Data Analysis

- Matt Grieshop
- MSU Department of Entomology

Lecture Overview

- Where do we start?
- Hypothesis development
- Measures of centrality and dispersion
- Graphical data analysis/exploration
- Hypothesis testing 101
- T tests: Continuous
- Chi Squared: Categorical

Where do we start?

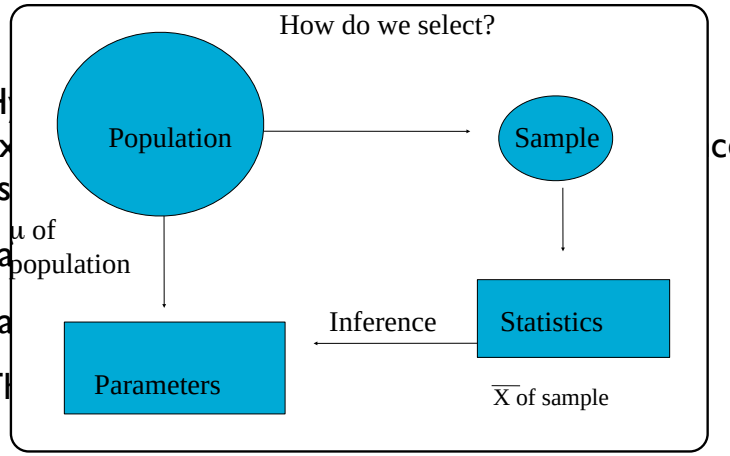
- Develop Hypotheses
- Summarize data by logical groups
- Visualize data
- Formal testing: Apply hypothesis tests if appropriate

Hypothesis Development

- Hypothesis: a supposition or proposed explanation made on the basis of limited evidence as a starting point for further investigation
- Samples vs. Populations
- Sample mean vs. population mean
- The central limit theory and inference

Hypothesis Development

- H
  ex
  as
- Sa
- Sa
- Th



ce

Hypothesis Development

- We never prove a hypothesis…..
- Start with a Null hypothesis
- Develop an Alternative hypothesis

Research Question
↓
Research Hypothesis
↓
Test

In research we

<u>NEVER</u>

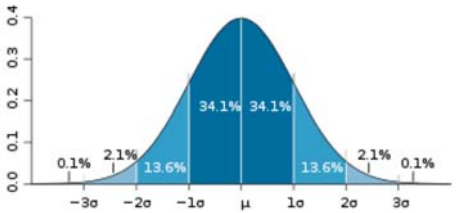
prove a hypothesis!

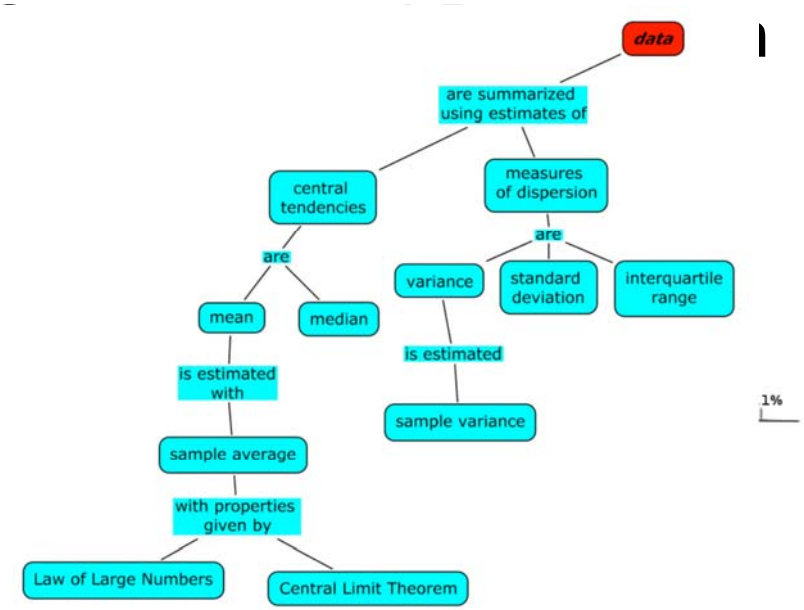
Hypothesis Development

- A. cucmeris populations are the same with and without A. coriaria
- A. cucumeris populations are different without A. coriaria
- A. cucumeris populations are higher without A. coriaria

Centrality and Dispersion

- Centrality provides an estimate of the middle point of a sample
- Dispersion provides an estimate of variability
- Centrality: Mean and Median
- Dispersion: Range, Variance, Standard Deviation, Standard Error





Centrality: Mean

- The mean ="the average"
- Sample total/sample size
- (92 + 135 + 131 + 306 + 218)/5 Mites from the *A. cucumeris* + *Atheta coriaria* treatment
- 882/5 = 176.4

Cucumeris Total	Cucumeris + Atheta Total
684	92
555	135
914	131
1076	306
1111	218

Dispersion

- Range: 92-306
- Standard Deviation (SD): a quantity calculated to indicate the extent of deviation for a group as a whole. The average absolute difference of individual samples to the mean.
- Standard Error of the Mean (SEM): The standard deviation divided by the square root of the sample size

Standard Deviation (Sample)

- Because this comes from a sample we reduce our demoninator by 1.
- Consider our previous data (92, 135, 131, 306, 218)
- We calculated the mean as 176.4
- (92-176.4)2 + (135-176.4)2 + (131-176.4)2 + (306-176.4)2 + (218-176.4)2
- 7123.36 + 1713.96 + 16796.16 + 1730.56 = 29425.2
- 29425.2/(5-1) = 29425.2/4 = 7356.3
- Square root of 7356.3 = 85.76
- SEM = 85.76/5^.5 = 38.35

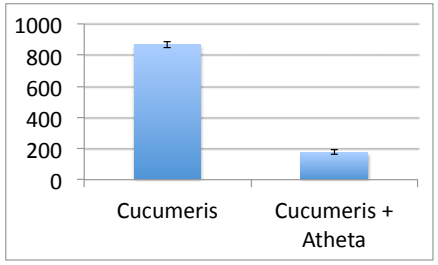
$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2},$$

Graphical Data Analysis

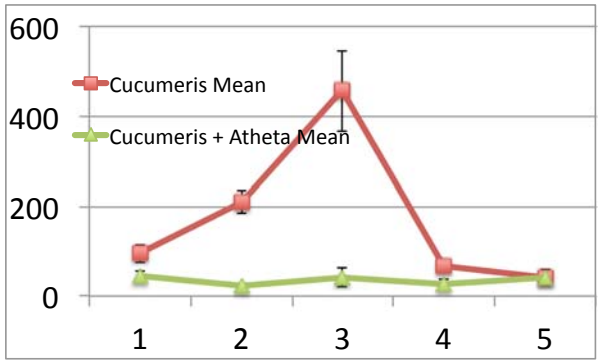
- Visualizing your data is a great step towards understanding it
- Bar charts and line charts for continuous data
- Stacked bar or pie charts for categorical data
- Dependent vs. independent variables and axes

Bar Charts

- A good tool for showing centrality and dispersion across discrete treatments or groups.
- What does this example tell us?

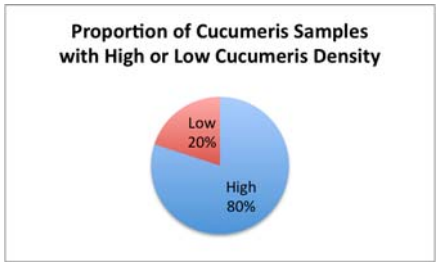


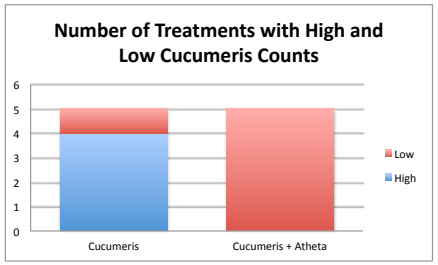
Line Charts



- A good tool fro showing centrality and dispersion over a continuous variable.
- E.g. time, temperature, height, weight, etc.
- What does this example tell us?

Stacked Bar or Pie Charts





- Good for showing the frequency of particular categories
- What do these examples tell us?

Six Steps of Hypothesis Testing

1. State the null hypothesis.
2. State the alternative hypothesis
3. Select a level of significance
4. Collect and summarize the sample data.
5. Refer to a criterion for evaluating the sample evidence.
6. Make a decision to keep/reject the null.

Six Steps of Hypothesis Testing

1. State the null hypothesis.
2. State the alternative hypothesis
3. Sele Alpha = .05 e
4. Collect and summarize the sample data.
5. Refer to a criterion for evaluating the sample evidence.
6. Make a decision to keep/reject the null.

Six Steps of Hypothesis Testing

1. State the null hypothesis.
2. State the alternative hypothesis
3. Select a level of significance
4. Collect and summarize the sample data.
5. Refer to a criterion for evaluating the sample evidence.
6. Make a decision to keep/reject the null.

Continuous vs. Categorical Data

- If we can summarize our data using centrality and dispersion we can use a t test (or related test) to formally test hypotheses
- If our data is categorical we have to use frequency based analyses
- i.e Chi Squared test

T test

- A t-test is any statistical hypothesis test in which the test statistic follows a Student's t distribution if the null hypothesis is supported.
- Assumptions
  - Independence
  - Normal distribution
  - Comparable dispersion (Variance)

T test in Excel

- =TTEST(array1,array2, tails,type)
- We will test the total mite counts for our two treatments
- We are using a 2 sample test and are not assuming equal variance so we will select type 3

Cucumeris Total	Cucumeris + Atheta Total
684	92
555	135
914	131
1076	306
1111	218

=TTEST(J2:J6,K2:K6,2,3)

Array 1

Array 2

Tails

Type

"P" Value → 0.001869615

T Test Interpretation and Reporting

- Excel provides us with a "P" value
- P value = *"The probability of obtaining a test statistic at least as extreme as the one that was actually observed, assuming that the null hypothesis is true."*
- The lower the p value the more likely our populations differ
- When reporting T tests we always provide the Test statistic (t), the Degrees of Freedom (df), and the P value.
- Degrees of Freedom for a two sample T test are equal to 2n-2.
- To get the test statistic in excel we use the TINV() function
- TINV(probability,degrees_freedom)

Cucumeris Total	Cucumeris + Atheta Total
684	92
555	135
914	131
1076	306
1111	218
0.001869615	

=TINV(J8,8)

P value

Degrees of Freedom

Pearson's Chi Squared Test

- Tests a null hypothesis stating that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.
- Useful for categorical data analysis
- Assumptions: Independence, Best when each potential category is >4.
- Does not provide a measure of direction or magnitude
- Null hypothesis is that populations are not associated —are independent

Chi Square in Excel

- CHISQ.TEST(actual_range,expected_range)
- Requires the calculation of "expected values"

Chi Square	Actual		
	High	Low	
Cucumeris	4	1	5
Cucumeris + Atheta	0	5	5
	4	6	

Chi Square	Expected	
	High	Low
Cucumeris	2	3
Cucumeris + Atheta	2	3

Chi Test	=CHISQ.TEST(I21:J22,I27:J28)
Chi Value	0.01974

Actual Values

Predicted Values

$$Fe = \frac{\text{Row Subtotal X Column Subtotal}}{\text{Total}}$$

"P" Value → **0.009823275**

Chi Square Interpretation and Reporting

- When reporting Chi Square tests we always provide the Test statistic (t), the Degrees of Freedom (df), and the P value.
- Degrees of Freedom = (Rows - 1)(Columns - 1) = (2 - 1)(2 - 1) = 1
- To get the test statistic in excel we use the CHISQ.INV() function
- CHISQ.INV (probability,deg_freedom)

Chi Square	Actual		
	High	Low	
Cucumeris	4	1	5
Cucumeris + Atheta	0	5	5
	4	6	

Chi Square	Expected	
	High	Low
Cucumeris	2	3
Cucumeris + Atheta	2	3

Chi Test	0.00982
Chi Value	=CHISQ.INV(I30,1)

P Value

Degrees of Freedom