Goal of the Lecture

To familiarize students with statistical concepts, definitions, and measures necessary to analyze habitat data.

Reading Assignments:


Lecture Structure

I. Definitions & Concepts

II. Statistical Measures
What is Statistics?

The subdiscipline of mathematics that uses mathematical theorems, principles, and techniques to analyze data.

Fundamental Concept:

- Is statistics necessary for a census?
- Sample vs. Census
- Collection of all individuals or possible measurements
- Subset of all individuals or possible measurements
- “Field of Statistics”
- “Point Estimates”

Statistical Definitions

Response Variable: The characteristic of interest measured. E.g., air temperature, g of seed, tree density

Experimental Unit: The natural or artificial entity that is measured. E.g., animal, plot (area), water column

Population: The entire collection of experimental units. E.g., all animals, all potential plots (1-1m²)

Sample: The subset of experimental units which are actually measured. E.g., animals measured, plots measured

Parameter: The unknown true measure of central tendency, variability, or relation. E.g., true mean weight (all animals) µ

Statistic: An estimate of a population parameter calculated from data. E.g., mean weight of measured animals µ

Probability: The chance of occurring. Provides the inferential framework (confidence!) for determining if differences exist in a response variable between ≥ 2 populations

Statistical Notation

For a response variable, x, we denote a sample from the population as:

\[ x_1, x_2, x_3, \ldots, x_n \]

where: \( x_i \) = measurement on EU #1, \( x_i \) = measurement on EU #2,..., \( x_n \) = measurement on the last EU

and, \( n \) = sample size

For more general notation, we let \( i \) = EU #, thus to denote the sum of all measurements in a sample:

\[ \sum_{i=1}^{n} x_i \]

Suppose, we have the following data set:

<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>|x|</th>
</tr>
</thead>
</table>
| 1 | 4 | 4
| 2 | 2 | 2
| 3 | 1 | 3

\[ \sum_{i=1}^{n} x_i = 78 + 74 + 82 + 97 + 95 = 426 \]
Measures of Central Tendency

Where is the center of our data?

Sample Median: The number where exactly 50% of the data lie above and below it.

Odd n:

Even n: (take average)

Sample Mean: The number such that the sum of deviations from each measurement to it \( \bar{x} \) = 0.

Odd n:

Even n:

Computationally:

The Sample Median is less affected by Outliers (compared to the sample mean)

Measures of Proportion

What proportion of the population is ________?

The true population proportion, \( P \), is estimated by calculating relative frequency.

Relative Frequency = Relative Frequency

Total Number of Individuals in Sample

<table>
<thead>
<tr>
<th>Value of x</th>
<th>Freq</th>
<th>( P ) Hat</th>
</tr>
</thead>
<tbody>
<tr>
<td>JUV</td>
<td>5</td>
<td>0.05</td>
</tr>
<tr>
<td>SUBADULTS</td>
<td>10</td>
<td>0.333</td>
</tr>
<tr>
<td>ADULT_F</td>
<td>5</td>
<td>0.185</td>
</tr>
<tr>
<td>ADULT_M</td>
<td>3</td>
<td>0.111</td>
</tr>
</tbody>
</table>

Measures of Variability

Population Mean

True Natural Variation of Individuals in a Population

Sample Standard Deviation, \( S \): Natural Variation of Measurements in Sample

How should we measure variation?

How about using deviations? If deviations are small, our measurements are clustered (precise)?

Now, a reasonable estimate of variation might be “Average Deviations” from the Mean:

But, if we square our deviations (all will be positive), the numerator will not equal zero. “Average Squared Deviations”

Which is not helpful.

Use \( S \) to estimate \( \sigma \)
Measures of Variability

Population Mean

Sample Variance, $S^2$: Average Squared Deviations from the Sample Mean

Sample Standard Deviation, $S$: Average Deviation from the Sample Mean

Sample Range: $\text{MAX} - \text{MIN}$

Calculating Variance & Standard Deviation:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

$$S = \sqrt{S^2}$$

Empirical Guidelines for Symmetric Distributions

If your data (thus the population) follow approximately a symmetric distribution (i.e., a bell-shaped curve), then:

- Approximately, 68% of your data should lie within 1 SD of mean
- Approximately, 95% of your data should lie within 2 SD of mean
- Approximately, 99% of your data should lie within 3 SD of mean

Measures of Variability

Population Proportion

Calculating the sample standard deviation for relative frequencies (i.e., proportions) is much easier.

$$S = \sqrt{\bar{q}_i \bar{p}_i \frac{1}{n}}$$

where, $\bar{q}_i = 1 - \bar{p}_i$

Demographic Data:
**Measure of Relation**

**Correlation**

**Goal:** Measure the linear relation between 2 response variables, X and Y.

**Accomplished by:** Calculating the correlation coefficient, \( r \).

**Interpreting \( r \):**

- Strong Positive: \( r > 0.5 \)
- Strong Negative: \( r < -0.5 \)
- Weak Positive: \( 0.1 < r < 0.5 \)
- Weak Negative: \( -0.5 < r < -0.1 \)
- No Relation: \( -0.1 < r < 0.1 \)

**Are body fat and clutch size \(+\) related?**

**Measure of Relation**

**Correlation Coefficient**

\[
\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}} = \frac{SS_{xy}}{SS_x SS_y} \sqrt{\frac{SS_y - \sum (y_i - \bar{y})^2}{SS_x - \sum (x_i - \bar{x})^2}}
\]

**Example:**

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>X*Y</th>
<th>X*X</th>
<th>Y*Y</th>
<th>X*Xbar</th>
<th>Y*Ybar</th>
<th>X*Ybar</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>-2</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>-2</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>-8</td>
<td>9</td>
<td>16</td>
<td>9</td>
<td>16</td>
<td>-8</td>
</tr>
</tbody>
</table>

\[
\hat{r} = \frac{11}{\sqrt{14}\sqrt{10}} = 0.93
\]