Statistics and Quantitative Biology

Tuesday, September 29, 2009

Data exploration and representation

Informal data analysis

```r
x1<-rnorm(30,0.8,0.1)
x2<-rnorm(30,0.7,0.1)
x3<-rnorm(30,0.8,0.2)
```
Percentiles and quantiles

- Median
- Interquartile range
- Whiskers
- Outlier

Percentiles and quantiles

- Median
- Interquartile range
- Whiskers
- Outlier

Percentiles and quantiles

- Median
- Interquartile range
- Whiskers
- Outlier

Percentiles and quantiles

- Median
- Interquartile range
- Whiskers
- Outlier
Trichodina, a common parasite of fish, infects the skin, gills, and sometimes the eyes. The infection can cause deformities, reduced growth, and even death. However, some fish species are more resistant to infection than others. The following table shows the observed frequencies of infection in different fish species. The table reports the number of fish that were infected or not infected.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number Infected</th>
<th>Number Not Infected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species A</td>
<td>30</td>
<td>70</td>
</tr>
<tr>
<td>Species B</td>
<td>40</td>
<td>60</td>
</tr>
<tr>
<td>Species C</td>
<td>50</td>
<td>50</td>
</tr>
</tbody>
</table>

Mosaic plot

Table: Enzyme concentration (mg/ml)

<table>
<thead>
<tr>
<th>Condition</th>
<th>Enzyme Concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Long days</td>
<td>1.525, 6.625</td>
</tr>
<tr>
<td>Long days</td>
<td>1.555, 9.9</td>
</tr>
<tr>
<td>Short days</td>
<td>1.385, 12.5</td>
</tr>
<tr>
<td>Short days</td>
<td>1.485, 11.625</td>
</tr>
<tr>
<td>Short days</td>
<td>1.255, 18.275</td>
</tr>
<tr>
<td>Short days</td>
<td>1.285, 13.225</td>
</tr>
</tbody>
</table>

From: Cobb. Introduction to design and analysis of experiments. Springer
A Down syndrome, the most common genetic birth defect associated with mental retardation, occurs equally across all races and levels of society. The effects of the disorder on physical and mental development are severe and are expressed throughout the life span. The individual's family is also affected emotionally, economically, and socially.


Doesn't it look suspicious?
A Table about 400 meters tall?
Describing data:
Proportion, location and spread

Measurements of location and spread

Mean
Median
Mode

Variance
Standard deviation
Standard error
Interquartile/interquartile range

Estimating mean from sample
Estimating the mean from the sample’s histogram
Quick Formulas

Mean
\[ \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \]

Variance
\[ s^2 = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1} \]

Standard deviation
\[ s = \sqrt{s^2} \]

Standard error of the mean
\[ \sigma_p = \frac{\sigma}{\sqrt{n}}, \quad SE_p = \frac{s}{\sqrt{n}} \]

Proportion
\[ \hat{p} = \frac{X}{n} \]

Mean(\(\hat{p}\)) = p
\[ \sigma(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} \]

We never have access to
the true probability distribution of population
nor
the true sampling probability distribution
We can only get estimates?
What is a good estimator?
What are the good estimators for the mean, the standard deviation, or for the proportion?
Rewind...

The central limit theorem

Let $x_1, x_2, x_3, \ldots, x_n$ be identical and independently distributed random variables with mean $\mu$ and variance $\sigma^2$.

Let $\Sigma_n$ be the sum of values of the $n$ variables.

When $n$ is very large $\Sigma_n$ follows approximately a Normal distribution with mean $n\mu$ and variance $n\sigma^2$. 

$\frac{n \cdot \mu}{\sqrt{n} \cdot \sigma}$
Histogram of 100000 dice draws

Histogram of 1000 means of 10 draws

\[ \mu = 3.5 \]
\[ \sigma = 1.87 \]

\[ \mu = \frac{3.5}{10} \]
\[ \sigma = \frac{1.87}{10} \]

\[ n = 100000 \]

\[ \text{Mean}(Y) = \frac{3.5}{\sqrt{10}} \]
\[ \text{SE}_Y = 0.54 \]
\[ n = 10 \]

\[ X \sim \text{Normal}(\mu, \sigma) \]

Compare \( Y \) and \( \mu \)
Compare \( s \) and \( \sigma \)