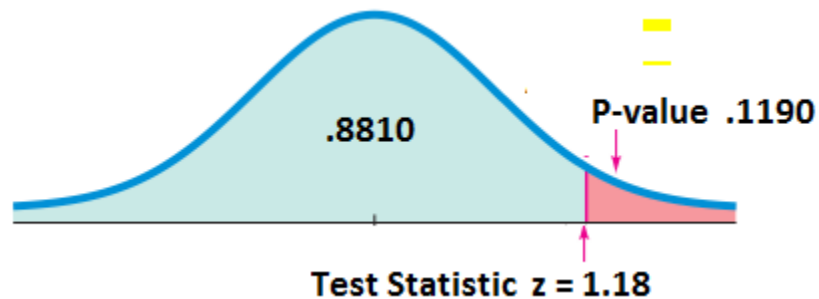
Example

$$\alpha = 0.05 \quad H_1: p > 0.25$$

Test Statistic of $z = 1.18$

Traditional Method:

Since, $z = 1.18$ It falls in Fail to Reject region. Hence, F.T.R. H_0

P-value Method:

Since P-value = .1190 > .05 Hence, F.T.R. H_0

8.3 Hypothesis Testing for Population Proportion

For testing Claims involving Percentages and Proportions (E.g. 35%, .20, MOST, etc.)

Requirements

1. Random Sample
2. $np \geq 5$ $nq \leq 5$
3. n = Sample Size (# of Trials)

$$\hat{p} = x / n \quad p = \text{Population Proportion} \quad q = 1 - p$$

Test Statistic

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Example: In a Sample of 300 Corporations, 183 of CEOs were male. Test the Claim that Most CEOs are male. Use a 0.05 Significance Level.

P-value Method

1. **CLAIM:** $p > .50$ H_1
 Opposite: $p \leq .50$ H_0
2. **Hypothesis H_0** $p = 0.50$
 H_1 $p > 0.50$
3. **Significance Level:** $\alpha = .05$

4. **Test Statistic**

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$\hat{p} = 183 / 300 = .61$$

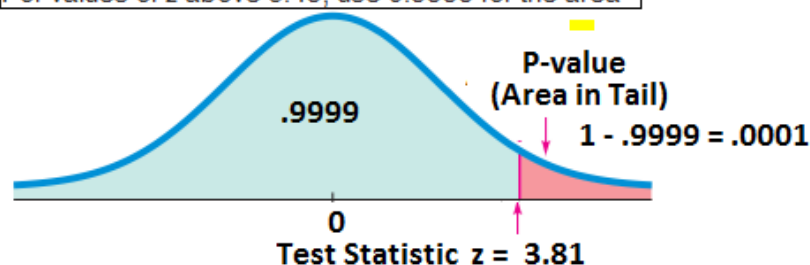
$$p = .50 \quad q = .50 \quad n = 300$$

$$z = (.61 - .50) / \sqrt{(.50)(.50) / 300} = 3.81$$

5. **Draw the graph**, note that H_1 has $>$ hence it is a right tail test.

As per Table

For values of z above 3.49, use 0.9999 for the area



6. **Make a Decision:**

P-value $\leq \alpha$ Reject H_0

P-value $> \alpha$ Fail to Reject H_0

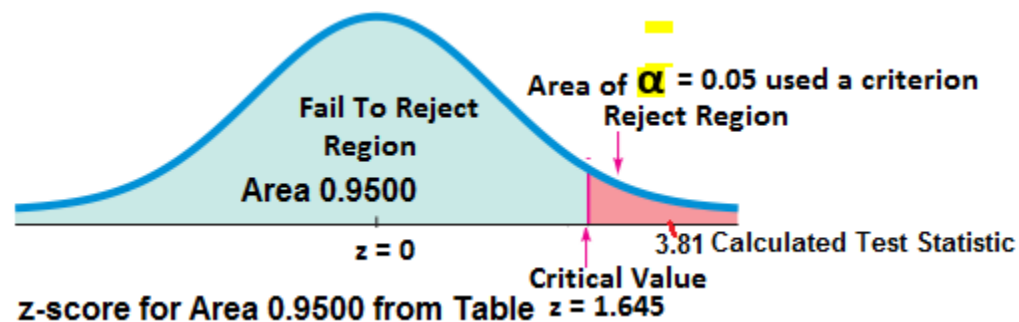
P-value ($= .0001$) $\leq \alpha$ ($= .05$) **Reject H_0 (Consequently H_1 True)**

7. **Interpret:** REJECT H_0 : "There is enough evidence to support the claim that Most CEOs are male".

Traditional Method

Steps 1 to 4 are identical to P-value Method. In step 4 of P-value method we got Test Statistic $z = 3.81$

Step 5. Draw the Graph



Step 6. Make a Decision: **Compare Test Statistic to Critical Value**

- If Test Statistic fall in Reject Region: Reject H_0
- If Test Statistic falls in Fail to Reject Region: Fail to Reject H_0

Test Statistic is higher than Critical Value, it falls in Reject Region.

Hence, **REJECT H_0**

ACCEPT H_1

Example 428 Green, 152 Yellow, Use a 0.05 Significance Level to test the claim that the Proportion of Yellow Pods is EQUAL to $\frac{1}{4}$.

P-value Method:

1. **CLAIM:** $p = .25$
OPP: $p \neq .25$
2. H_0 $P = .25$
 H_1 $P \neq .25$
3. **Significance Level** $\alpha = 0.05$
4. **Test Statistic**

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$x = 152 \quad n = 152 + 428 = 580 \quad p = .25 \quad q = .75$$

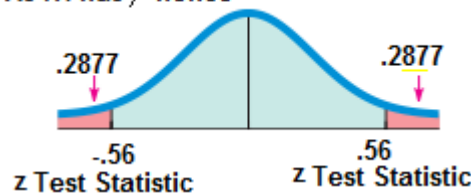
$$\hat{p} = x/n = 152 / 580 = .26$$

$$z = (.26 - .25) / \sqrt{(.25)(.75) / 580} = .56$$

5. **Draw the Graph:**

Since H_1 has \neq hence it is a **Two Tail Test**. In this case, we state z Critical Value twice, one positive one negative. Our P-value is the area associated with each of those Test Statistic. So, we are going to add the area in the right tail and the area in the left tail i.e. whatever those areas (probabilities) are, we are going to add up. Since Z-score table gives cumulative area from Left up to the z-score point hence to find the area for this z Test Statistic we look up into the Negative Z-Scores Table at $-z$ score (i.e. $-.56$) and in the Table for negative z-score of $-.56$ we get an area of $.2877$. (Note: Due to being symmetric the area in the right tail is also going to be the same for the same positive z score. However, we can also find the area by looking into the Positive Z score Table, look for area of z-score $(.56)$, this is $.7123$ and then, since Z Scores Table gives Cumulative Area from the left, to find the area in the right tail we subtract this value from 1 i.e. area in right tail = $1 - .7123 = .2877$.

As H_1 has \neq hence it is a **Two Tail Test**



6. **Make a Decision:**

P-value (Probability) is the area = $.2877 + .2877 = .5754$

We are going to compare probability to learn if the evidence is strong enough or not strong enough.

P-value (i.e. Probability i.e. Area) $\leq \alpha$ Reject H_0

P-value (i.e. Probability i.e. Area) $> \alpha$ Fail to Reject H_0

P-value ($= .5754$) $> \alpha$ ($= .05$) **Fail To Reject H_0** (Consequently we have proved nothing)

7. We fail to reject the Claim. Does it mean the opposite is True? No, it means nothing, it may be or it may not be, but failed to reject the Null.

Traditional Method

Steps 1 to 4 are identical to P-value Method. In step 4 of P-value method we got **Test Statistic $z = .56$**

Step 5. Draw the Graph

To draw the graph we have to find the Critical Values for given Significance Level. **Since H_1 has \neq** we have to do Two Tail Test. Accordingly, we divide the **Significance Level of .05** into Two Tails, i.e. this gets divided as areas in two tails and becomes .025 in each tail.

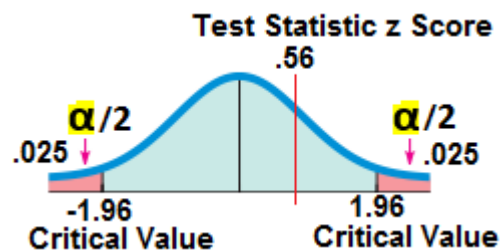
Left Tail: From the **Table of Negative Z-score for Area .025 is -1.96**

Table A-2 Standard Normal (z) Distribution: Cumulative Area from the LEFT

z	.00	.01	.02	.03	.04	.05	.06	.07
-3.50 and lower	.0001							
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244

Right Tail: From the **Table of Positive Z-scores for Area .025** Note: As Table give area from Left hence we find it for $1 - 0.025 = 0.9750$ and Z score for this area is 1.96

z	.00	.01	.02	.03	.04	.05	.06
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750



Step 6. Make a Decision: Compare Test Statistic to Critical Value

- If Test Statistic fall in Reject Region: Reject H_0
- If Test Statistic falls in Fail to Reject Region: Fail to Reject H_0

Test Statistic is IN Fail To Reject Region.

Hence, **FAIL TO REJECT** H_0

8.4 Hypothesis Testing for Population Mean (σ Known)

For testing Claims involving Population Mean, Standard Deviation is known.

Requirements

1. Random Sample
2. σ Known
3. n (Sample Size) > 30 OR Population is Normally Distributed

Test Statistic

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

Example

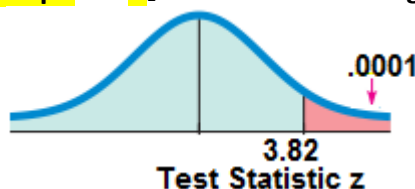
You sample 465 M&Ms, the Sample had a Mean of 0.8635. The Population Standard Deviation is 0.0565. Test the claim that the Mean weight is greater than .8535. Significance Level of .01.

1. CLAIM: $\mu > .8535$ H_1
OPP: $\mu \leq .8535$ H_0
2. $H_0: \mu = .8535$
 $H_1: \mu > .8535$
3. Significance Level $\alpha = .01$
4. **Test Statistic Z** (As per formula given above for Test Statistic)

$$z = (.8635 - .8535) / (0.0565 / \sqrt{465}) = 3.82$$

P-value Method

5. **Draw the Graph:** As H_1 has $>$ hence it is a Right Tail Test.



As per Table, for values of z above 3.49, use 0.9999 for the area on left, hence **area in Right Tail** = $1 - .9999 = .0001$

6. P-value (Probability is the area) = .0001

We are going to compare probability to learn if the evidence is strong enough or not strong enough.

P-value (i.e. Probability i.e. Area) $\leq \alpha$ Reject H_0

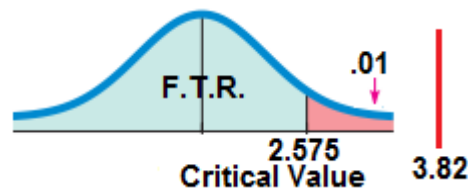
P-value (i.e. Probability i.e. Area) $> \alpha$ Fail to Reject H_0

P-value (= .0001) $\leq \alpha$ (= .01) **Reject H_0** (Consequently we have proved nothing)

Traditional Method

5. **Draw the Graph:** As H_1 has $>$ hence it is a Right Tail Test.

To draw the graph we have to find the Critical Value for given Significance Level. Since H_1 has $>$ we have to do Right Tail Test. Accordingly, the **Significance Level of .01** is in the Right Tail.



Significance Level = Area in Right Tail = **.01**

Area on the Left = $1 - .01 = 0.99$

As per Table, for Confidence Level 0.99 the **Critical Value = 2.575**

6. **Make a Decision:** Compare Test Statistic to Critical Value

- If Test Statistic fall in Reject Region: Reject **H_0**
- If Test Statistic falls in Fail to Reject Region: Fail to Reject **H_0**

Test Statistic 3.82 is IN Reject Region.

Hence, **REJECT H_0**

7. There is enough to Support the Claim ... that

8.5 Hypothesis Testing for Population Mean (σ Unknown)

For testing Claims involving Population Mean, Standard Deviation is not known.

Requirements

1. Random Sample
2. σ is not known (s is known)
3. n (Sample Size) > 30 OR Population is Normally Distributed

Test Statistic

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Traditional Method Only

Example

You sample 39 soda cans. The Mean volume of soda was 12.11 oz. with a Sample Standard Deviation of .27 oz. Test the claim that the Mean volume of soda is Greater Than 12 oz. with a .01 Significance Level.

1. **CLAIM:** $\mu > 12$ H_1
OPP: $\mu \leq 12$ H_0
2. $H_0: \mu = 12$
 $H_1: \mu > 12$
3. Significance Level $\alpha = .01$
4. Calculate Test Statistic – t-score

$$\bar{x} = 12.11 \quad s = .27 \quad \mu = 12 \quad n = 39$$

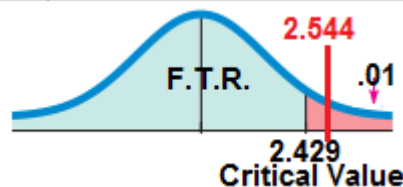
$$T = (12.11 - 12) / (.27/\sqrt{39}) = 2.544$$

5. Draw the Graph: H_1 has $>$ hence it is a right tail test.

Look into t-Table for Critical Value related to the Significance Level

Table A-3 t Distribution: Critical t Values

	0.005	0.01	Area in One Tail 0.025
Degrees of Freedom	0.01	0.02	Area in Two Tails 0.05
1	63.657	31.821	12.706
2	9.925	6.965	4.303
36	2.719	2.434	2.028
37	2.715	2.431	2.026
38	2.712	2.429	2.024
39	2.708	2.426	2.023
40	2.704	2.423	2.021



6. t-score 2.544 falls in reject region, hence, **REJECT H_0**
7. There is enough evidence to support the Claim that Mean level of soda is > 12 oz.

8.6 Hypothesis Testing for Standard Deviation or Variance

Testing Claim about Variance and Standard Deviation

To test a claim about the value of the variance or the standard deviation of a population, then the test statistic will follow a chi-square distribution with $n-1$ degrees of freedom, and is given by the following formula.

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

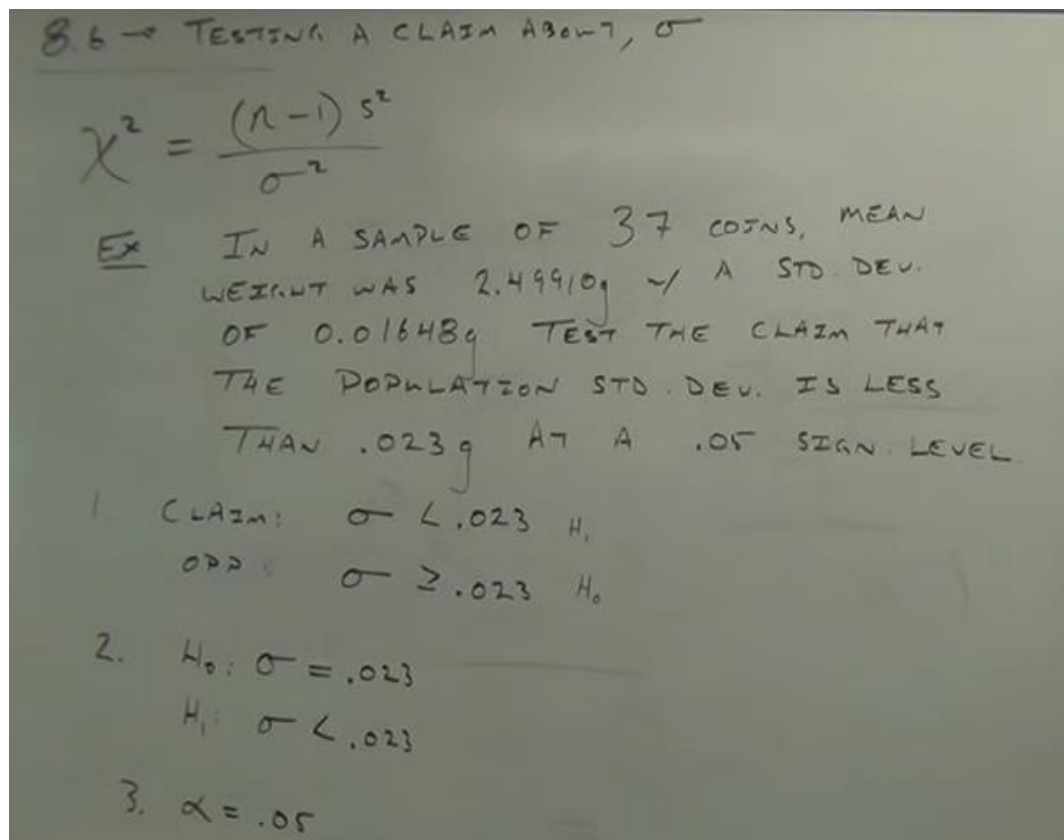
Ex: The television habits of **30 children** were observed. The **sample mean** was found to be **48.2** hours per week, with a **standard deviation of 12.4** hours per week. Test the claim that the **standard deviation was at least 16** hours per week.

The hypotheses are:

$$H_0: \sigma = 16$$

$$H_a: \sigma < 16$$

- We shall choose $\alpha = 0.05$
- The test statistic $\chi^2 = (n-1)s^2 / \sigma_0^2 = (30-1)12.4^2 / 16^2 = 17.418$
- The p-value is $p = \chi^2 \text{cdf}(0, 17.418, 29) = 0.0447$
- Since $p < \alpha$, we reject H_0
- The variation in television watching was less than 16 hours per week.



4. $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ $n = 37$
 $s = .01648$
 $\sigma = .023$

$$\chi^2 = \frac{(37-1)(.01648)^2}{(.023)^2}$$

D.F. = 36

$$\chi^2 = 18.48$$

5.

6. REJECT H_0

7. THERE IS ...

Testing Difference between Two Variances or Two Standard Deviations

Ex: Samples from two makers of ball bearings are collected, and their diameters (in inches) are measured, with the following results:

Acme: $n_1=80$, $s_1=0.0395$

Bigelow: $n_2=120$, $s_2=0.0428$

Assuming that the diameters of the bearings from both companies are normally distributed, test the claim that there is no difference in the variation of the diameters between the two companies.

Soln:

Since sample variances are related to chi-square distributions, and the ratio of chi-square distributions is an F-distribution, we can use the F-distribution to test against a null hypothesis of equal variances. Note that this approach does not allow us to test for a particular magnitude of difference between variances or standard deviations.

Given sample sizes of n_1 and n_2 , the test statistic will have n_1-1 and n_2-1 degrees of freedom, and is given by the following formula.

$$F = \frac{s_1^2}{s_2^2}$$

If the larger variance (or standard deviation) is present in the first sample, then the test is right-tailed. Otherwise, the test is left-tailed. Most tables of the F-distribution assume right-tailed tests, but that requirement may not be necessary when using technology.

The hypotheses are:

$$H_0: \sigma_1 = \sigma_2$$

$$H_a: \sigma_1 \neq \sigma_2$$

- We shall choose $\alpha = 0.05$
- The test statistic is $F = s_1^2/s_2^2 = 0.0395/0.0428 = 0.8517$
- Since the first sample had the smaller standard deviation, this is a left-tailed test. The p-value is $p = \text{Fcdf}(0, 0.8517, 79, 119) = 0.2232$
- Since $p > \alpha$, we fail to reject H_0

There is insufficient evidence to conclude that the diameters of the ball bearings in the two companies have different standard deviations.

What is a Chi Square Test?

There are **two types of chi-square tests**. Both use the chi-square statistic and distribution for different purposes:

- A **chi-square goodness of fit test** determines if a sample data matches a population. For more details on this type, see: [Goodness of Fit Test](#).
- A **chi-square test for independence** compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of [categorical variables](#) differ from each another.
 - A **very small chi square test statistic** means that your observed data fits your expected data extremely well. In other words, there is a relationship.
 - A **very large chi square test statistic** means that the data does not fit very well. In other words, there isn't a relationship.

Test Statistic for Goodness-of-Fit Tests

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

O represents the *observed frequency* of an outcome, found from the sample data.

E represents the *expected frequency* of an outcome, found by assuming that the distribution is as claimed.

"**Chi-squared test**", also written as **χ^2 test**, could be used as the description of any statistical hypothesis test where the sampling distribution of the test statistic is, under some circumstances, approximately, or is simply hoped to be approximately, a **chi-squared distribution**, when the null hypothesis is true.

Chi-squared Distribution

The **chi-square distribution** (also **chi-squared** or **χ^2 -distribution**) with k degrees of freedom is the **distribution of a sum of the squares of k independent standard normal random variables**. It is used in the common chi-square tests for goodness of fit of an observed distribution to a theoretical one,

If Z_1, \dots, Z_k are [independent](#), [standard normal](#) random variables, then the sum of their squares,

$$Q = \sum_{i=1}^k Z_i^2$$

is distributed according to the chi-square distribution with k degrees of freedom. This is usually denoted as $Q \sim \chi^2(k)$ or $Q \sim \chi_k^2$.