Introduction to Data and Analysis

Wildlife Management is a very quantitative field of study.
Results from studies will be used throughout this course and throughout your career.

Sampling design influences the strength of inference of data.

- Population – a group of organisms occupying specific area at a specific time, the group to which inferences are made.
- Sample – The subset of the population that is measured.
- Parameters – An unknown quantity that characterizes a population.
- Statistics – Derived quantities that describe the sample.
- Inference
- Description
- Estimation
- Parameters
- Statistics
### Scales of Measurement and Types of Data

<table>
<thead>
<tr>
<th>Type of questions we ask determine the type of data needed</th>
</tr>
</thead>
</table>
| Nominal Data  
  - Classification data – m/f  
  - No ordering – m not > f  
  - Arbitrary labels |
| Ordinal Data  
  - Ordered but differences between values not important – political parties given labels 0,1,2; Likert scales etc. |
| Interval Data  
  - Ordered, constant scale  
  - Differences important, ratios do not make sense  
  - Temperature, dates |
Scales of Measurement and Types of Data

- **Ratio Data**
  - Ordered, constant scale, natural zero
  - Height, weight, age, length

- **Discrete** - Only certain specific values are valid, points between these values are not valid. For example, counts of people (only integer values allowed), the grade assigned in a course (F, D, C-, C, C+,...).

- **Continuous** - All values in a certain range are valid. For example, height, weight, length, etc.

Summary Statistics

I. **What are statistics?**

Statistics deals with variation and attempts to draw conclusions from data despite variation.
2 major roles
1. Condense variable information into a summary to convey information (descriptive stats)
2. Assess whether given variability in data are consistent with your hypothesis (inferential stats)

What is the scale of measurement? Ratio

What are some useful statistics to describe the data?

II. Descriptive Statistics
A. Sample specific – Sample size, minimum, maximum value
B. Location – where on a scale do the data fall
1. Mean – the average of a sample
   \[
   \bar{x} = \frac{\sum x}{n}
   \]
   Advantage – simple to compute and interpret
   Disadvantage – heavily influenced by extremes
   If data are skewed then not good measure

<table>
<thead>
<tr>
<th>Trt</th>
<th>Mass</th>
<th>Hem</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>67.6</td>
<td>46.13</td>
</tr>
<tr>
<td>F</td>
<td>71.23</td>
<td>44.23</td>
</tr>
<tr>
<td>F</td>
<td>70.7</td>
<td>46.1</td>
</tr>
<tr>
<td>F</td>
<td>73.6</td>
<td>47.2</td>
</tr>
<tr>
<td>F</td>
<td>76.78</td>
<td>42.53</td>
</tr>
<tr>
<td>F</td>
<td>67.28</td>
<td>39.9</td>
</tr>
<tr>
<td>F</td>
<td>68.6</td>
<td>41.3</td>
</tr>
<tr>
<td>F</td>
<td>68.16</td>
<td>39.48</td>
</tr>
<tr>
<td>F</td>
<td>71.95</td>
<td>46.33</td>
</tr>
<tr>
<td>F</td>
<td>70.66</td>
<td>42.1</td>
</tr>
<tr>
<td>F</td>
<td>63.68</td>
<td>52.9</td>
</tr>
<tr>
<td>F</td>
<td>76</td>
<td>41.33</td>
</tr>
<tr>
<td>F</td>
<td>70.73</td>
<td>43</td>
</tr>
<tr>
<td>F</td>
<td>66.89</td>
<td>41.5</td>
</tr>
<tr>
<td>F</td>
<td>70.08</td>
<td>41.47</td>
</tr>
<tr>
<td>C</td>
<td>59.34</td>
<td>36.2</td>
</tr>
<tr>
<td>C</td>
<td>70.74</td>
<td>36.92</td>
</tr>
<tr>
<td>C</td>
<td>72.54</td>
<td>38.96</td>
</tr>
<tr>
<td>C</td>
<td>66.7</td>
<td>45.55</td>
</tr>
<tr>
<td>C</td>
<td>67.5</td>
<td>42.45</td>
</tr>
<tr>
<td>C</td>
<td>65.23</td>
<td>34.07</td>
</tr>
<tr>
<td>C</td>
<td>70.3</td>
<td>43.6</td>
</tr>
<tr>
<td>C</td>
<td>69.75</td>
<td>33.95</td>
</tr>
<tr>
<td>C</td>
<td>64.48</td>
<td>32.76</td>
</tr>
<tr>
<td>C</td>
<td>59.35</td>
<td>36.6</td>
</tr>
<tr>
<td>C</td>
<td>91.1</td>
<td>45.01</td>
</tr>
<tr>
<td>C</td>
<td>70.53</td>
<td>43.9</td>
</tr>
<tr>
<td>C</td>
<td>60.58</td>
<td>35.78</td>
</tr>
<tr>
<td>C</td>
<td>97.37</td>
<td>36.57</td>
</tr>
<tr>
<td>C</td>
<td>99.7</td>
<td>40.88</td>
</tr>
<tr>
<td>C</td>
<td>73.18</td>
<td>34.52</td>
</tr>
</tbody>
</table>
2. Median – middle value, 50% less than and 50% more than
   Rank data from smallest to largest – median is rank \( n + 1/2 \)
   Odd
   14 17 18 20 21
   Even
   14 17 18 20

3. Mode – most frequent value, commonest
   Used very infrequently, mostly by Ornithologists
   6
   5
   2
   6
   7
   6
   3
   6
   7
   5
C. Dispersion

Spread of data around a central location

1. Range – difference between max. and min., very sensitive to extreme values (same units as original data)

2. Percentiles – description of sample distribution
   Data put in ascending order
   28, 32, 34, 34, 36, 38, 41, 42, 44, 45, 50, 51, 52
   The pth percentile has at least p% of the values above that point and 100-p% below.
   - 25th Percentile = 0.25(13) = 3.25 = 4th = 34
   - 85th Percentile = 0.85(13) = 11.05 = 12th = 51

3. Standard Deviation – measure of mean deviation of observations from the mean of the distribution (same units as original data) (mean distance from the mean)

4. Variance – quantifies how far each observation is from mean. No units associated with variance. Average of the squared deviations
   Important measure in statistics.
4. Standard Error – often used synonymously with standard deviation, standard deviation of mean.

5. Coefficient of Variation (CV) – std dev expressed as % of mean.

When populations differ (considerably) in means direct comparisons of variance or std deviations not useful.

e.g. larger organisms vary more in size than smaller ones (std dev of elephant tails will be greater than std dev of mouse tails)

CV - compares relative amounts of variation in populations with different means.

---

Sampling

Wolf Data

---

Sampling

- There are 2 basic principles of sampling
Sampling

- **Replication** - number of random, independent experimental units drawn from the research population.
  - Provides estimate of experimental error (provides several observations on experimental units receiving the same treatment)
  - Increases precision of experiment by reducing standard errors.

- **Randomization** – independence of data, treatments are assigned to experimental units in such a way that any unit is equally likely to receive any treatment.

Goal for sampling
- Make inference about population characteristics from our sample
Goal for sampling
- Sample is not a complete enumeration and thus uncertainty associated with estimates

Steps for designing sampling scheme
1. Identify population of interest = Target Population
2. How much of target population can be sampled?
   - The portion of the target population that could be included in our sampling is the sampling frame (spatial, temporal)

Steps for designing sampling scheme
Do you think you can make inferences about the wolf harvest in AK (target population) from the haphazard sample you collected?

Simple Random Sample
- A sample drawn from the population in such a way that every item (individual) has the same probability of being included.
  - If the population size is N then each individual has a 1/N probability of being included in the sample
  - Every possible combination of n sampling units is equally likely to be selected
Simple Random Sample

- Systematic sources of bias are not included (on average) leading to accurate estimates of population parameters.
- How do we obtain these?
  - Lottery
  - Random Number Table
  - Random Number Generator (excel)

Suppose we wanted to estimate the number of snakes on Antelope Island
Can we count them all?

One way to obtain a random sample is to divide the island into a grid and randomly select a sample of these cells.

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
</tr>
</tbody>
</table>

I used a random number generator and had it select 7 cells

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>16</td>
<td>17</td>
<td>18</td>
<td>19</td>
<td>20</td>
<td>21</td>
</tr>
<tr>
<td>22</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>27</td>
<td>28</td>
</tr>
<tr>
<td>29</td>
<td>30</td>
<td>31</td>
<td>32</td>
<td>33</td>
<td>34</td>
<td>35</td>
</tr>
</tbody>
</table>
If I could count all the snakes in each of these randomly selected cells I could estimate the abundance ($A$) of snakes on the island.

$$A = N\bar{X}$$

Where $N$ = total number of cells, and $\bar{X}$ = mean number of snakes in cells sampled.

### Assignment for next week’s lab

### Bias, Accuracy and Precision

How can we judge the “quality” of the estimates you just calculated?

There are 3 main criteria that we use to judge the quality of estimators and estimates.

- Precision – a measure of the closeness to each other of repeated measurements of the same quantity (measures = standard error, confidence intervals)

- Degree of reproducibility
Bias, Accuracy and Precision

- **Precision** – Larger variation in the population leads to lower precision of an estimate, whereas a larger sample size leads to higher precision in the estimators.

- **Bias** – describes how far the average value of the estimator is from the true population value. An unbiased estimator centers on the true value for the population.

- **Accuracy** – ultimate measure of a quality of an estimate. Refers to the small size of deviations of the estimator from the true population value. Degree of veracity.

If an estimate is both unbiased and precise then it is said to be accurate (measure = estimator with a small mean-squared error).
Unbiased and precise = accurate

Unbiased but not precise = not accurate

Biased but precise = not accurate

Biased and not precise = not accurate
Recall our discussion of dispersion -

Population Variance ($\sigma^2$) is

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

Obviously we would have to sample all individuals in the population to know this parameter.

Spread of a population is more typically characterized by the

Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$

Standard deviation is expressed in same units as quantity measured.

Population Standard Deviation gets larger as the spread among individuals increases.

We rarely are able to measure the entire population, so we need sample the population and estimate the population variance and standard deviation.

Estimators for Population Variance ($s^2$) and Population Standard Deviation ($s$) look similar.
Estimate of Population Variance ($s^2$) is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Estimate of Population Standard Deviation

$$s = \sqrt{s^2} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

Standard deviation is expressed in same units as quantity measured.

Remember the estimate of the population mean (estimate of population standard deviation) is based on our sample.

There is uncertainty associated with the estimate.

This uncertainty is called precision and is estimated with sampling standard error.

Standard error (SE) is not an estimate of a population parameter!

SE measures uncertainty with a population parameter (e.g. SE of Mean).
SE is estimated as

\[ SE = \frac{s}{\sqrt{n}} = \sqrt{\frac{s^2}{n}} \]

III. Inferential Statistics
A. Hypothesis testing

Goal is to determine if 2 samples differ (were samples drawn from same population).

- 2 independent samples drawn from same population
- Calculated means estimated from same population
- Differences result of chance / sampling error

Null hypothesis – means of 2 populations are equal
Alternative hypothesis – means of 2 populations are different.

Test allows us to say with some level of certainty (probability) if we can reject the null and accept the alternative.
Need to determine what magnitude of error we are willing to live with.

<table>
<thead>
<tr>
<th>Null hypothesis</th>
<th>Accepted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct</td>
<td>Type I error</td>
</tr>
<tr>
<td>False</td>
<td>Type II error</td>
<td>Correct</td>
</tr>
</tbody>
</table>

Level of probability traditionally chosen to be 0.05
You are willing to take a 5% chance of rejecting the null when it is in fact true.

Type I error = $\alpha$

Tests calculate $p$ value for you, but you must determine $\alpha$ BEFORE the experiment.

B. T – test

Ho – $\bar{x}_1 = \bar{x}_2$; means drawn from same population

Ha – $\bar{x}_1 \neq \bar{x}_2$; means differ

Traditional test to determine if means from two samples are different from one another.

1. Assumptions of test
   - Individuals sampled randomly from population
   - Variances from each sample are not significantly different
   - Data are normally distributed

2. Procedures
   - SPSS
A. Definitions

- Data – information pertinent to answering some question
- Population – group to which you are trying to generalize
  - Observational (wing lengths of House Flies)
  - Experimental (wing lengths of males on standard diet)
- Samples – the proportion of population that is measured

Definitions (cont.)

- Experimental Unit – the “thing” that is measured; the smallest unit that is independent of other units and to which we can randomly assign a treatment.
- Random Sample – sample drawn so that all members of a population have equal and independent chance of being included in sample.

<table>
<thead>
<tr>
<th>X</th>
<th>f</th>
<th>x = X - ( \bar{x} )</th>
<th>( x^2 )</th>
<th>( fx^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.4</td>
<td>2</td>
<td>-0.6</td>
<td>0.36</td>
<td>0.72</td>
</tr>
<tr>
<td>3.7</td>
<td>8</td>
<td>-0.3</td>
<td>0.09</td>
<td>0.72</td>
</tr>
<tr>
<td>4.0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4.3</td>
<td>8</td>
<td>0.3</td>
<td>0.09</td>
<td>0.72</td>
</tr>
<tr>
<td>4.6</td>
<td>2</td>
<td>0.6</td>
<td>0.36</td>
<td>0.72</td>
</tr>
</tbody>
</table>

\( \bar{x} = 4.0 \)  \( n = 25 \)  \( \sum 2.88 \)

\[ s^2 = \frac{\sum fx^2}{n} = \frac{2.88}{25} = 0.115 \text{ (average of squared deviations)} \]

\[ s = \sqrt{0.115} = 0.3394 \]