29 Sept. 15

This Time: Intro, types of variables: samples & populations

Next Time: numerical & graphical descriptive methods

### Disease? (Dichotomous)

<table>
<thead>
<tr>
<th>y</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>0</td>
<td>N</td>
</tr>
</tbody>
</table>

Sample size $N = \frac{4}{93} = 4.3\%$

1 row for each deer

1 column

$N = 93$

### General Data Set

1 row for each subject

1 column for each variable (thing we measure)

Average $= \frac{\sum \frac{y}{n}}{\frac{1}{n}} = \text{mean} = \% \text{ of the deer with disease}$

With subscript notation

### Sample Mean $\bar{y}$

- $\bar{y} = \frac{y_1 + y_2 + \ldots + y_n}{N}$
- $N = 93$

### Calculations

- $\bar{y} = \frac{\sum y_i}{n}$
- $\frac{1}{N} \sum y_i$

### Mean $y_i$

- $\frac{1}{N} \left( y_1 + y_2 + \ldots + y_n \right)$

- $\sum_{i=1}^{n} y_i$

- $\sum_{i=1}^{n} (y_i - \bar{y})$

- $\sum_{i=1}^{n} y_i$

### Mean $\bar{y}$

- $\bar{y} = \frac{1}{N} \sum_{i=1}^{n} y_i$

- $\sum_{i=1}^{n} (y_i + y_2 + \ldots + y_n)$
Population

all ucsc deer as of 31 Dec. 2006

\[
\begin{bmatrix}
1.5 \\
& 0.5
\end{bmatrix}
\]

\(N = 800\)

\(\text{population size}\)

\(\text{mean } = \sigma \text{ is a parameter}\)

\(\text{unknown}\)

\(\text{A pop. numerical summary}\)

\(\text{like at random}\)

Sample

The observed deer

\[
\begin{bmatrix}
disease? \\
0 \\
1 \\
0 \\
\vdots \\
1 \\
0
\end{bmatrix}
\]

\(n = 93\)

\(\text{sample size}\)

\(1 = \text{yes}\)

\(0 = \text{no}\)

\(\text{row for each deer}\)

\(\frac{\hat{p}}{n} = 4.3\%\)

\(\text{sample mean}\)

\(\text{is about (approx.)}\)

\(\text{equal to}\)

\(\overline{y} \text{ is a good estimate of } \sigma\)

\(\overline{y} = \hat{\sigma} \leftrightarrow \text{"theta hat" sample estimate of } \sigma\)

basic sampling principle = try to make sampled & unsampled subjects in pop.

as similar as possible in all relevant ways.

How do this? Choose sample at random from pop.
At random with replacement (put first draw back in):
inddependent identically distributed (IID) sampling
has easier math

At random without replacement:
simple random sampling (SRS)
more informative than IID

<table>
<thead>
<tr>
<th>Variable</th>
<th>Possible Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>eye color</td>
<td>blue (1), brown (0)</td>
</tr>
<tr>
<td>hair color</td>
<td>black, brown, red, white</td>
</tr>
<tr>
<td>plant</td>
<td>height (cm) 7.42, 8.20,...</td>
</tr>
<tr>
<td></td>
<td>leaves</td>
</tr>
</tbody>
</table>