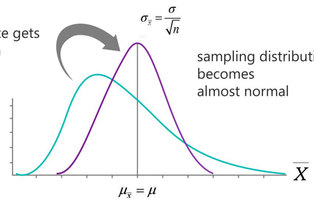
Definitions :

- **Data** - Values collected by direct or indirect observation
- **Population** – Complete set of all observations in existence
- **Sample** – Slice of population meant to represent, as accurately as possible, that population
- **Inferential Statistics** - procedures employed to arrive at broader generalizations or inferences from sample data to populations
- **Measure** – Measurement of population/sample, an example would be some "score" (a.k.a. an observation)

Central Limit Theorem

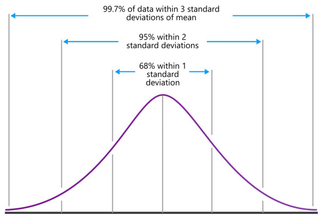
As sample size gets large enough ($n \geq 30$)...



sampling distributinon becomes almost normal

A Normal Distribution

- A.K.A. "Bell Curve"
- Way to visualize how volume of a population is distributed based on some measurement
- Largest volume is packed around middle
- Volume curves down towards zero to left and right
- Symmetrical around middle
- Interesting Fact: The Mean, Median, and Mode are all the same and at the exact center



Is My Data Special?

Null Hypothesis in Layman's Terms:

There is nothing different, or special, about this data

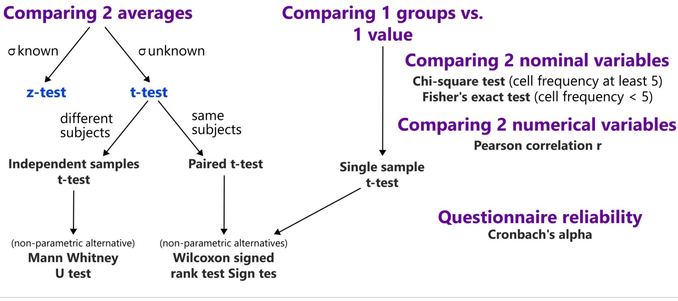
- **Best used when you need to know if your data is different or somehow special**
- Always start out assuming Null Hypothesis is **True**
- Goal is to either "reject" or fail to reject Null Hypothesis
- If **Fail To Reject** Null Hypothesis then there is nothing really different about the data
- If we **Reject** Null Hypothesis then we are confident that what we see is different or special
- On the curve above, we can only say that an observation is different/special if it falls in either of shaded regions (called "tails)
- The tails are 2 Standard Deviations away from (either above or below) the Mean
- Assumes dealing with a normal distribution!

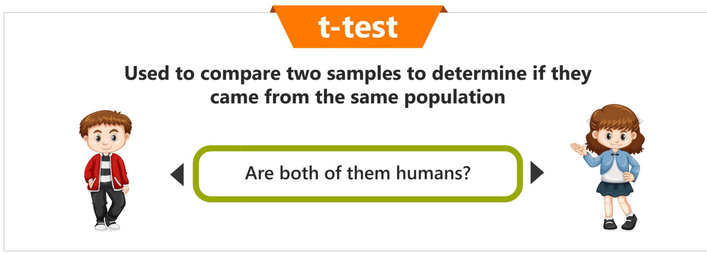
★ **Type I Error (false positive)** – In hypothesis testing when you incorrectly reject the Null Hypothesis.

★ **Type II Error (false negative)** – In hypothesis testing, when you incorrectly fail to reject the Null Hypothesis.

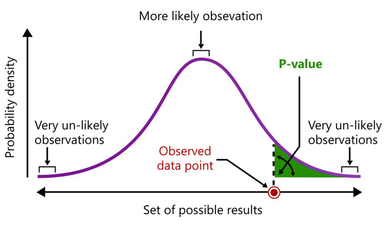
- **Confusing Confidence Intervals...**
  ...with probability. 95% confidence just means that 95% of the time the true (population) value will be within the limits.

- **Multiple Inference...Faking it 'till you're making it!**
  Running a hypothesis test over and over, the same way on the same data, until you get a "significant" result greatly increases chances you will get a false positive (Type I Error) result because... there is always the chance of getting a randomly significant result.





P - value



A **P-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

CHI-SQUARED

For Feature Selection

To use X^2 for feature selection, we calculate X^2 between each feature and the target, and select the desired number of features with the best X^2 scores.

The intuition is that if a feature is independent to the target it is uninformative for classifying observations.

# of observations in class i

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

# of expected observations in class i if there was no relationship between the feature and target.