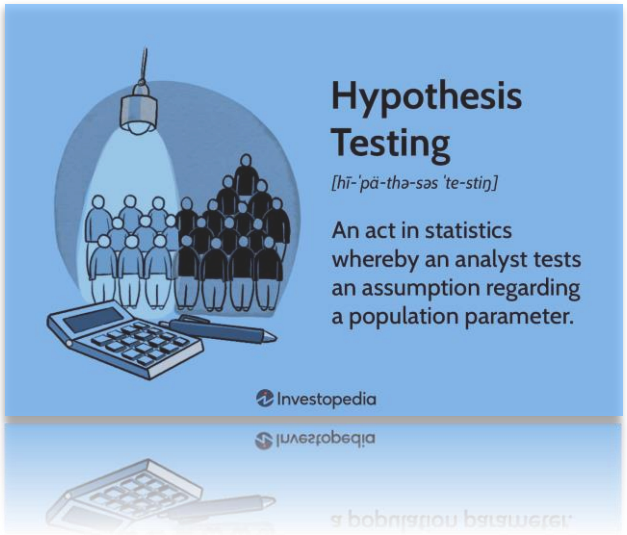
DATA, MSML, BIOI 602
Principles of Data Science



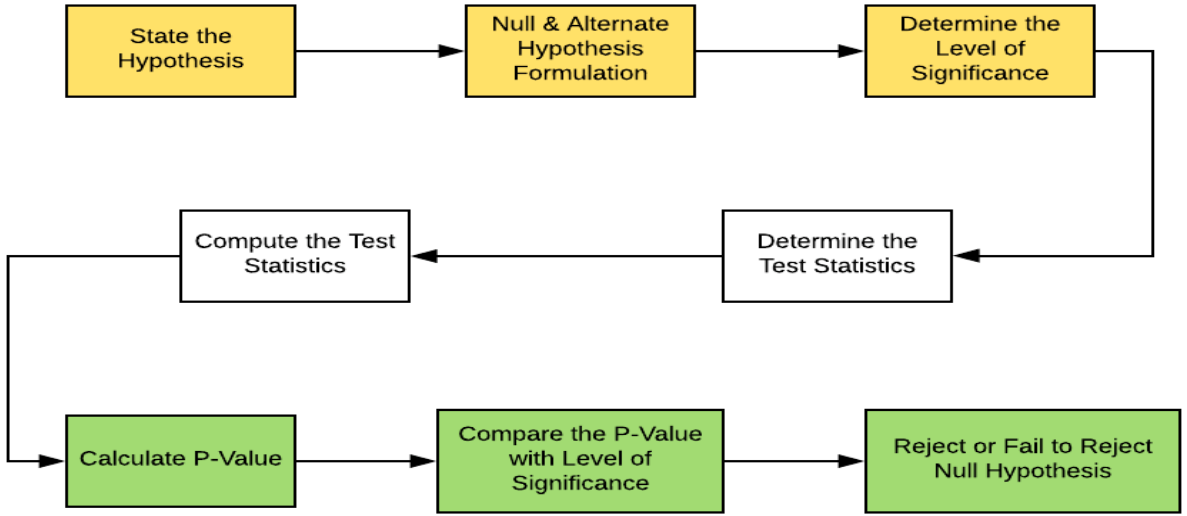
[Slides from Fardina Fathmiul Alam]

“Data alone is not interesting.
It is the interpretation of the data that we are really interested in.”

Hypothesis Testing (Making Informed Decisions with Data)

Hypothesis Testing

Hypothesis testing is a statistical method used to make informed decisions or draw conclusions about a *population* based on a *sample of data*.

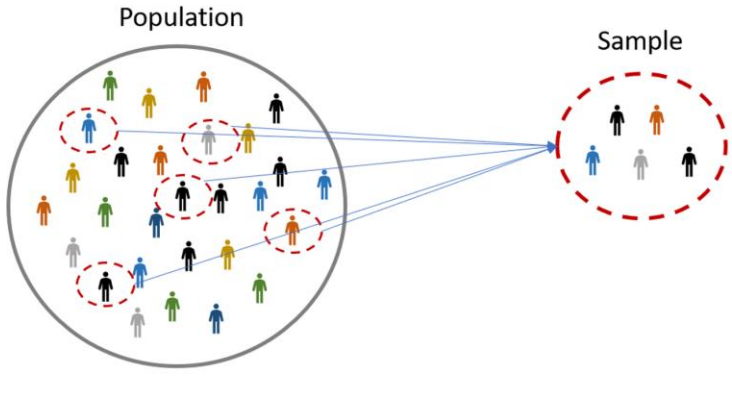


Hypothesis Testing Workflow

Hypothesis Testing

Hypothesis testing is a statistical method used to make informed decisions or draw conclusions about a *population* based on a *sample of data*.

RECAP:
Population: The entire group that you are interested in studying.
Sample: A subset of the population used for analysis.

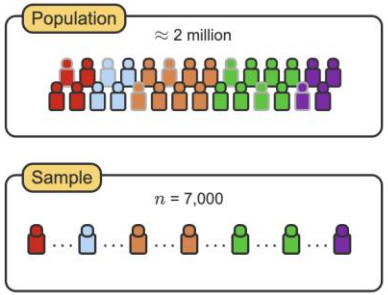


Main Steps of Hypothesis Testing

1. State the Null Hypothesis
2. State the Alternative Hypothesis
3. Pick a Level of Significance α
4. Choose a Test
5. Collect Data
6. Calculate a test statistic
7. Calculate P-Value and compare with α
8. Draw a Conclusion

Collect Data

Obtain a representative sample from the population.



Remember the importance of recognizing whether data is collected through an experimental design or observational study.

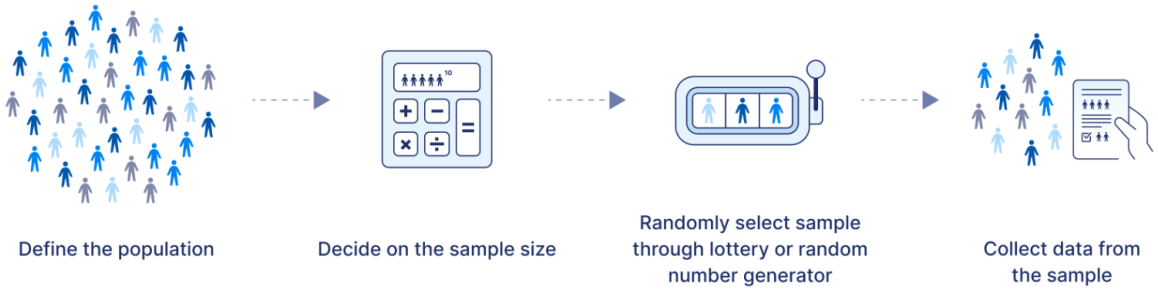
Sampling Methods

A sampling method is a process by which individual items/event (observational units) are selected from the population to be included in the sample. Common sampling methods:

- **Random sampling:** Choosing names randomly from a list for a survey
- **Stratified sampling:** Surveying transportation preferences involves dividing the population into three income strata—low, middle, and high-income households—and then randomly selecting households from each stratum in a city.
- **Cluster sampling:** Estimating the average income of households in a city by dividing the city into clusters based on neighborhoods or districts, then randomly selecting specific neighborhoods as clusters and surveying all households within the selected neighborhoods.
- **Systematic sampling:** Selecting every 10th person from a list of customers
- **Convenience sampling:** Surveying people in a shopping mall (ease of access, availability), Recruiting volunteers from a specific organization or community etc.

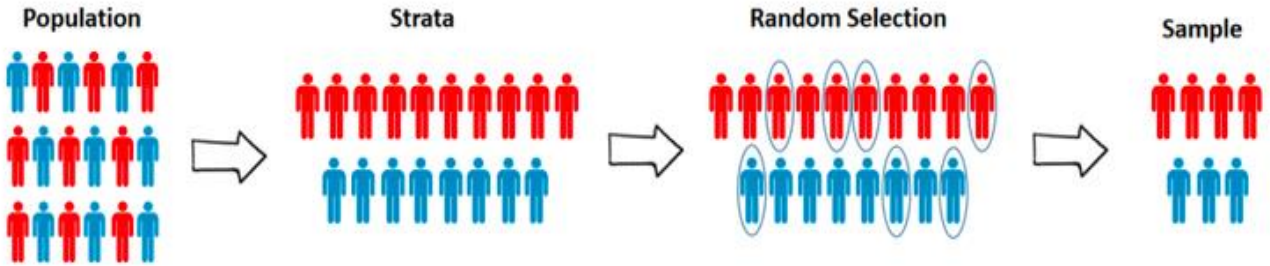
Different Sampling Methods

- **Random sampling:** Observational units are chosen completely at random from the entire population. Each unit has an equal chance of being selected, ensuring that the sample represents the population well.



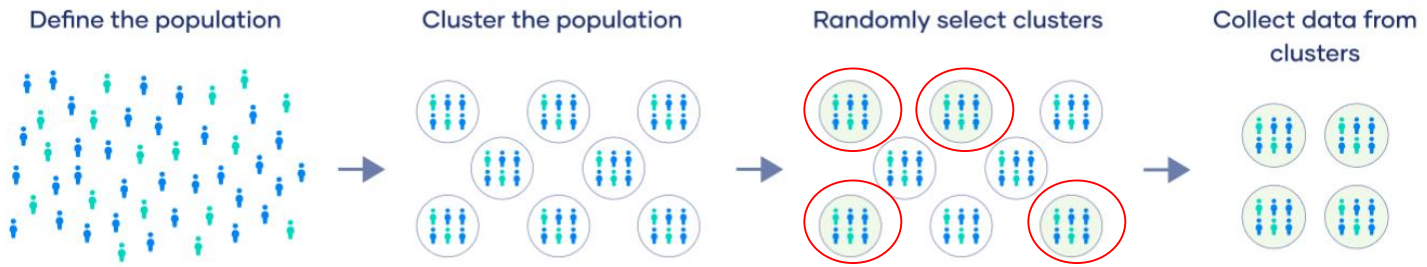
Different Sampling Methods

- **Stratified sampling**: The population is divided into distinct groups or strata based on specific characteristics (e.g., gender, age, sex, race, education level, or income) that are relevant to the study. Then, random samples are taken from each stratum. This ensures representation from all important subgroups in the population, regardless of their individual similarities or differences.



Different Sampling Methods

- **Cluster sampling:** The population is divided into clusters based on a certain characteristic (such as geographic location). Then a random selection of clusters is made, and all units (individuals) within the selected clusters are included in the sample. This method is particularly useful when obtaining a complete list of the population is difficult, as clusters provide a more manageable sampling frame.

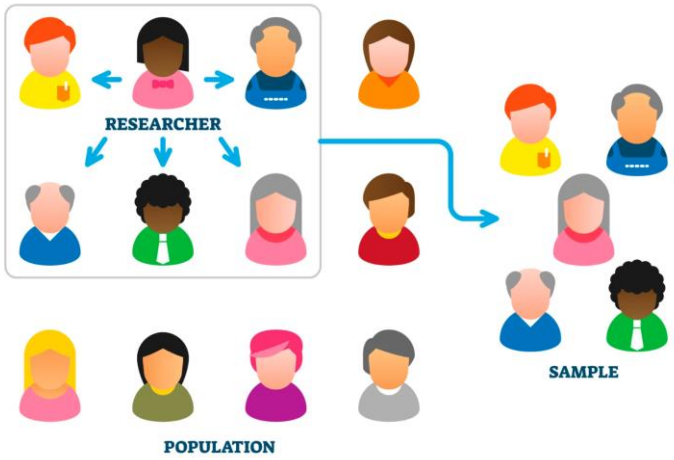


Different Sampling Methods

- **Systematic sampling:** Researchers select a random starting point in the population and then choose every kth unit thereafter until the desired sample size is reached. This method is efficient and easier to conduct compared to simple random sampling, especially when a complete list of the population is available.

Different Sampling Methods

- **Convenience sampling**: This method involves selecting units that are readily available and accessible to the researcher. It is a non-probabilistic sampling technique where researchers select individuals who are readily available and accessible to participate in a study.
    - This method involves choosing participants based on their convenience or accessibility, rather than using random selection procedures.
    - While convenient, this method may introduce bias because it may not represent the entire population accurately. It's often used in exploratory research or when other sampling methods are impractical.



STATE/ FORMULATE THE HYPOTHESIS

Hypothesis

Hypothesis → A **premise** or **claim** that we want to test/investigate.

Hypothesis

Hypothesis → A **premise** or **claim** that we want to test/investigate.

Why do hypothesis testing? ***Sample mean may be different from the population mean***

It is a statement about one or more populations.

It is usually concerned with the parameters (such as the mean (average), variance (spread), or proportion (percentage)) of the population (help in understanding the overall characteristics of the population).

The hospital administrator may want to **test the hypothesis** that the *average length of stay* of patients admitted to the hospital is **5 days**

Back to Hypothesis

There are two hypotheses involved in hypothesis testing

1. Null Hypothesis: This is the hypothesis to be tested
    a. Def: A statement of no effect or no difference (currently established).
    b. Represents currently accepted value for a parameter
    c. Suggests that the observed data does not deviate from the population
    d. Denoted as H_o

2. Alternative Hypothesis (also known as Research Hypothesis): Statement you will adopt if the evidence (data) is strong enough to reject the null hypothesis.

    a. Contradicts the null hypothesis.
    b. Suggests that the observed data is DIFFERENT than the population
    c. Represents what you think might be true
    d. Denoted as H_a

Example:

Let's say, **It is believed** that a candy making machine makes chocolate bars that are on average **5 gram in weight**. A worker claims that the machine after maintenance no longer makes 5 gram bar. Write down **H_0** and **H_a.**

Null Hypothesis H_0 : $\mu = 5$ gram.

Alternative Hypothesis: H_a : $\mu \neq 5$ gram.

H_0 and H_a. are mathematical opposite (mutually exclusive)

Possible Outcomes of the Test

Let's say, It is believed that a candy making machine makes chocolate bars that are on average 5 gram in weight. A worker claims that the machine after maintenance no longer makes 5 gram bar. Write down **H0** and **Ha.**

Ho: μ = 5 gram. ; Ha : μ ≠ 5 gram.

This test can lead to TWO possible outcome of the test

Possible Outcomes of this test:

1. **Reject** the null hypothesis
2. **Fail to Reject** the null hypothesis

Criteria: Based on evidence from sample data.

Possible Outcomes of the Test

Let's say, It is believed that a candy making machine makes chocolate bars that are 5 gram in weight. A worker claims that the machine after maintenance no longer makes 5 gram bar. Write down **H0** and **Ha.**

Ho: μ = 5 gram. ; Ha : μ ≠ 5 gram.

This test can lead to TWO possible outcome of the test

Possible Outcomes of this test:

1. **Reject** the null hypothesis
2. **Fail to Reject** the null hypothesis

Q: How do you do the testing? How do you actually decide to reject the null or not?

Possible Outcomes of the Test

Let's say, It is believed that a candy making machine makes chocolate bars that are 5 gram in weight. A worker claims that the machine after maintenance no longer makes 5 gram bar. Write down **H0** and **Ha.**

Ho: μ = 5 gram. ; Ha : μ ≠ 5 gram.

This test can lead to TWO possible outcome of the test

Possible Outcomes of this test:

1. **Reject** the null hypothesis
2. **Fail to Reject** the null hypothesis

Do Test Statistics → this is computed from sample data (during the execution of a statistical test) used to decide possible outcome

PRACTICE

WRITING NULL AND ALTERNATIVE HYPOTHESIS

Doctors believe that the average teen sleeps on average of no longer than **10 hours per day**. A researcher believes that teens on average sleep longer.

Write down **H0** and **Ha.**

Choose a Statistical Test and calculate Test Statistics

Choose a Statistical Test:

Choose the statistical test that is suitable for your data and research question.

refers to the overall method or procedure used to decide whether to reject the null hypothesis

The choice of the appropriate statistical test depends on various factors, such as

- ❏ The research question
- ❏ The specific hypothesis being tested
- ❏ The type of data (e.g., categorical or continuous), and
- ❏ The number of groups being compared etc.

Calculate a Test Statistic:

Definition: A test statistic is a numerical value calculated from sample data during the hypothesis testing process.

- It quantifies the evidence from the data against the null hypothesis and helps determine **whether to reject or fail to reject the null hypothesis**.
- It measures how close the sample results align with the null hypothesis or how much the sample result deviate from what is expected if the null hypothesis is true.

The test statistic basically tells you if your data agrees or disagrees with your starting idea (the null hypothesis). It shows how strong your evidence is either for or against your idea.

Different hypothesis tests **use different test statistics** based on the probability model assumed in the null hypothesis. Common tests and their test statistics include:

Hypothesis Test	Test Statistic
Z-test	Z-statistic
T-Test	t-statistic
ANOVA	F-statistic
Chi-Square Tests	Chi-square-statistic

Q: How do you do the testing?

Let's say, It is believed that a candy making machine makes chocolate bars that are on average 5 gram in weight. A worker claims that the machine after maintenance no longer makes 5 gram bar. Write down **H0** and **Ha.**

Ho: μ = 5 gram. ; Ha : μ ≠ 5 gram.

Possible Outcomes of this test:

1. Reject the null hypothesis
2. Fail to Reject the null hypothesis

Do Test Statistics → this is calculated from sample data used to decide possible outcome

Example: We sample 50 chocolate bars and get average value of mass of the bar.

- Then we calculate test statistics (depends on what type of problem you have)
- It help to determine that the data you have statistically significant enough to reject this null hypothesis or not.

What step helps us **decide whether to reject** the null hypothesis or not, and **how confident** can we be in our decision?

Determine the Significance Level (Alpha):

Significance Level

Significance Level: (usually denoted as α) represents is a measure of the strength of the evidence used as a threshold **for deciding whether to reject or fail to reject the null hypothesis in a hypothesis test.**

It is the probability of rejecting the null hypothesis when it's actually true.

- Common values for **α are 0.05 or 0.01**.

The significance level helps you **determine how confident you need to be in your results before** you make a decision.

Significance Level: Example

Let's say you're testing a new treatment for headaches.

Your null hypothesis (H0) : the treatment has no effect, and

Your alternative hypothesis (H1) : the treatment reduces headaches.

- **You set your significance level (α) at 0.05**, meaning **you're willing to accept a 5% chance of mistakenly** concluding that the treatment works when it actually doesn't.

- After conducting your study, let's say you calculate a p-value of 0.03.
  - Since the **p-value is less than α (0.03 < 0.05)**, you reject the null hypothesis and
  - indicates that the observed data is statistically significant, providing enough evidence to support the alternative hypothesis.
    - conclude that the treatment likely reduces headaches.

Back to the previous example

Q: How confident you are with your decision?

Ho: μ = 5 gram. ; Ha : μ ≠ 5 gram.

Possible Outcomes of this test: Reject the null hypothesis or Fail to Reject the null hypothesis

Example: We sample 50 chocolate bars and get average value of mass of the bar.

- Then we calculate test statistics (depends on what type of problem you have)
- It help to determine that the data you have statistically significant enough to reject this null hypothesis or not.

Let's say:

Monday → Received average value of mass of the bar 5.12 gram

Wednesday → Received average value of mass of the bar 5.75 gram

Friday → Received average value of mass of the bar 7.82 gram

How much I am confident that I should reject the Ho or not! → we need a CONCRETE WAY to look at the Ho

Understanding Level of Confidence (C) or Significance (Alpha)

How confident are we in our decision?

We use variable **C (Confidence Level):** Represents the degree of certainty in our decision-making process; could be 95%, 99%.

Level of Significance (α): Indicates the likelihood of making a Type I error, which is rejecting the null hypothesis when it's true.

- **α** is a term used in scientific research to **describe when to reject the null hypothesis.** Often times it is .05 or .01.
- This gives us a **quick way of describing whether or not it is likely an effect exists.**

Understanding Level of Confidence (C) or Significance (Alpha)

How confident are we in our decision?

Interpreting Confidence Level:

- **High Confidence Level (e.g., 99%):** If we're 99% confident in our decision and reject the null hypothesis, we're highly certain that rejecting the null is the correct choice.
- **Low Confidence Level (e.g., 50%):** If our confidence level is only 50%, we're less convinced and might hesitate to reject the null hypothesis.

Understanding Level of Confidence (C) or Significance (Alpha)

How confident are we in our decision?

Remember: **Level of significance; alpha=1-C** (always sum up to 1)

If Level of Confidence=99% , C=0.95 , so alpha = 1-C =0.95 = 0.05

- C and α are often used **interchangeably.**
- **They both convey the same concept: How certain we are about the correctness of our decision.**

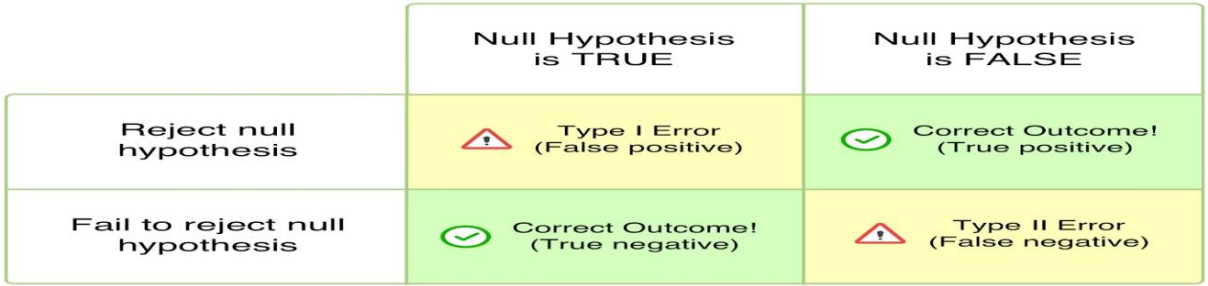
How to decide C or Alpha value? depends on several factors, including the nature of the research, the consequences of making Type I and Type II errors, and the standard practices in the field. A general guide for choosing the value:

- Evaluate the impact of Type I and Type II errors on your study and adjust alpha or C based on which error is more critical.
- Check if your field typically follows specific alpha (e.g., α = 0.05). or C values, as preferences may vary based on disciplinary standards

Type I and Type II Errors

Hypothesis testing involves the risk of making errors.

- Type I Error (False Positive): Occurs when you incorrectly reject a true null hypothesis (α) (**We reject the null hypothesis when the it is true.**) The significance level (α) represents the probability of making a Type I error.
- Type II Error (False Negative): Occurs when you fail to reject a false null hypothesis (β). (**we accept the null hypothesis when it is not true**)



Type I and Type II Errors

In simpler terms:

- **Type I Error (α):** We say there is an effect when there isn't (α is the probability of this happening).
- **Type II Error (β):** We say there isn't an effect when there is.

These errors are important to consider because they affect the validity of the conclusions drawn from hypothesis testing.

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	⚠ Type I Error (False positive)	✓ Correct Outcome! (True positive)
Fail to reject null hypothesis	✓ Correct Outcome! (True negative)	⚠ Type II Error (False negative)

Example: What is Type I and II error?

Let's say you're testing whether a new drug is effective in reducing blood pressure.

Your null hypothesis (H0) : the drug has no effect on blood pressure, and

Your alternative hypothesis (H1) : the drug reduces blood pressure.

Collect data from a sample of patients and conduct a statistical test.

Type I error → concluding that the drug is effective at reducing blood pressure when, in reality, it has no effect. (**a false positive result, indicating the drug works when it doesn't.**)

Type II error → failing to detect that the drug is effective at reducing blood pressure when, in fact, it does reduce blood pressure (**a false negative result, indicating the drug doesn't work when it actually does.)**

PRACTICE

WRITING NULL AND ALTERNATIVE HYPOTHESIS

A company stated that their straw machine makes straws that are 4mm in diameter. But a worker believes that the machine no longer makes straws of this size and samples 100 straws to perform a hypothesis test with 99% confidence

Write down **H0**, **H, N, C, alpha.**

Set Criteria for the Decision

RECAP: The significance level (α) represents the t**hreshold below which you** would reject the null hypothesis (probability value used to define the (unlikely) sample outcomes) i**f the null hypothesis is true.**

- It is the probability of making a Type I error in hypothesis testing.

Interpretation: If you set your significance level at α = 0.05 (5%), it means you are willing to accept a 5% chance of incorrectly rejecting a true null hypothesis (a false positive).

Common Values: Common values for alpha are 0.05 (5%) and 0.01 (1%), although other levels may be chosen depending on the desired balance between Type I and Type II errors.

Next: Set Criteria for the Decision

After determining the significance level, the next step is to calculate the test statistic and then compare it to either the **critical value** or calculate the **p-value** to make a decision about the null hypothesis (whether to reject or fail to reject the null hypothesis).

Determine the critical region (designing a decision rule)

In hypothesis testing, **critical region (or** region of rejection**)** is represented by set of values of the test statistic, for which null hypothesis is rejected.

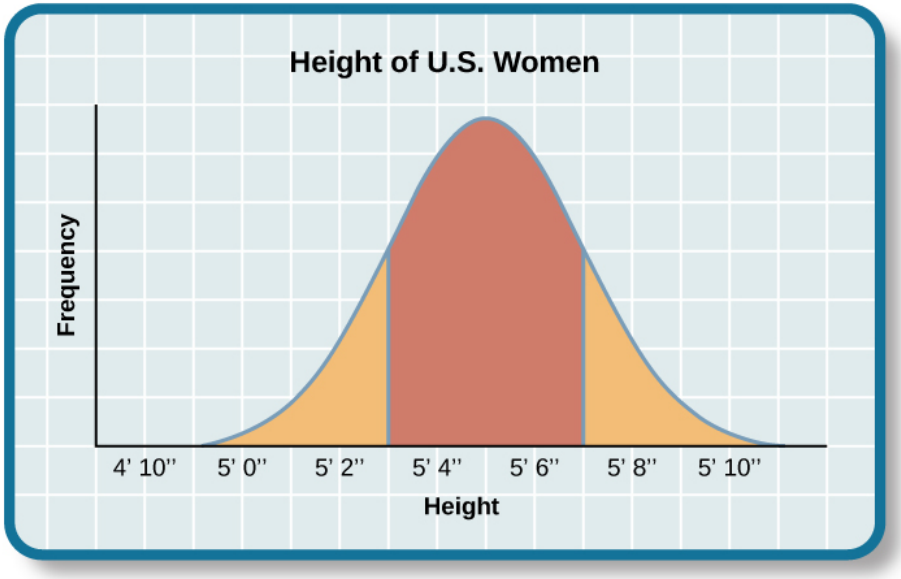
It represents the boundary beyond which the null hypothesis is not supported by the data.

- The critical region takes different boundary values for different levels of significance.

Example: Human Heights

Suppose we're investigating whether aliens are abducting humans based on their height.

Ques: If an alien abduct someone, what is our most *likely* height?

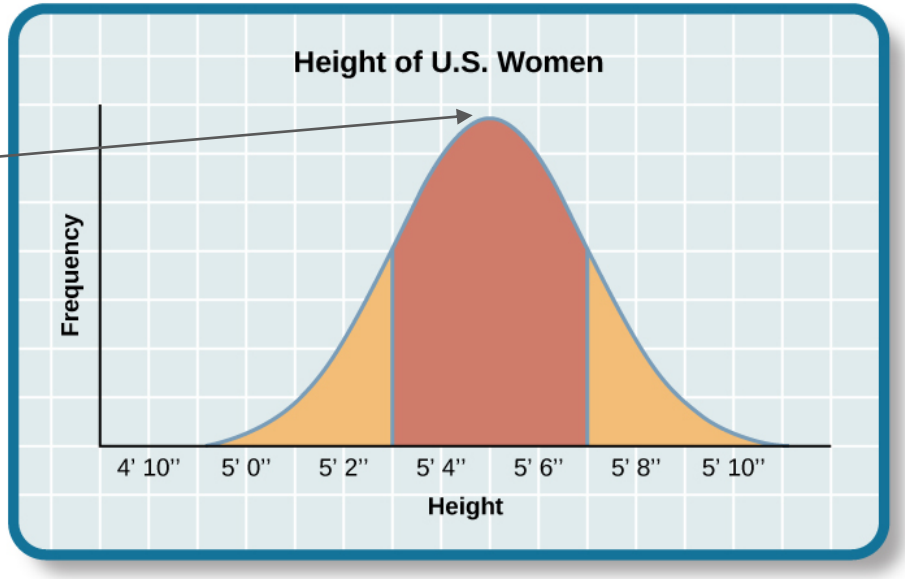


Human Heights

Suppose we're investigating whether aliens are abducting humans based on their height.

If an alien abduct someone, what is our most *likely* height?

If heights are *normally distributed*, the most likely human we got is the mode.



Human Heights

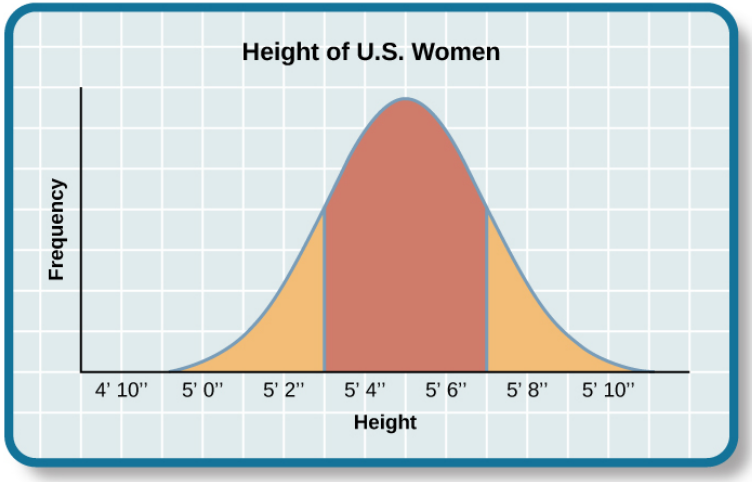
Suppose we're investigating whether aliens are abducting humans based on their height.

Set up:

Null Hypothesis (H0): The height of abducted humans is just like the most common height in the human population.

Question: Based on the setup of the null hypothesis, where would we expect to observe rejection of the null hypothesis in our analysis?

Consider the unusual or extreme regions that represent values of the test statistic that are unlikely to occur (far from the mean or most common height in the human population) if the null hypothesis is true.

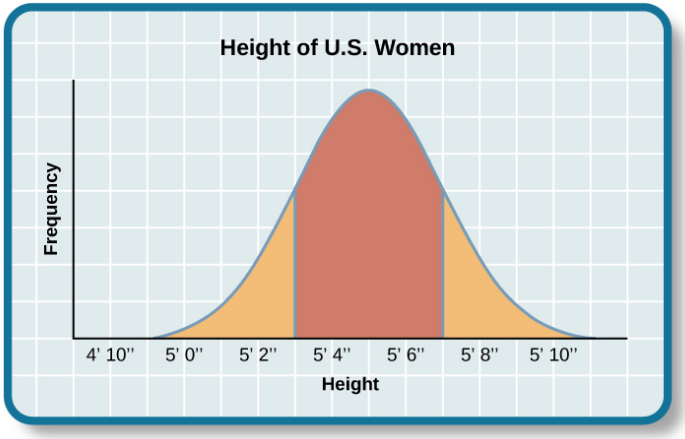


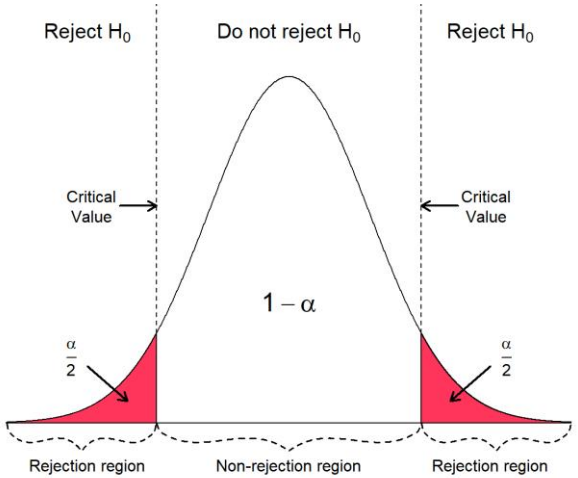
Critical Region

The critical region is the range of values where, if our test statistic falls within it, we'll reject the null hypothesis.

The critical value is the boundary point that separates this critical region from the rest of the possible values of the test statistic.

In simpler terms, it's like setting a line in the sand—if our test statistic crosses that line, we make a decision about the null hypothesis.

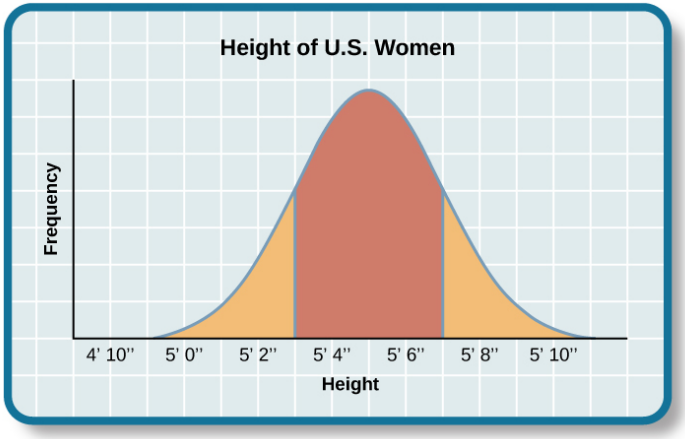


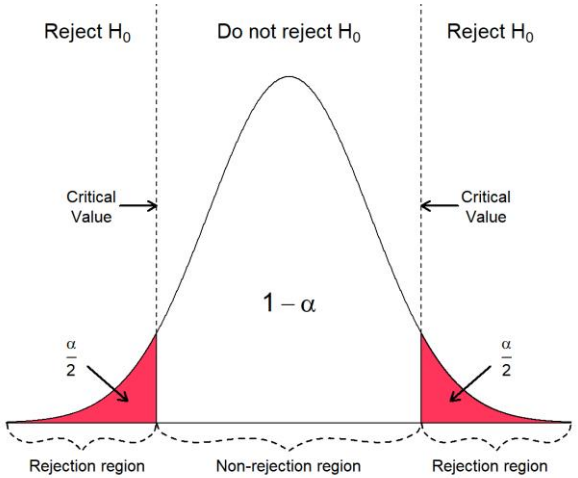


Critical Region and Alpha

The critical value (the boundary point) is determined by the alpha value.

- α (alpha), threshold used to determine the likelihood of making a Type I error. It represents the maximum probability of rejecting the null hypothesis when it is actually true.
- This significance level α is the acceptable level of risk that researchers are willing to take when making a decision about the null hypothesis.





More: Critical Region

A critical value is a specific value or range of values that defines the boundary of the critical region in a hypothesis test.

Extreme sample values that are very unlikely to be obtained if the null hypothesis is true.

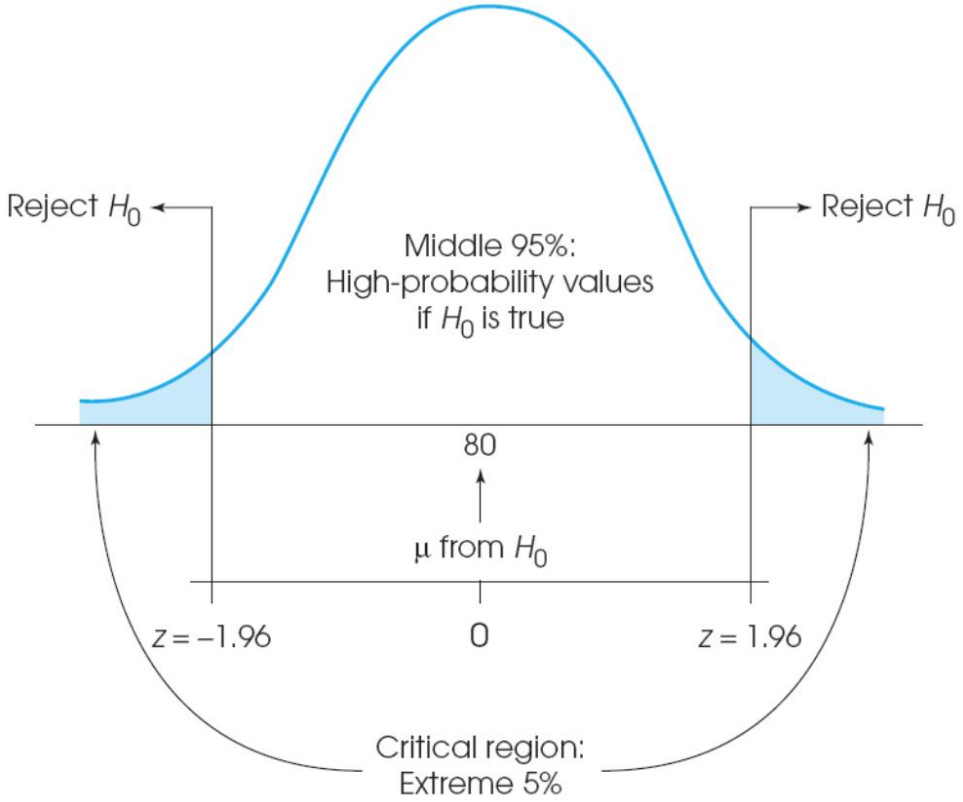
- It is the region of values that corresponds to the rejection of the null hypothesis at some chosen probability level.
- Boundaries determined by alpha level
- If sample data falls within this region (the shaded tails), reject the null hypothesis

SET CRITERIA FOR DECISION

Critical Region Boundaries

In a hypothesis test, you compare the *calculated test statistic* to the *critical value(s)* to make a decision:

- If the test statistic falls in the critical region (beyond the critical value), you reject the null hypothesis.
- If the test statistic falls outside the critical region, you fail to reject the null hypothesis.



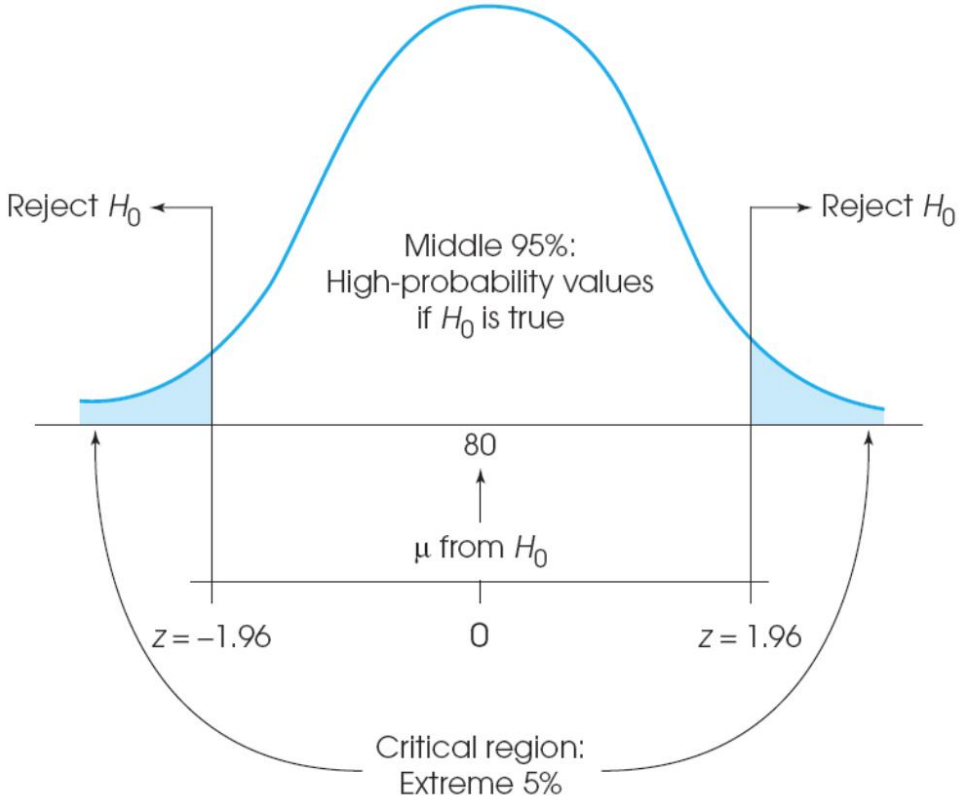
Critical Region Boundaries

Assume normal distribution

Example: if significance level of 5%:

α = .05, boundaries of critical region divide middle 95% from extreme 5%

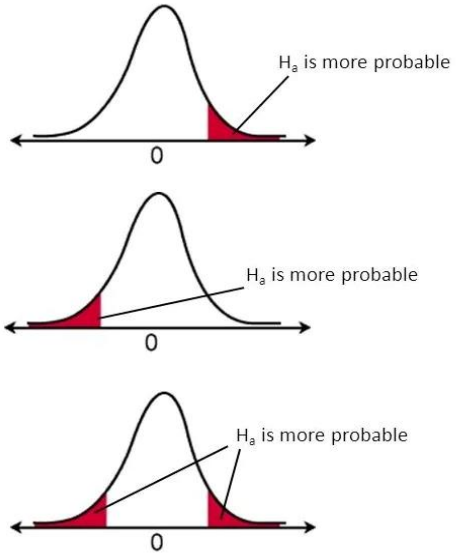
- 2.5% in each tail (2-tailed test which tests for differences in both directions)



One tailed (one sided test), Two tailed (left, right sided test)

In hypothesis testing, two-tailed tests, left-tailed tests, and right-tailed tests are different ways to specify the directionality of the test and the critical region.

Connection to the critical region in how they determine which values of the test statistic lead to the rejection of the null hypothesis.



Right-tail test

$H_a: \mu > \text{value}$

Left-tail test

$H_a: \mu < \text{value}$

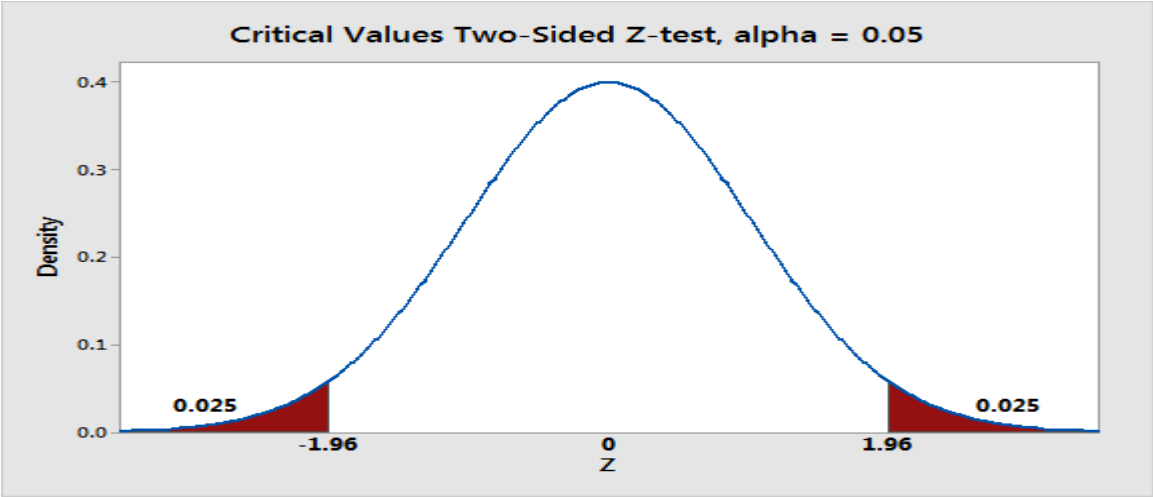
Two-tail test

$H_a: \mu \neq \text{value}$

Two-sided hypothesis tests

In a two-sided test, the null hypothesis is rejected if the test statistic is too small or too large.

- Tests whether the sample mean is significantly different from the population mean, in either direction.
- Thus, the rejection region for such a test consists of two parts: one on the left and one on the right.

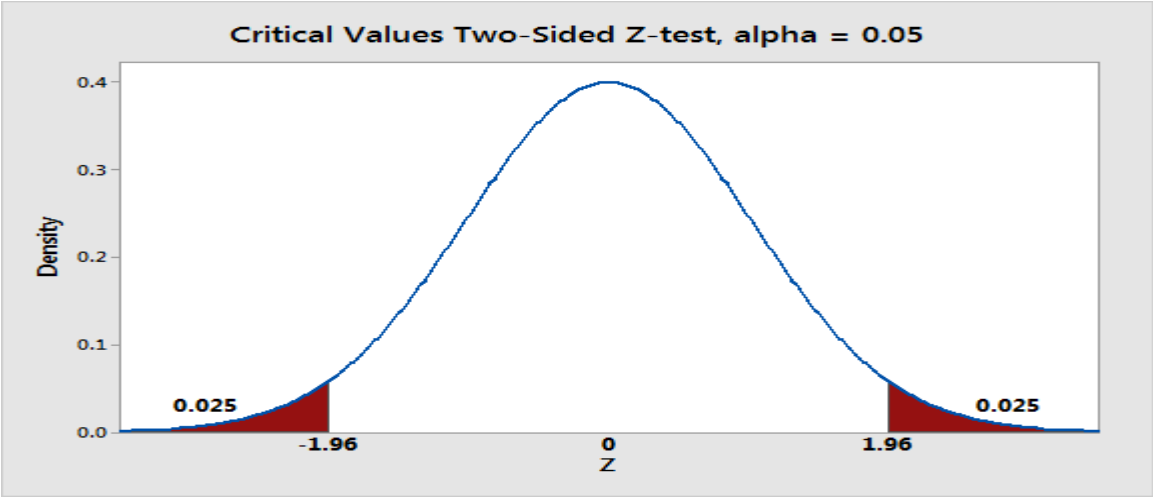


H0: μ = μ0
Ha: μ ≠ μ0

Two-sided hypothesis tests

Two-sided hypothesis tests have two rejection regions.

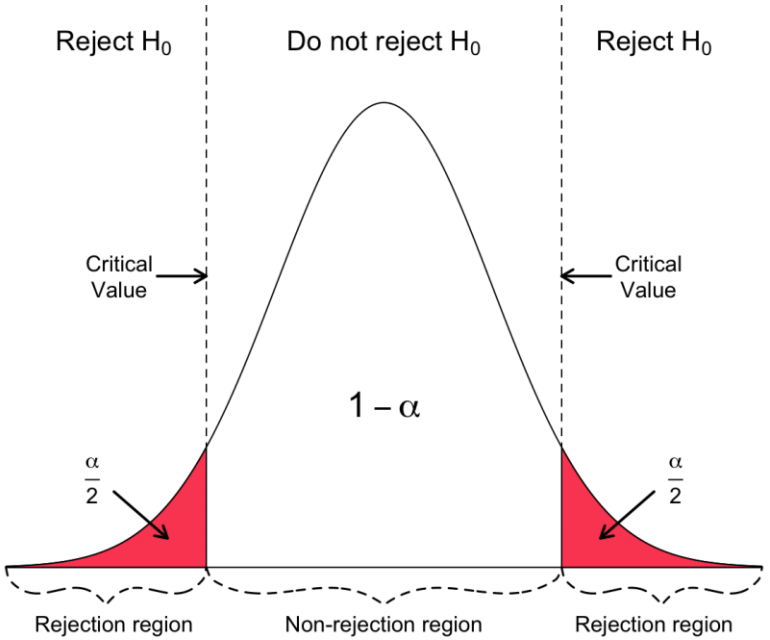
- you'll need two critical values that define them.
- Because there are **two rejection regions**, we must split our significance level in half →Each rejection region has a probability of α / 2, making the total likelihood for both areas equal the significance level.



H0: μ = μ0
Ha: μ ≠ μ0

Two-sided hypothesis tests

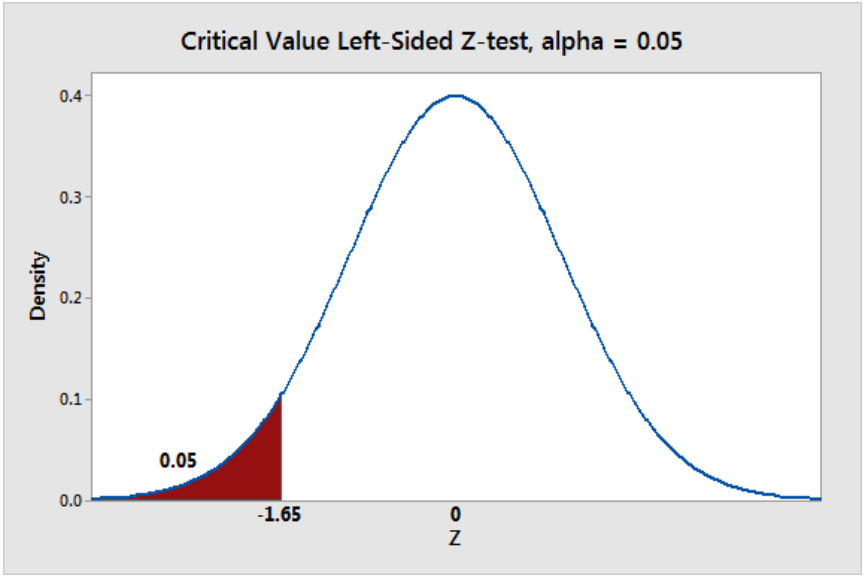
- The critical area of a distribution is two-sided and tests whether a sample is greater than or less than a certain range of values.
- If the sample being tested falls into either of the critical areas, **the alternative hypothesis is accepted instead of the null hypothesis.**



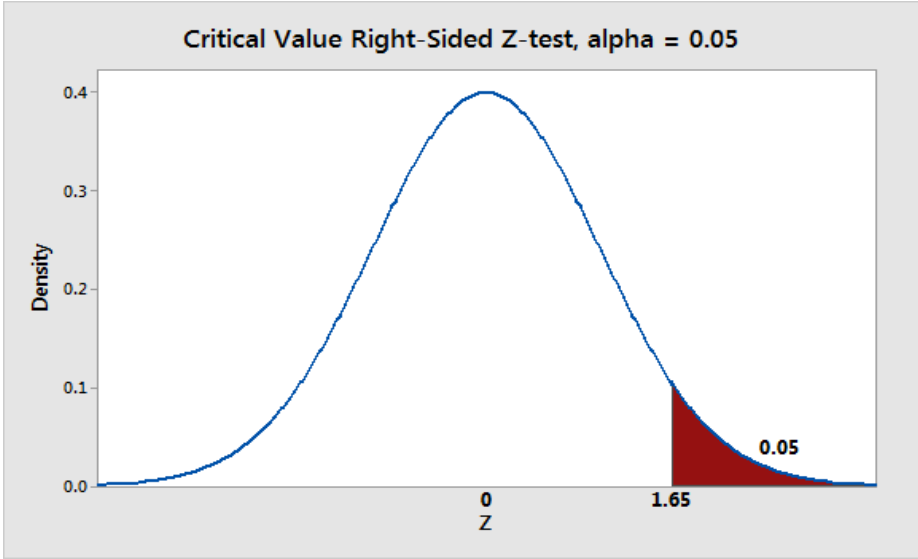
H0: μ = μ0
Ha: μ ≠ μ0

One-Sided Tests

One-tailed tests have one rejection region and, hence, only one critical value. The total α probability goes into that one side.



H0: μ ≥ μ0 ; Ha: μ < μ0



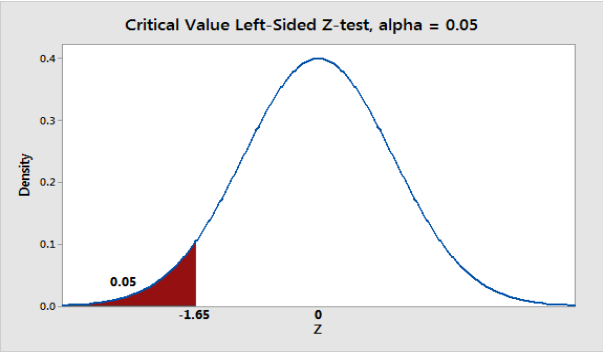
H0: μ ≤ μ0 vs. Ha: μ > μ0

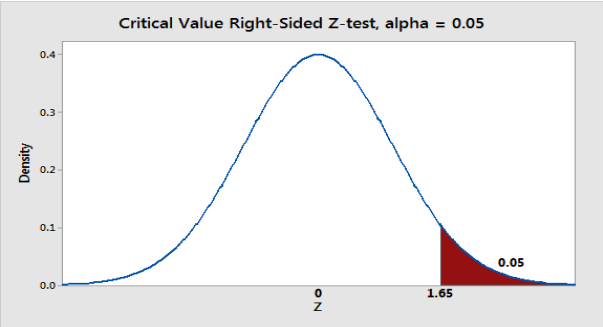
One-Sided Tests

Left-Tailed Test: A Hypothesis Test where the rejection region is located to the extreme left of the distribution. A left-tailed test is conducted when the alternative hypothesis (HA) contains the condition: $\mu < \mu_0$

Right-Tailed Test: The rejection region is located to the extreme right of the distribution. A right-tailed test is conducted when the alternative hypothesis (HA) contains the condition $\mu > \mu_0$

$H_0: \mu \geq \mu_0$; $H_a: \mu < \mu_0$





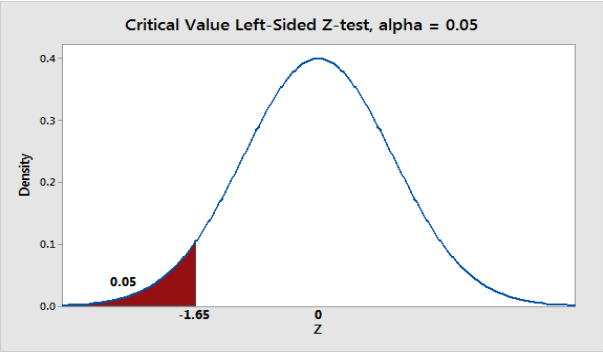
$H_0: \mu \leq \mu_0$ vs. $H_a: \mu > \mu_0$

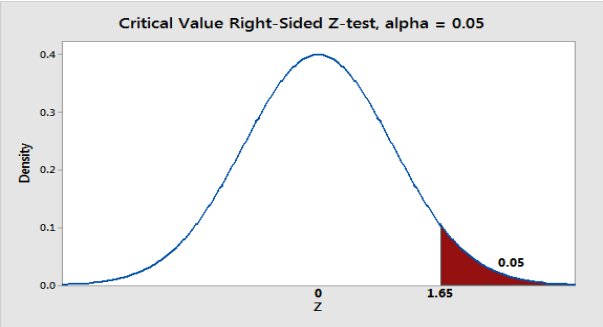
One-Sided Tests

The critical area of a distribution is one-sided so that it is either greater than or less than a certain value, but not both.

If the sample being tested falls into the one-sided critical area, the alternative hypothesis will be accepted instead of the null hypothesis.

H0: μ ≥ μ0 ; Ha: μ < μ0

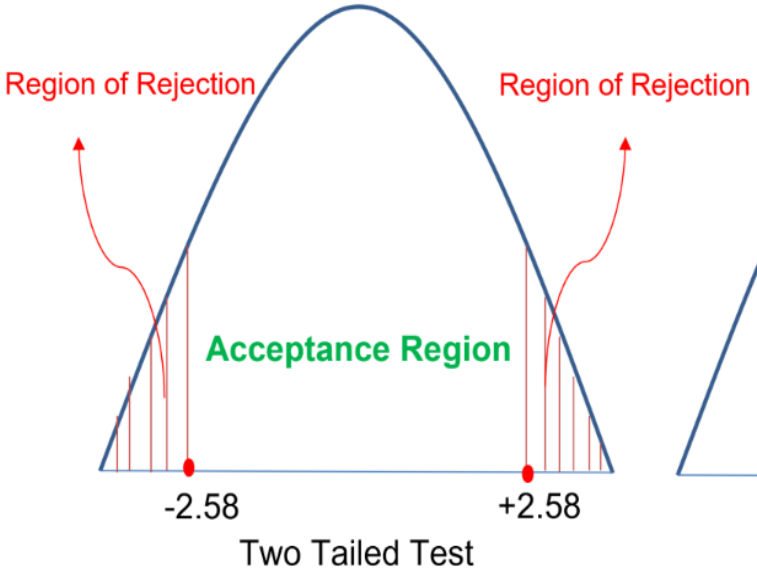


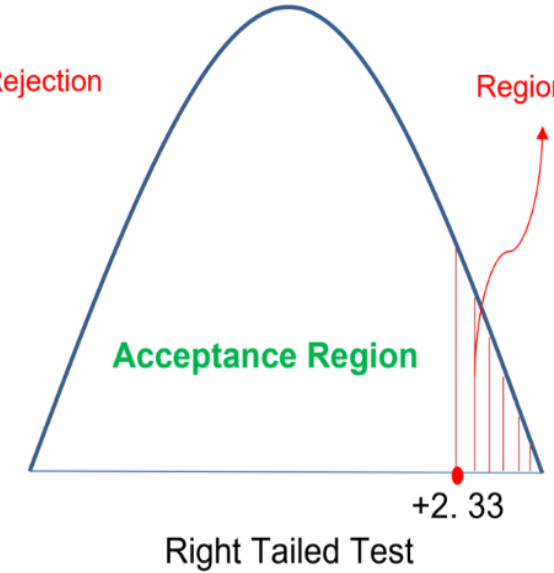


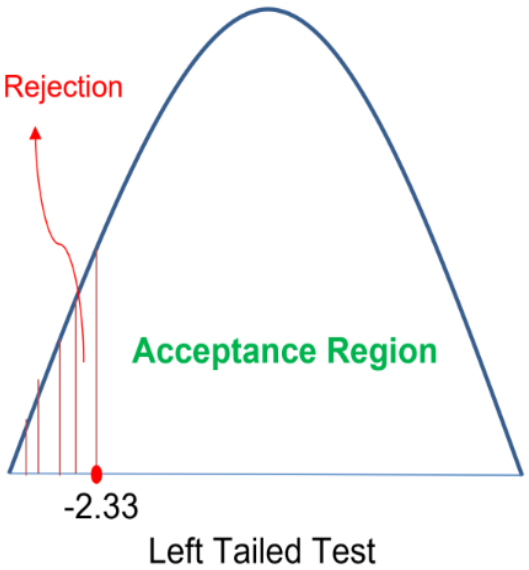
H0: μ ≤ μ0 vs. Ha: μ > μ0

Critical Regions in Hypothesis Testing

Level of Significance	1%	5%	10%
Two Tailed Test	-2.58, +2.58	-1.96, +1.96	-0.645, +0.645
Right Tailed Test	+2.33	+1.645	+1.28
Left Tailed Test	-2.33	-1.645	-1.28







With Level of Significance = 1%

How to Find a Critical Value

Unfortunately, the formulas for finding critical values are very complex. Typically, you don’t calculate them by hand. You can use some statistical software or statistical tables (Z-table, T distribution table, Chi-square table, F-table) to find them.

ALTERNATIVE: P-Value (Another approach to evaluating the significance of test results.)

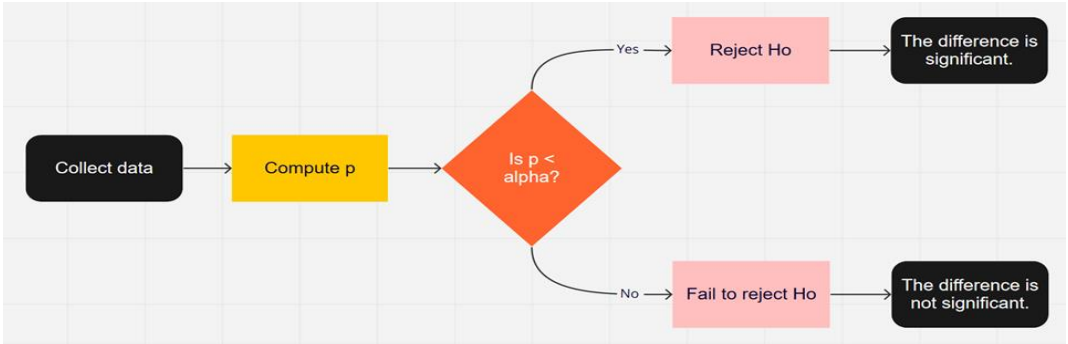
The Use of P –Values in Decision Definition Making

p-value: the probability that the null hypothesis is correct

Smaller p-values indicate stronger evidence against the null hypothesis.

Compare the test statistic with the critical region using p-value:

- If p-value ≤ α, reject the null hypothesis.
- If p-value > α, fail to reject the null hypothesis.



P-Value - Example

Suppose you are testing the following hypothesis at a significance level (α) of 5% and you got the p-value as 3%, and your sample statistic is $\bar{x} = 25$

H_0: $\mu = 20$

H_1: $\mu > 20$

Given, α=5%, we are fine to reject our null hypothesis 5 out of 100 times even though it is true.

P-value is 3% which is less than α
0.03 < 0.05 → We reject the null Hypothesis (extremely strong evidence against the null hypothesis.)

What about P-Value is 6%?

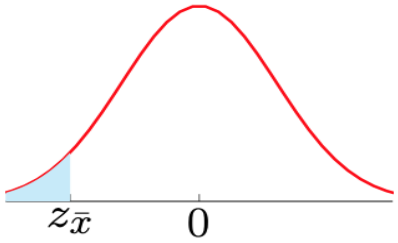
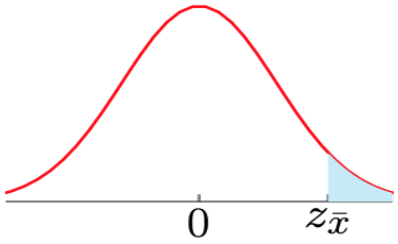
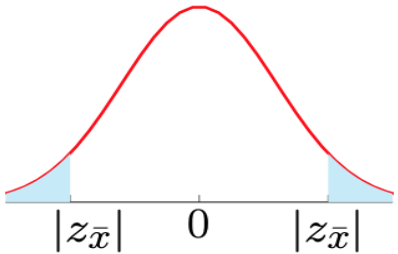
PRACTICE

WRITING NULL AND ALTERNATIVE HYPOTHESIS

The school board claims that atleast 60% of students bring a phone at school. A teacher believe that this number is too high and randomly samples 25 students to test at at level of significance of .02

Write down **H0**, **H, C, alpha.**

P-Values and Types of Tests:

Graph	Test	Conclusion
	1. Left-tailed Test $H_0 : \mu = k \quad H_1 : \mu < k$ P-value $= P(z < z_{\bar{x}})$ This is the probability of getting a test statistic as low as or lower than $z_{\bar{x}}$	If P-value $\leq \alpha$, we reject H_0 and say the data are statistically significant at the level α. If P-value $> \alpha$, we do not reject H_0.
	2. Right-tailed Test $H_0 : \mu = k \quad H_1 : \mu > k$ P-value $= P(z > z_{\bar{x}})$ This is the probability of getting a test statistic as high as or higher than $z_{\bar{x}}$	If P-value $\leq \alpha$, we reject H_0 and say the data are statistically significant at the level α. If P-value $> \alpha$, we do not reject H_0.
	3. Two-tailed Test $H_0 : \mu = k \quad H_1 : \mu \neq k$ P-value $= 2P(z > \|z_{\bar{x}}\|)$ This is the probability of getting a test statistic either lower than $\|z_{\bar{x}}\|$ or higher than $\|z_{\bar{x}}\|$	If P-value $\leq \alpha$, we reject H_0 and say the data are statistically significant at the level α. If P-value $> \alpha$, we do not reject H_0.

RECAP:

A hypothesis test is a method that evaluates population claims (or hypothesis) using sample data.

- The conclusion of a hypothesis test determines whether the observed (sample) data support the claim or suggest an alternative explanation.
    - **The null hypothesis** typically represents the claim of no effect or no difference
        - E.g., The new drug has no effect on blood pressure.
    - **The alternative hypothesis** represents the claim of an effect or difference.
        - E.g., The new drug reduces blood pressure.

We are interested in "whether we can **reject** or **fail to reject** the null hypothesis?"

RECAP:

We are interested in “whether we can **reject** or **fail to reject** the null hypothesis?”

How to answer this question?

We calculate the statistics from our sample data. The statistics used here called Test Statistics (TODAY’s topic). These test statistics help us evaluate hypotheses about population parameters.

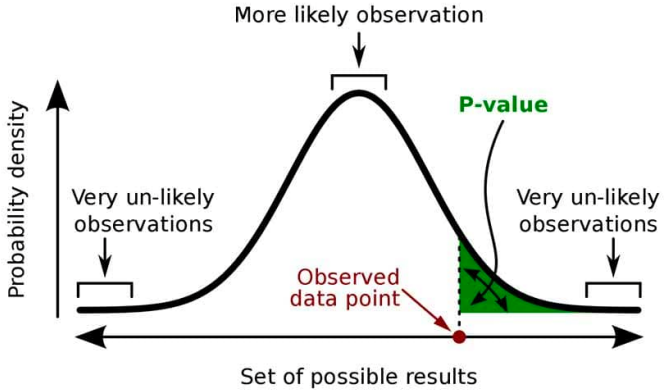
How? By applying different types of test statistics, we obtain p-values, which are probability values used to test the hypothesis statement.

- Draw a conclusion about the null hypothesis based on the statistical evidence provided by the p-value
  - **if p-value ≤ predefined significant level (α) [threshold]**, we reject the null hypothesis (H0).
  - **if p-value > α**, we fail to reject the null hypothesis (H0).

RECAP:

The p-value is a measure utilized in hypothesis testing to quantify the strength of evidence against the null hypothesis.

- It represents the probability of observing the test statistic, or a more extreme value, if the null hypothesis is true.
- **A low p-value indicates that** the observed data are unlikely under the null hypothesis, suggesting strong evidence against it.
- Typically, a significance level (alpha) is chosen, such as 0.05, and
- if the p-value is less than or equal to alpha, the null hypothesis is rejected.



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

Different Types of Statistical Tests

Statistical Tests

A statistical hypothesis test is a method of statistical inference used to decide whether the data at hand sufficiently support a particular hypothesis.

- Test statistics such as z-tests,one sample t-tests, chi-square tests, etc., are commonly used to assess hypotheses about population parameters based on sample data.

NOTE: Generally, in hypothesis tests, test statistic means to obtain all of the sample data and convert it to a single value. For example, Z-test calculates Z statistics, t-test calculates t-test statistic, and F-test calculates F values etc., are the test statistics. Test statistics need to compare to an appropriate critical value (cv) or p-value. A decision can then be made to reject or not reject the null hypothesis.

Does knowing more help us?

Yes! If we have an **idea of the standard deviation of the underlying population** (or even just have enough data to make an estimate), we can use a **z-test instea**d, which give more accurate results.

According to the theory, we cannot use **z-tests for sample sizes under 30 elements**.

Z-Test

The Z-test compares a sample mean to a population mean.

- It is used when you have a large sample size (typically n > 30) and you know the population standard deviation (σ) or can make a reasonable assumption about it.

Assumption: It assumes that the sample data is normally distributed or that the sample size is large enough for the Central Limit Theorem to apply

Compute the **Z-Test Statistic** using the sample mean, μ_1, the population mean, μ_0, the number of data points in the sample, n and the population's standard deviation, σ:

$$z = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}}$$

T-Test

T-test is a statistical test used to determine if there is a significant difference between the means of two groups.

- It is used when you have a smaller sample size (typically $n < 30$) or when you don't know the population standard deviation (σ) and must estimate it from the sample.

In General, the type of test statistic used depends on the number of samples being compared

- **One Sample:** when there is only one sample that needs to be compared with a given value.
- **Two Samples**, when there are two or more samples to be compared. In this case, tests can include correlation tests and tests for differences between samples.

Additionally, samples can be paired or not paired.

- Paired samples are also called **dependent samples (***observations that are related or matched in some way***)**, while not paired samples are also called **independent samples (***not related or matched***)**.

One sample T-Test

Determine if the mean of a **single sample** is significantly different from a known or hypothesized population mean.

- Commonly used when you have collected data from a single group or sample and want to compare its mean to a specific value or a hypothesized population mean.

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

$\bar{x}$ = obersved mean of the sample
μ = assumed mean
s = standard deviation
n = sample size

How to run a one-sample t test:

```
import numpy as np

from scipy import stats

 stats.ttest_1samp(your_data, popmean=0.5)
```

>>> TtestResult(statistic=2.456308468440, pvalue=0.017628209047638, df=49)

Two Sample t-test

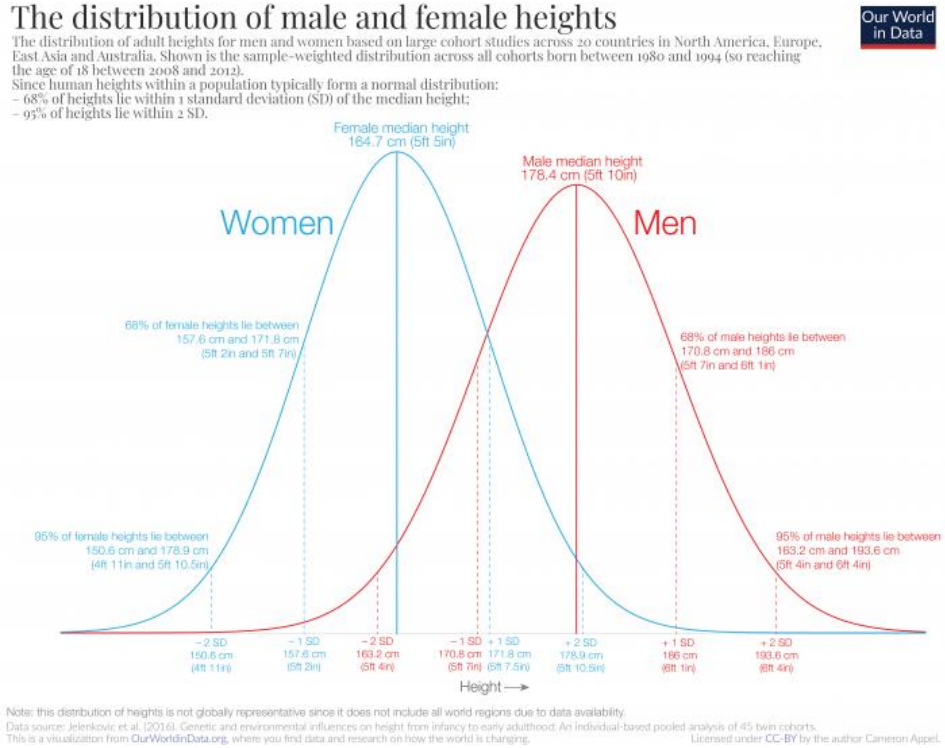
The two sample t test, (also referred to as the unpaired t test), is used to compare the means of two different samples.

Example: We have noticed most humans fall into one of two distinct categories– male or female. We would like to know if our sample of males is taller than our sample of females.

Can we just take the average of the two samples?

Two-Sample T-test

- NO! Simply taking the average of the two samples is not sufficient to determine if one group is taller than the other (does not account for variability within each group).
- Instead, we would conduct a statistical test to determine whether there is a statistically significant difference between the two groups.



Two Sample T-Test

- **Null hypothesis:** Men and women are the same height
- **Alternative hypothesis:** Men and women are different heights
- **p-value:** the probability that we would see these observations if the null hypothesis is true/correct

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$\bar{x}_1$ = observed mean of 1st sample
$\bar{x}_2$ = observed mean of 2nd sample
s_1 = standard deviation of 1st sample
s_2 = standard deviation of 2nd sample
n_1 = sample size of 1st sample
n_2 = sample size of 2nd sample

When can we use the T-test?

- We can use the test for continuous data obtained from a random sample that follows a normal distribution.

Our Assumptions: For the t-test we assume:

- The data is normally distributed
- We care about **the mean** → ex: whether there is a significant difference in the means of the two groups.

Q: What if my data isn't nearly normally distributed?

- If your **sample sizes are very small**, you might not be able to test for normality. You might need to rely on your understanding of the data.
- When you cannot safely assume normality, you can **perform a nonparametric test** that doesn't assume normality.

Paired Sample t test

The paired sample t test is used to compare the means of **two related groups** of samples.

- It is used in a situation where you have two values (i.e., a pair of values) for the same group of samples.
- Often these two values are measured from the same samples either at two different times, under two different conditions, or after a specific intervention.

Paired Sample t test: Example

The aliens monitor a bunch of humans, test them for intelligence, and then run one half of them through a machine to make them smarter. Afterwards, they want to know if their machine worked.

This would be called a **paired t-test.**

Null Hypothesis: ?

Alternative Hypothesis: ?

Paired Sample t test: Example

The aliens monitor a bunch of humans, test them for intelligence, and then run one half of them through a machine to make them smarter. Afterwards, they want to know if their machine worked.

Null Hypothesis: The machine did nothing

Alternative Hypothesis: The machine came from a different distribution

Paired Sample t test: Example

The aliens monitor a bunch of humans, test them for intelligence, and then run one half of them through a machine to make them smarter. Afterwards, they want to know if their machine worked.

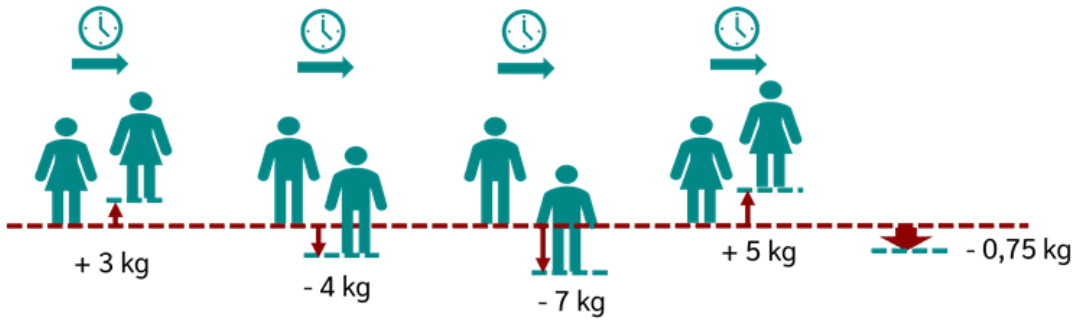
Null Hypothesis: The machine did nothing

Alternative Hypothesis: The machine has effect on human intelligence

Ques: The aliens get a p-value of .05. What can they conclude?

$$t = \frac{\bar{x}_{diff} - 0}{s_{\bar{x}}} \qquad s_{\bar{x}} = \frac{s_{diff}}{\sqrt{N}}$$

Paired Sample t test: Example



Number of cases

$n = 4$

Degrees of freedom

$df = n - 1 = 3$

Mean

$$\bar{x} = \frac{+3 - 4 - 7 + 5}{4} = -0.75$$

Standard deviation

$$s = \sqrt{\frac{(3 + 0.75)^2 + (-4 + 0.75)^2 + (-7 + 0.75)^2 + (5 + 0.75)^2}{4 - 1}} = 5.68$$

Standard error of the mean

$$s_e = \frac{s}{\sqrt{n}} = \frac{5.68}{2} = 2.84$$

t-Value

$$t = \frac{\bar{x} - 0}{s_e} = \frac{\bar{x}}{s_e} = \frac{-0.75}{2.84} = -0.26$$

What about this?

We are monitoring **birds from two different places on the planet,** and get the following results:

Bird Type	Location A	Location B
Grackle	7	13
Pigeon	2	7
Sea pigeon	15	1
One of those big fish-beak things	13	0
Big long bird	22	0
Bat	3	4

Each bird type and location falls into distinct categories, making them categorical variables suitable for analysis using methods like the **chi-square test**.

We want to find out if two different places on Earth have the same types of birds

Do these locations have the same underlying bird population?

Enter the Chi Square Test! A test for checking if two sets of categorical variables come from the same distribution.

Null hypothesis: ?

Alternative hypothesis: ?

Do these locations have the same underlying bird population?

Enter the Chi Square Test! A test for checking if two sets of categorical variables come from the same distribution.

Null hypothesis: The bird populations observed in Location A and Location B are the same.

Alternative hypothesis: The bird populations observed in Location A and Location B are different.

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Recap: Tests so Far

- **One sample t-test:** Tells how likely it is a single sample of normally distributed data would be generated by a specific mean
- **Two sample t-test:** Tells how likely it is two samples would be generated by a population with the same mean
- **Chi-squared test:** Estimates the chances two sets of categorical data come from the same distribution

Multiple Groups

The Aliens decide to kidnap humans to study, but we don't know what humans eat! We have five different food mixes we want to try. We split the humans up into five groups and feed each group a different mix, and then measure how much the humans grow over the next few years.

Ques: How do Aliens know if the mixes have different effects?

Anova (Analysis of Variance) Test

ANOVA is a powerful statistical test for comparing the **means of multiple groups (three or more groups (more than two))** to determine if there are significant differences among them.

We use a **anova** test.

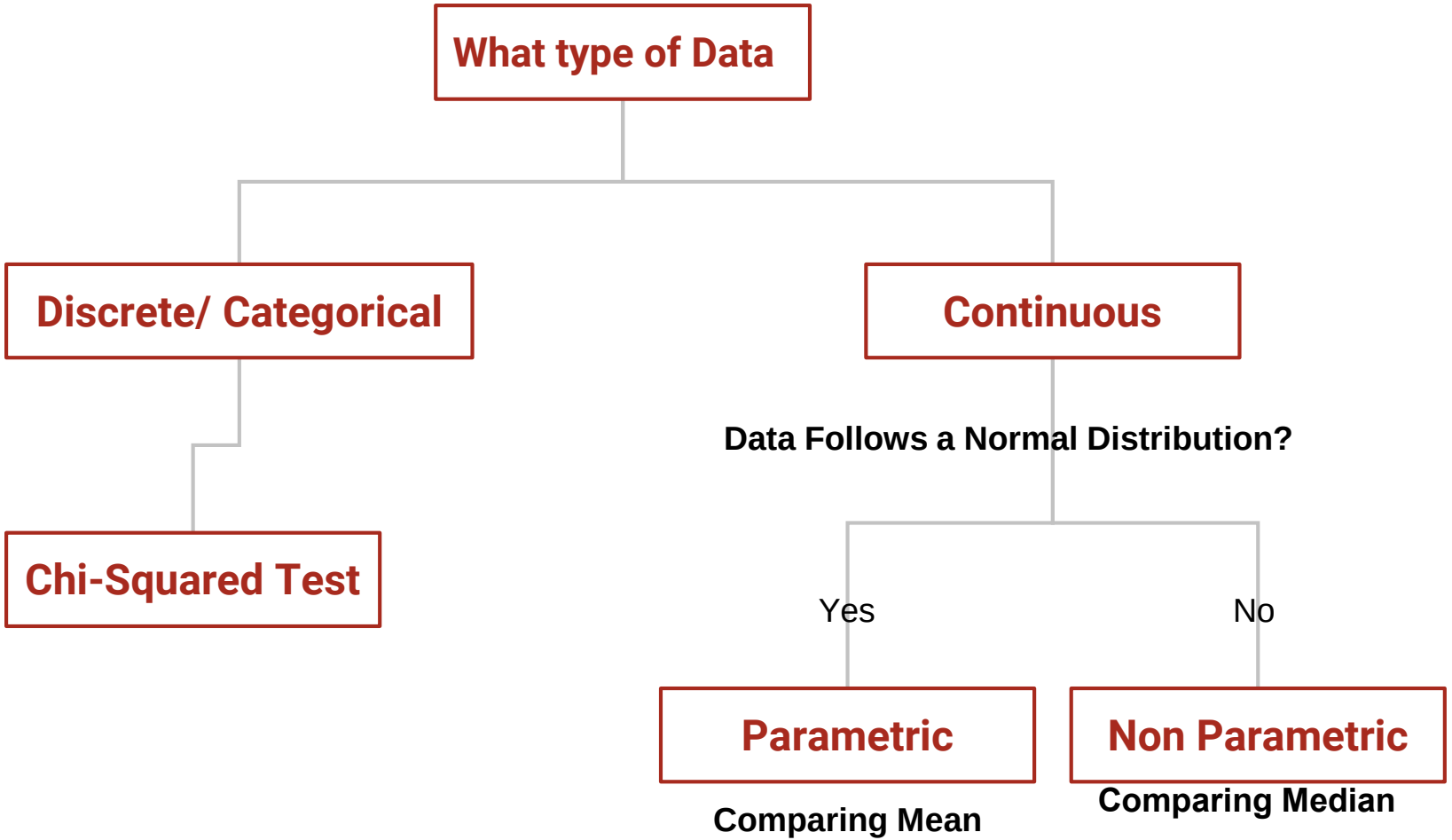
- Null hypothesis:
- Alternative hypothesis:

Anova Test

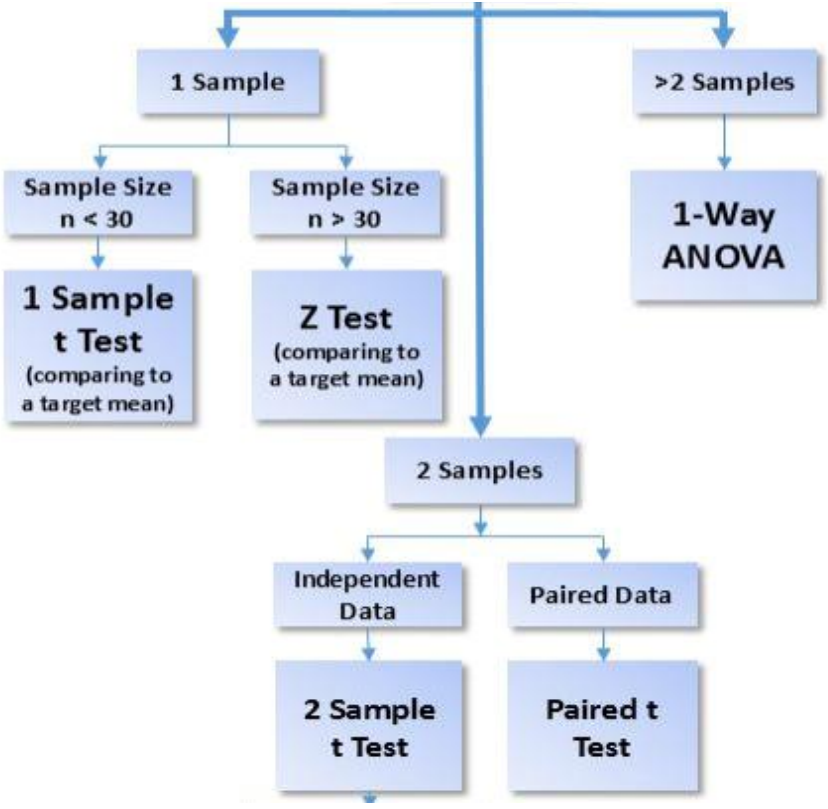
We use a **anova** test.

- **Null hypothesis:** There is no difference between any of the groups
- **Alternative hypothesis:** There is a difference between at least one of the groups

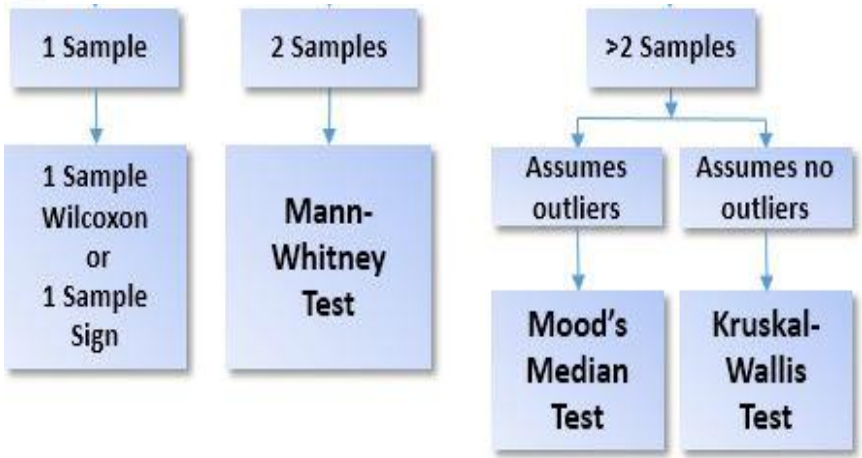
Notes: In t-tests and z-tests, we typically compare means of two groups using individual datasets or assess the mean of a single group against a known value. **ANOVA** evaluates differences in means across three or more groups as a whole, considering both within-group and between-group variability.


What type of Data
Discrete/ Categorical
Continuous
Chi-Squared Test
Data Follows a Normal Distribution?
Yes
No
Parametric
Non Parametric
Comparing Mean
Comparing Median

Parametric Tests (Comparing Means)



Nonparametric Tests (Comparing Medians)



Kruskal-Wallis Test

This nonparametric test is an extension of the one-way ANOVA and is used to compare the medians of more than two independent groups (three or more independent groups) when the data do not meet the assumptions of normality required for ANOVA.

Mann-Whitney U Test

This nonparametric test is used to

- compare two independent groups
- when the data do not meet the assumptions of normality required for a t-test.

Wilcoxon Signed-Rank Test

This nonparametric test is used to

- compare two related or paired groups.
- when the data do not meet the assumptions of normality required for a paired t-test.

Considerations: (How to decide an appropriate statistical test?)

- What are you curious about?
  - Mean? Standard deviation? Frequency?
- Is your data categorical or continuous?
- Do you have one or two samples?
- Is your data normally distributed?
  - If it is, you would use a **parametric** test. If it is *very* non-normal, you would use a non-parametric test
- Is your data paired? Is there a before and after?