Calculating Descriptive Statistics

Statistics 1403

Spring 2016

1 Calculating the Basic Moment Statistics

univariate

\[ y = \frac{1}{n} \sum_{i=1}^{n} y_i \]  \[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

...average...

...deviation...

bivariate

\[ r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \]
2 Sufficient Statistics to the Rescue!

Consider this small sample of APGAR scores for newborn infants:

\[ x = \{9, 7, 2, 8, 7, 9\} \]

The sample mean is simple to calculate

\[ \bar{x} = \frac{9 + 7 + 2 + 8 + 7 + 9}{6} = 7 \]

However, the sample variance has a more tedious formula

\[ s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

This can be rearranged for calculation using the sufficient statistics \( n, \sum_i x_i, \) and \( \sum_i x_i^2, \) so

\[ s_x^2 = \frac{1}{n-1} \left[ \sum_{i=1}^{n} x_i^2 - \left( \sum_{i=1}^{n} x_i \right)^2 / n \right] \]

Then work in two steps

1. find the sufficient statistics

\[ n = 6 \quad \sum_i x_i = 9 + 7 + 2 + 8 + 7 + 9 = 42 \quad \sum_i x_i^2 = 9^2 + 7^2 + 2^2 + 8^2 + 7^2 + 9^2 = 328 \]

2. then find the sample mean and variance
2. then find the sample mean and variance

\[ \bar{x} = \frac{42}{6} = 7 \quad s_x^2 = \frac{328 - 42^2/6}{6 - 1} = 6.8 \]

\[ s_x = \sqrt{6.8} \]

3. Your Turn

Now consider the 3-month APGAR scores for the same 6 infants

\[ y = \{8, 4, 3, 7, 8, 9\} \]

1. find the sufficient statistics

\[ n = 6 \quad \sum_i y_i = 39 \quad \sum_i y_i^2 = 283 \]

2. then find the sample mean and variance

\[ \bar{y} = \frac{39}{6} = 6.5 \]

\[ s_y^2 = \frac{(283 - \frac{39^2}{6})}{(6-1)} \]

\[ = 5.9 \]
4 On to Correlation

Correlation can be done in the same way, adding just one more sufficient statistic, $\sum x_i y_i$.

$n = 6$

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>9</th>
<th>7</th>
<th>2</th>
<th>8</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>8</td>
<td>4</td>
<td>3</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>$x_i y_i$</td>
<td>72</td>
<td>28</td>
<td>6</td>
<td>56</td>
<td>56</td>
<td>81</td>
</tr>
</tbody>
</table>

Then use the sufficient statistic version of the correlation formula

$$r_{xy} = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sqrt{[\sum x_i^2 - (\sum x_i)^2/n][\sum y_i^2 - (\sum y_i)^2/n]}}$$

$$\Sigma x_i = 42$$
$$\Sigma x_i^2 = 328$$
$$\Sigma y_i = 35$$
$$\Sigma y_i^2 = 283$$

$$\frac{249 - \frac{42 \times 35}{6}}{\sqrt{328 - \frac{42^2}{6}}}$$
\[ \frac{\Sigma y_i}{\Sigma y_i^2} = 283 \]
\[ \Sigma x_i y_i = 2.99 \]

\[ -1 \leq \frac{\Sigma x_i y_i}{\Sigma x_i^2} \leq +1 \]

5  Doing it with R

```r
> x <- c(9, 7, 2, 8, 7, 9)
> y <- c(8, 4, 3, 7, 8, 9)
> mean(x)
> var(x)
> mean(y)
> var(y)
> cor(x, y)
```
> mean(y)
> var(y)
> cor(x, y)