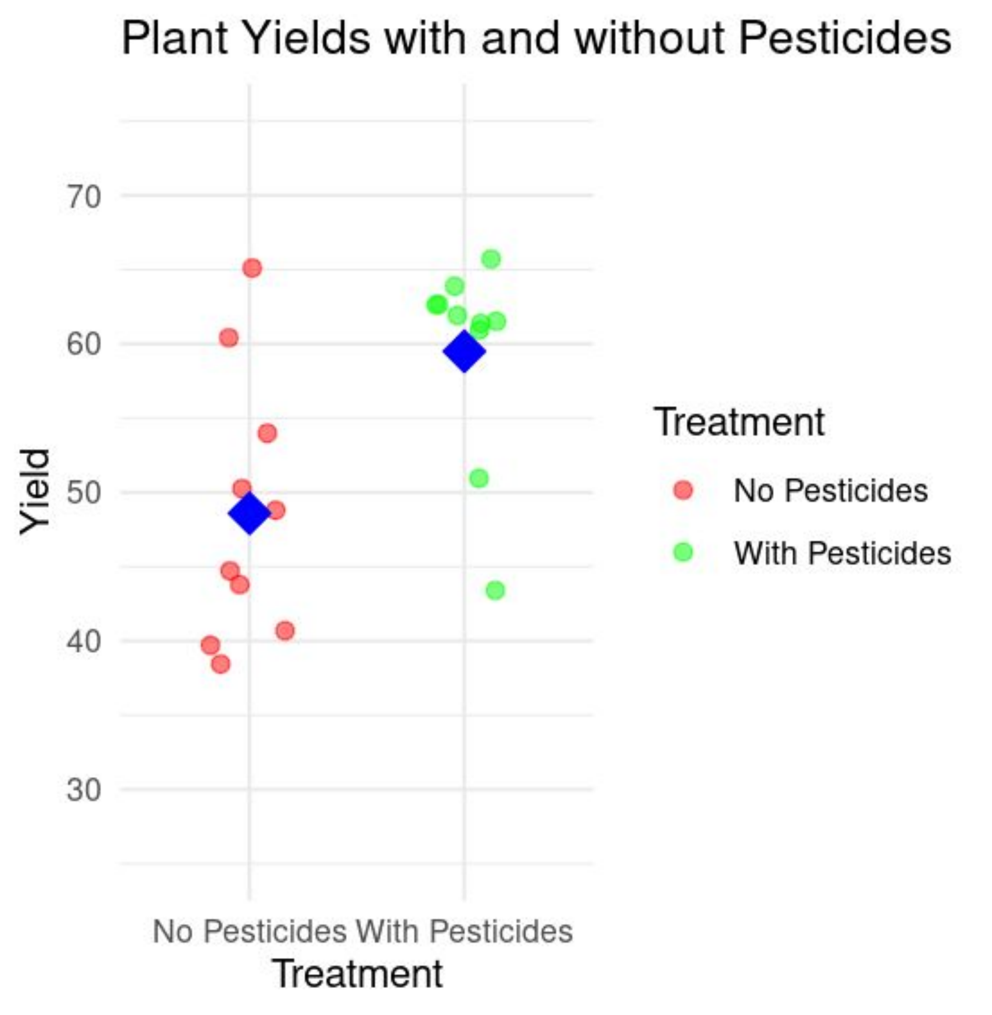
Hypothesis Testing in R

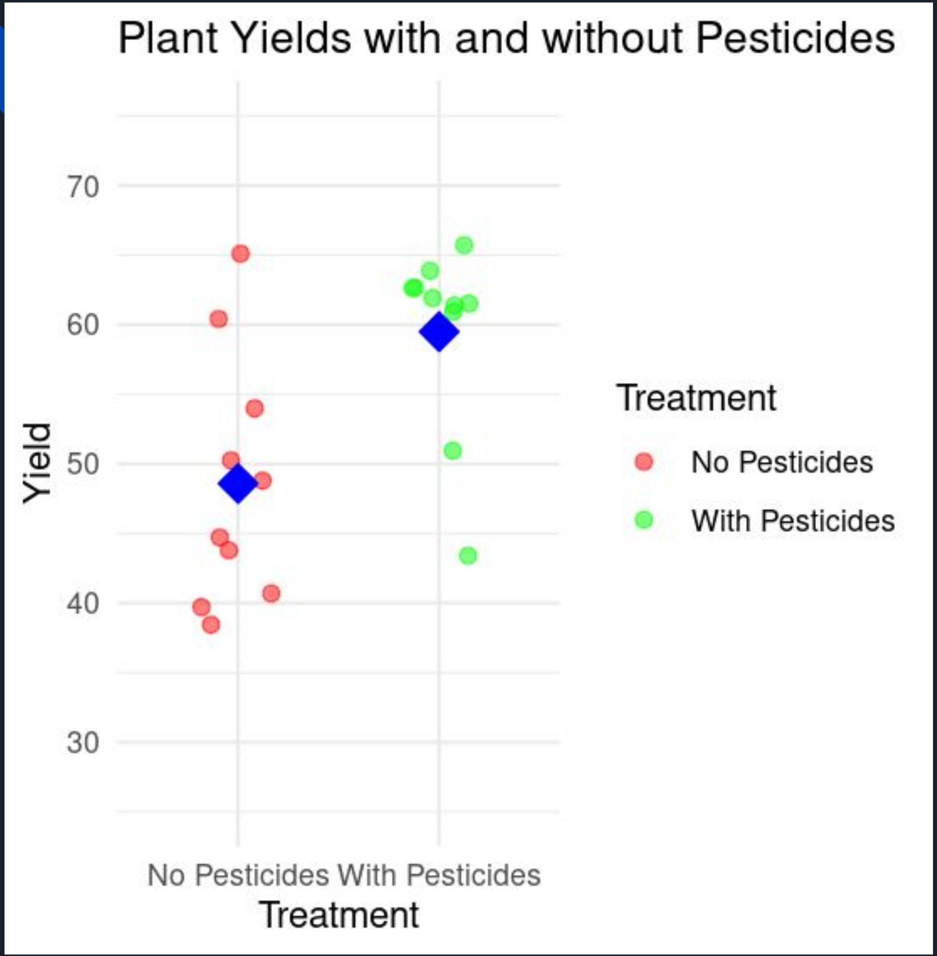
By: Michael Remington

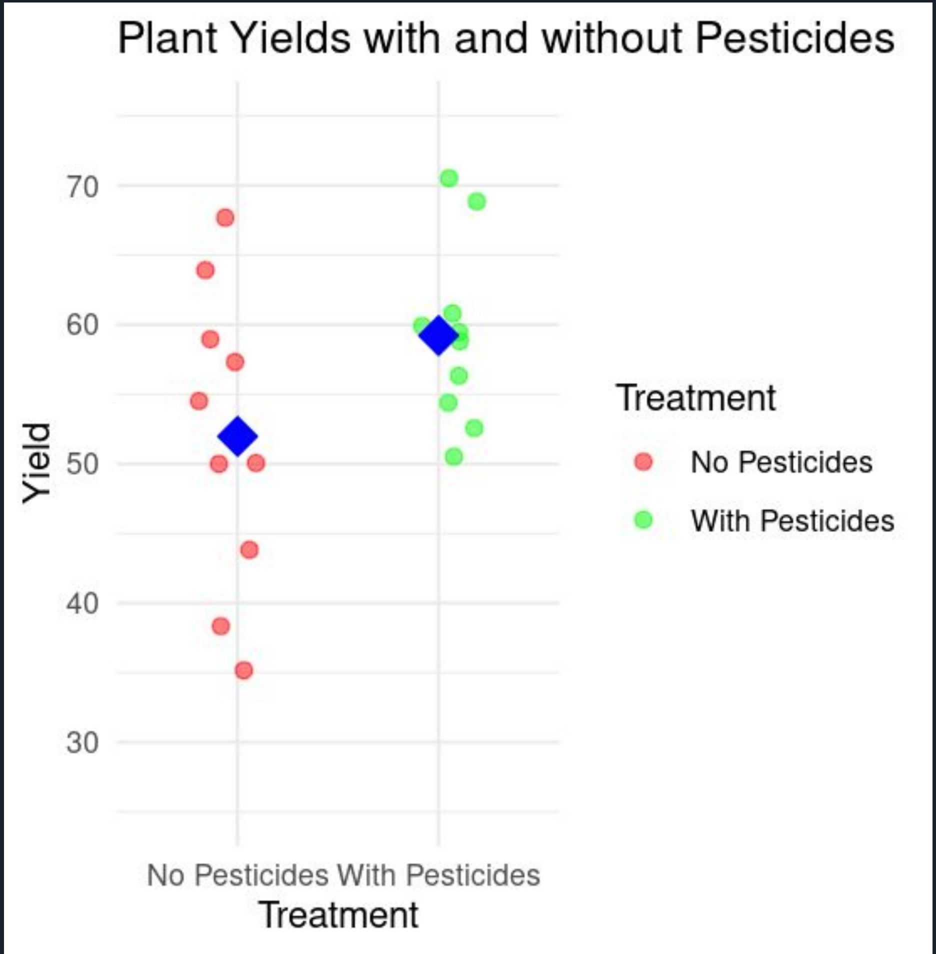
Table Of Contents

1. Introduction to Hypothesis Testing
    - Definition of hypothesis testing
    - Types of hypotheses (null and alternative)
    - Type I and Type II errors
    - Significance level and p-value
2. Common Statistical Tests
    - T-test f/ Z Test
    - Two-sample t-test for difference between means
    - Chi-square test for goodness of fit and independence
    - Anova
3. Power Analysis
4. Introduction to A/B Testing
5. Practical Example and Exercise
    - Walk through a real-world example of A/B testing
    - Hands-on exercise: Conduct a hypothesis test or analyze the results of an A/B test using a provided dataset
    - Discussion of results and interpretation

Intro to Hypothesis Testing


Plant Yields with and without Pesticides
70
60
50
40
30
Yield
No Pesticides With Pesticides
Treatment
Treatment
No Pesticides
With Pesticides


Plant Yields with and without Pesticides
Yield
70
60
50
40
30
No Pesticides With Pesticides
Treatment
Treatment
No Pesticides
With Pesticides


Plant Yields with and without Pesticides
Yield
70
60
50
40
30
No Pesticides With Pesticides
Treatment
Treatment
No Pesticides
With Pesticides

What is Hypothesis Testing?

Hypothesis Testing: a method of statistical inference used to decide whether the data sufficiently support a particular hypothesis

<u>Workflow</u>

1. Defining a null hypothesis (H0) and an alternative hypothesis (H1).
2. Collect Data
3. Choosing a significance level.
4. Determining a test statistic
5. Calculating the p-value.
6. Comparing the p-value to α to decide whether to reject H0.

Alternative vs Null Hypothesis

Null Hypothesis (Ho):

- There is no difference in [outcome variable] between [groups/conditions]
- There is no difference in average yield between Pesticides and No Pesticides
- There is no association to average yield when adding pesticides and not adding pesticides.

Alternative Hypothesis (H1):

- There is a difference in [outcome variable] between [groups/conditions]
- There is a difference in average yield between Pesticides and No Pesticides
- There is a significant effect on average yield by adding pesticides

Workflow

1. Defining a null hypothesis (H0) and an alternative hypothesis (H1).
2. Collect Data
3. Choosing a significance level (α), often 0.05.
4. Determining a test statistic based on the sample data.
5. Calculating the p-value
6. Comparing the p-value to α to decide whether to reject H0.

Data collection

1. Define Your Population Clearly

2. Determine your control group, experimental groups

3. Techniques for sampling

   - Random Sampling Techniques
   - Cluster Sampling for Large Populations
   - Systematic Sampling
   - Avoid Convenience sampling when possible

4. Ensure Sample Size is Adequate

   - Power analysis
   - Cost Benefit

5. Clean Data!!!!!!!!

   - Clearly define categories and data inputs
   - Consistent sampling

6. Optional: Repeated Measures

   - Constant sample times
   - Account for seasonality

Workflow

1. Defining a null hypothesis (H0) and an alternative hypothesis (H1).
2. Collect Data
3. Choosing a significance level (α), often 0.05.
4. Determining a test statistic based on the sample data.
5. Calculating the p-value
6. Comparing the p-value to α to decide whether to reject H0.

Type 1 and Type 2 errors

- Type I Error: Rejecting the null hypothesis when it is actually true. Also known as a '**false positive.’**
- Type II Error: Failing to reject the null hypothesis when it is actually false. Also known as a '**false negative.'**

	Truth		
Decision Based on Data		**H0 True**	**H0 False**
	Reject H0	**Type I Error (α)** **False Positive**	**Correct Decision (Power)** **True Positive**
	Fail to Reject H0	**Correct Decision ($1-\alpha$)** **True Negative**	**Type II Error (β)** **False Negative**

Assessing Confidence

- <u>α (alpha)</u>
  - represents the significance level, the probability of committing a Type I error.
- <u>β (beta)</u>
  - represents the probability of committing a Type II error.
- <u>Power ($1-\beta$)</u>
  - The probability of correctly rejecting a false null hypothesis.

<table>
<tr><td></td><td colspan="3">Truth</td></tr>
<tr><td rowspan="3">Decision Based on Data</td><td></td><td>H0 True</td><td>H0 False</td></tr>
<tr><td>Reject H0</td><td>Type I Error (α) False Positive</td><td>Correct Decision (Power) True Positive</td></tr>
<tr><td>Fail to Reject H0</td><td>Correct Decision (1-α) True Negative</td><td>Type II Error (β) False Negative</td></tr>
</table>

Type 1 and Type 2 errors

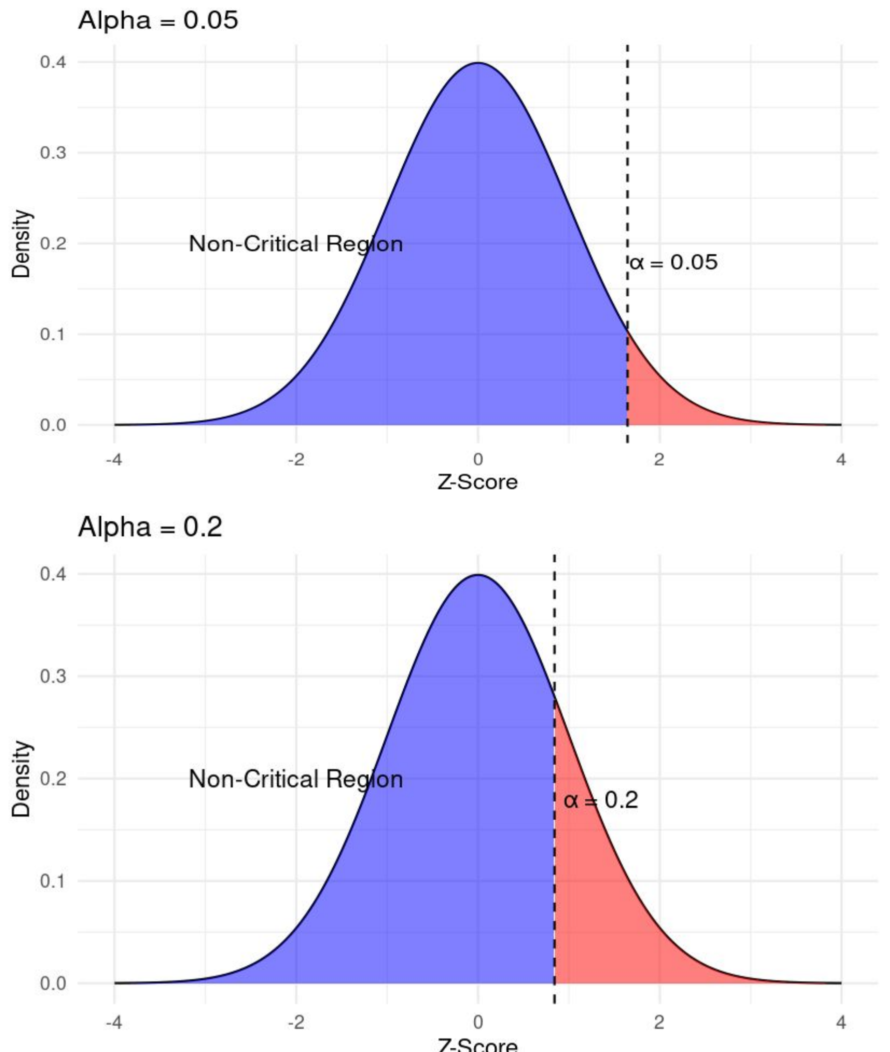
	Truth		
Decision Based on Data		**Patient Does Not Have Cancer**	**Patient Does Have Cancer**
	Diagnosing A Patient as Having Cancer	**False Positive**	**True Positive**
	The Patient is Healthy	**True Negative**	**False Negative**

Understanding α (Alpha) And P values

P Value Definition: The p-value measures the probability of observing the collected data, or something more extreme, if the null hypothesis is true.

α (alpha):

- the predetermined threshold for significance in hypothesis testing.
- It represents the probability of committing a Type I error which is rejecting the null hypothesis when it is actually true.


Alpha = 0.05
0.4
0.3
0.2
0.1
0.0
Density
Non-Critical Region
α = 0.05
-4
-2
0
2
4
Z-Score
Alpha = 0.2
0.4
0.3
0.2
0.1
0.0
Density
Non-Critical Region
α = 0.2
-4
-2
0
2
4
Z-Score

Workflow

1. Defining a null hypothesis (H0) and an alternative hypothesis (H1).
2. Collect Data
3. Choosing a significance level (α), often 0.05.
4. Determining a test statistic based on the sample data.
5. Calculating the p-value
6. Comparing the p-value to α to decide whether to reject H0.

What is a Test Statistic?

Test statistic

- a numerical value used to evaluate the strength of the evidence against the null hypothesis
- compares observed data to what is expected under the null hypothesis, quantifying the difference or effect observed in the sample.

Example

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

t: The t-statistic.

x̄: The sample mean.

μ: The population mean or the mean value under the null hypothesis.

s: The sample standard deviation.

n: The sample size.

Types Of Test statistics

Z-test Statistic

- For comparing sample means to a population mean with known population standard deviation (large samples).

T-test Statistic

- One-sample t-test: Tests the mean of a single group against a known mean.
- Two-sample t-test (independent): Compares the means of two independent groups.
- Paired t-test: Compares the means of two related groups.

Chi-square (χ2) Statistic

- Goodness-of-fit test: Determines if a sample matches the expected distribution.
- Test of independence: Assesses whether two categorical variables are independent.

F-test Statistic

- For comparing variances of two populations.
- ANOVA (Analysis of Variance): Compares means among three or more groups.

Knowing data types and their distribution

- What is your hypothesis?
  - Difference in variance
  - Difference in mean
- Number of groups you have in your experiment
- Categorical versus continuous data
- What's your sample size?
- Is your data normally distributed?
  - Parametric versus non-parametric tests
- What does the variance of your groups look like?
- Does your experiment have covariates or random effects?

Example: Chi Squared Statistic

- Compared observed Data to expected

	Snack A	Snack B	Total
Children	30	10	40
Adults	20	40	60
Total	50	50	100

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

- Expected frequency for Children preferring Snack A:
  - (40×50)/100=20
- Expected frequency for Children preferring Snack B:
  - (40×50)/100=20
- Expected frequency for Adults preferring Snack A:
  - (60×50)/100=30
- Expected frequency for Adults preferring Snack B:
  - (60×50)/100=30

Expected Frequencies

- Expected frequency for Children preferring Snack A:
  - 20
- Expected frequency for Children preferring Snack B:
  - 20
- Expected frequency for Adults preferring Snack A:
  - 30
- Expected frequency for Adults preferring Snack B:
  - 30

	Snack A	Snack B	Total
Children	30	10	40
Adults	20	40	60
Total	50	50	100

Chi Squared Values

- For Children preferring Snack A:
  a. (30−20)^2/20=5
- For Children preferring Snack B:
  a. (10−20)^2/20=5
- For Adults preferring Snack A:
  a. (20−30)^2/30=3.33
- For Adults preferring Snack B:
  a. (40−30)^2/30=3.33

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

$\chi 2=5+5+3.33+3.33=\boxed{16.66}$

Workflow

1. Defining a null hypothesis (H0) and an alternative hypothesis (H1).
2. Collect Data
3. Choosing a significance level (α), often 0.05.
4. Determining a test statistic based on the sample data.
5. Calculating the p-value
6. Comparing the p-value to α to decide whether to reject H0.

Degrees of Freedom

Degrees of Freedom: is the number of parameters of the system that may vary independently.

Free Values that can vary (Observations, vs parameters)

Chi-Squared df = (rows−1)×(columns−1),

<u>For our example</u>

(2-1)*(2-1) = 1 df

P value

<u>Chi Squared Distribution</u>

1 df

x^2 = 16.66

P value < 0.001

	P										
DF	0.995	0.975	0.2	0.1	0.05	0.025	0.02	0.01	0.005	0.002	0.001
1	.0004	.00016	1.642	2.706	3.841	5.024	5.412	6.635	7.879	9.55	10.828
2	0.01	0.0506	3.219	4.605	5.991	7.378	7.824	9.21	10.597	12.429	13.816
3	0.0717	0.216	4.642	6.251	7.815	9.348	9.837	11.345	12.838	14.796	16.266
4	0.207	0.484	5.989	7.779	9.488	11.143	11.668	13.277	14.86	16.924	18.467
5	0.412	0.831	7.289	9.236	11.07	12.833	13.388	15.086	16.75	18.907	20.515
6	0.676	1.237	8.558	10.645	12.592	14.449	15.033	16.812	18.548	20.791	22.458
7	0.989	1.69	9.803	12.017	14.067	16.013	16.622	18.475	20.278	22.601	24.322
8	1.344	2.18	11.03	13.362	15.507	17.535	18.168	20.09	21.955	24.352	26.124
9	1.735	2.7	12.242	14.684	16.919	19.023	19.679	21.666	23.589	26.056	27.877
10	2.156	3.247	13.442	15.987	18.307	20.483	21.161	23.209	25.188	27.722	29.588
11	2.603	3.816	14.631	17.275	19.675	21.92	22.618	24.725	26.757	29.354	31.264
12	3.074	4.404	15.812	18.549	21.026	23.337	24.054	26.217	28.3	30.957	32.909
13	3.565	5.009	16.985	19.812	22.362	24.736	25.472	27.688	29.819	32.535	34.528
14	4.075	5.629	18.151	21.064	23.685	26.119	26.873	29.141	31.319	34.091	36.123
15	4.601	6.262	19.311	22.307	24.996	27.488	28.259	30.578	32.801	35.628	37.697
16	5.142	6.908	20.465	23.542	26.296	28.845	29.633	32	34.267	37.146	39.252
17	5.697	7.564	21.615	24.769	27.587	30.191	30.995	33.409	35.718	38.648	40.79
18	6.265	8.231	22.76	25.989	28.869	31.526	32.346	34.805	37.156	40.136	42.312
19	6.844	8.907	23.9	27.204	30.144	32.852	33.687	36.191	38.582	41.61	43.82
20	7.434	9.591	25.038	28.412	31.41	34.17	35.02	37.566	39.997	43.072	45.315

Workflow

1. Defining a null hypothesis (H0) and an alternative hypothesis (H1).
2. Collect Data
3. Choosing a significance level (α), often 0.05.
4. Determining a test statistic based on the sample data.
5. Calculating the p-value
6. Comparing the p-value to α to decide whether to reject H0.

Putting it all together

Alpha = 0.05

P value < 0.001

0.001 < 0.05

Reject the null hypothesis that there is no statistical difference in snack preference between adults and children

- Reject the null versus fail to reject null

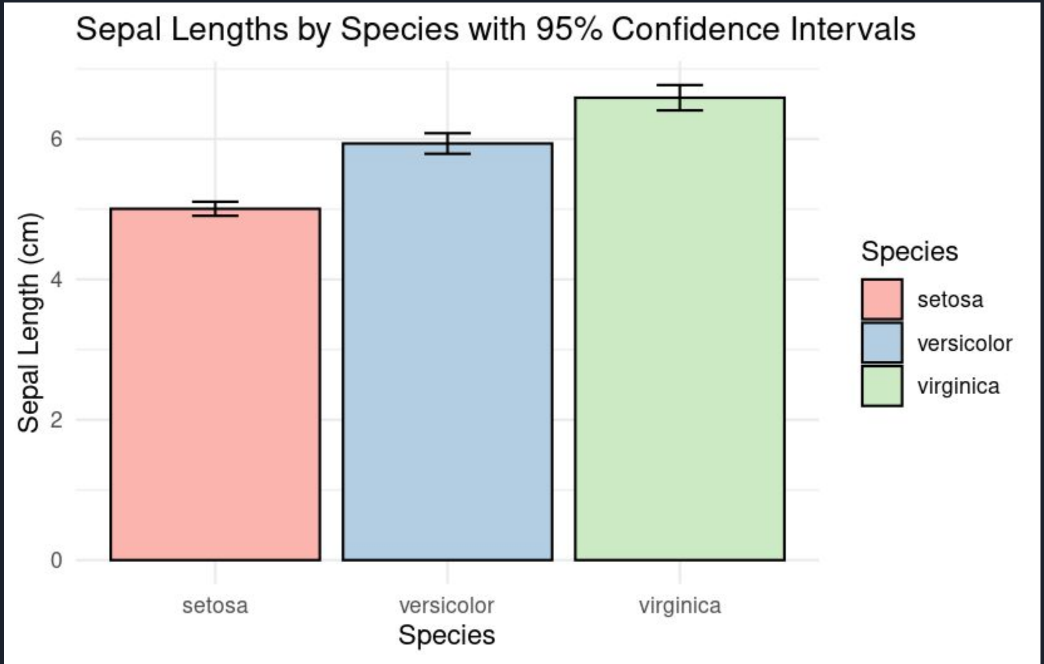
Workflow

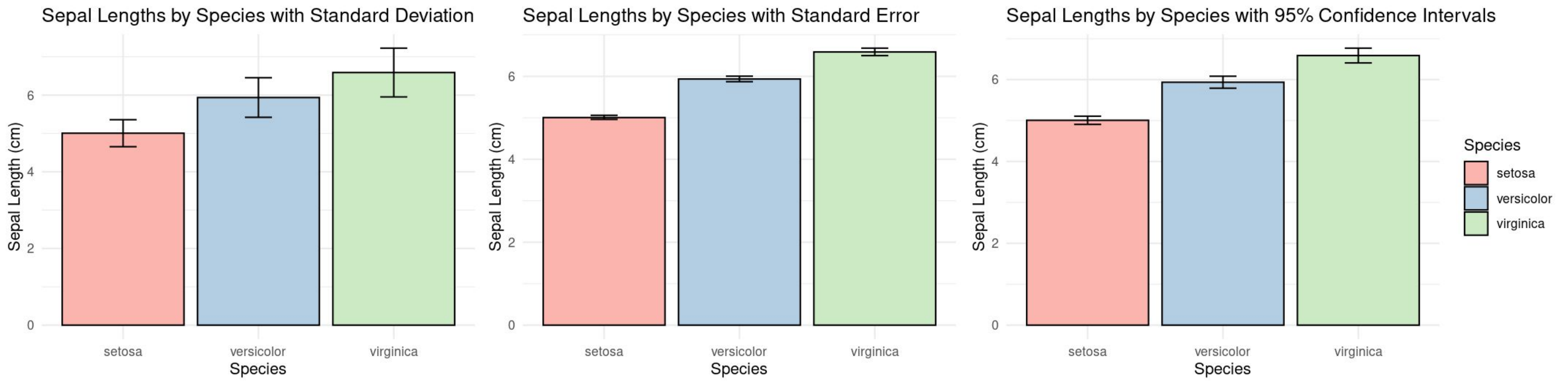
1. Defining a null hypothesis (H0) and an alternative hypothesis (H1).
2. Collect Data
3. Choosing a significance level (α), often 0.05.
4. Determining a test statistic based on the sample data.
5. Calculating the p-value
6. Comparing the p-value to α to decide whether to reject H0.

Confidence Intervals

Types of Error Bars

- Standard deviation (SD)
- Standard error (SE): SD/sqrt(N)
- Confidence intervals (CI): mean ±(Zscore×SE)




Sepal Lengths by Species with Standard Deviation
Sepal Lengths by Species with Standard Error
Sepal Lengths by Species with 95% Confidence Intervals
Sepal Length (cm)
0
2
4
6
setosa
versicolor
virginica
Species

Power Analysis

Statistical Power: The probability that a test will correctly reject the null hypothesis (H0) when the alternative hypothesis (H1) is true. In simpler terms, it's the likelihood that a study will detect an effect when there is an effect truly present.

Power = $(1-\beta)$

- Used to determine the appropriate sample size
- Determined by the effect size
    - Magnitude of the change in the relationship
- Can be done before or after the analysis
- Minimizes False Negatives/Type two errors

<u>Other Practical Uses</u>

- Grant Proposals and Study Design
- Interpreting Non-significant Results

A/B Testing

- A form of hypothesis testing for practical application comparing two variables
- Data driven decision making
- Examples: marketing campaigns, web design, medical research

See R Code