# Statistics
# Homework 1: Descriptive Statistics

Javier Tasso

1. Solve the following exercises of chapter 1 of Devore, Berk, Carlton's *Modern Mathematical Statistics with Applications* (third edition): 7, 8, 21, 23, 25, 29, 33, 40, 41, 45, 49, 53, 55, 61, 65, 67, 69, 77, 89.

2. Consider a sample $x_1, x_2, \ldots, x_n$ from variable $x$ and construct the variable $y$ as follows: $y_i = ax_i + b$ (where $a \neq 0$).

   (a) Show that $\bar{y} = a\bar{x} + b$.

   (b) Show that $S_y^2 = a^2 \cdot S_x^2$

3. What's the value of $c$ that minimizes $\sum_{i=1}^{n}(x_i - c)^2$?

4. Show that the sample variance $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$ is equivalent to the shortcut formula $S_x^2 = \frac{(\sum_{i=1}^{n} x_i^2) - n\bar{x}^2}{n-1}$.

5. Show that $\mathrm{COV}(x, x) = S^2$. The sample covariance of a variable with itself is equal to the sample variance.

6. Show that the sample covariance $\mathrm{COV}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$ has also a shortcut formula given by $\mathrm{COV}(x, y) = \frac{(\sum_{i=1}^{n} x_i y_i) - n\bar{x}\bar{y}}{n-1}$

7. Consider the usual estimates for the slope and intercept of the regression line $\hat{\beta}_1 = \frac{S_{xy}}{S_x^2}$ and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$.

   (a) Argue that is previously demeaned (and in that case $\bar{y} = 0$ and $\bar{x} = 0$) then $\hat{\beta}_0 = 0$.

   (b) Continue assuming that $\bar{y} = 0$ and $\bar{x} = 0$. Find the value of $b$ that minimizes $\sum_{i=1}^{n}(y_i - bx_i)^2$.

8. Bivariate data often arises from the use of two different techniques to measure the same quantity. As an example, the accompanying observations on x = hydrogen concentration (ppm) using a gas chromatography method and y = concentration using a new sensor method were read from a graph in the article *A New Method to Measure the Diffusible Hydrogen Content in Steel Weldments Using a Polymer Electrolyte-Based Hydrogen Sensor* (Welding Res., July 1997: 251s–256s).

   (a) Construct a scatterplot.

   (b) Calculate the correlation coefficient.

   (c) Calculate the slope an intercept of the regression line.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| x | 47 | 62 | 65 | 70 | 70 | 78 | 95 | 100 | 114 | 118 |
| y | 38 | 62 | 53 | 67 | 84 | 79 | 93 | 106 | 117 | 116 |
| x | 124 | 127 | 140 | 140 | 140 | 150 | 152 | 164 | 198 | 221 |
| y | 127 | 114 | 134 | 139 | 142 | 170 | 149 | 154 | 200 | 215 |

**Answers**

1. You can find answers to the odd numbered problems at the end of the textbook.

   8. Categorical: Gender, educational level. Numerical: Age, income, WTP, WTP for the second wine.

   40. a. Mean: 474.4, Median: 507.5
       b. Mean: 484.4, Median: No change
       c. Mean: 554.375, Median: 525
       d. 494.17
       e. 473.125

2. (a) Three steps: sum all the $y_i = ax_i + b$ both sides, apply properties of the summation, and divide by $n$.

$$
\begin{aligned}
y_i &= ax_i + b \\
\sum_{i=1}^{n} y_i &= \sum_{i=1}^{n} (ax_i + b) \\
\sum_{i=1}^{n} y_i &= a\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} b \\
\sum_{i=1}^{n} y_i &= a\sum_{i=1}^{n} x_i + nb \\
\frac{\sum_{i=1}^{n} y_i}{n} &= \frac{a\sum_{i=1}^{n} x_i + nb}{n} \\
\bar{y} &= a\bar{x} + b
\end{aligned}
$$

   (b) Subtract $\bar{y}$ from $y_i$, square that deviation, sum, and divide by $n-1$.

$$
\begin{aligned}
y_i - \bar{y} &= a(x_i - \bar{x}) \\
(y_i - \bar{y})^2 &= a^2(x_i - \bar{x})^2 \\
\sum_{i=1}^{n}(y_i - \bar{y})^2 &= a^2\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} &= \frac{a^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \\
S_y^2 &= a^2 S_x^2
\end{aligned}
$$

3. Differentiate wrt to $c$ and set equal to 0.

$$\sum_{i=1}^{n} -2(x_i - c) = 0$$
$$\sum_{i=1}^{n} (x_i - c) = 0$$
$$\sum_{i=1}^{n} x_i - nc = 0$$
$$c^* = \frac{\sum_{i=1}^{n} x_i}{n}$$
$$c^* = \bar{x}$$

4. Expand and use the fact that $\sum_{i=1}^{n} x_i = n\bar{x}$.

$$(n-1)S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$
$$(n-1)S^2 = \sum_{i=1}^{n} (x_i^2 + \bar{x}^2 - 2\bar{x}x_i)$$
$$(n-1)S^2 = \sum_{i=1}^{n} (x_i^2) + n\bar{x}^2 - 2\bar{x}\sum_{i=1}^{n} (x_i)$$
$$(n-1)S^2 = \sum_{i=1}^{n} (x_i^2) + n\bar{x}^2 - 2\bar{x}^2$$
$$(n-1)S^2 = \sum_{i=1}^{n} (x_i^2) - n\bar{x}^2$$
$$S^2 = \frac{\sum_{i=1}^{n} (x_i^2) - n\bar{x}^2}{n-1}$$

5. Write down the definition of the covariance and replace $y$ by $x$.

$$\mathrm{COV}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

If you replace $y_i - \bar{y}$ by $x_i - \bar{x}$ you get the variance.

6. Expand and follow similar steps as in question 4.

$$
\begin{aligned}
(n-1)\mathrm{COV}(x,y) &= \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\
(n-1)\mathrm{COV}(x,y) &= \sum_{i=1}^{n}(x_iy_i+\bar{x}\bar{y}-x_i\bar{y}-\bar{x}y_i) \\
(n-1)\mathrm{COV}(x,y) &= \sum_{i=1}^{n}(x_iy_i)+n\bar{x}\bar{y}-\bar{y}\sum_{i=1}^{n}(x_i)-\bar{x}\sum_{i=1}^{n}(y_i) \\
(n-1)\mathrm{COV}(x,y) &= \sum_{i=1}^{n}(x_iy_i)+n\bar{x}\bar{y}-n\bar{y}\bar{x}-n\bar{x}\bar{y} \\
(n-1)\mathrm{COV}(x,y) &= \sum_{i=1}^{n}(x_iy_i)-n\bar{x}\bar{y} \\
\mathrm{COV}(x,y) &= \frac{\sum_{i=1}^{n}(x_iy_i)-n\bar{x}\bar{y}}{n-1}
\end{aligned}
$$

7. (a) Simply plug $\bar{y}=0$ and $\bar{x}=0$ in the formula.

   (b) Take derivatives, set equal to 0 and isolate $b$.

$$
\begin{aligned}
\sum_{i=1}^{n}(-2)(y_i-bx_i)x_i &= 0 \\
\sum_{i=1}^{n}(y_i-bx_i)x_i &= 0 \\
\sum_{i=1}^{n}(x_iy_i-bx_i^2) &= 0 \\
\sum_{i=1}^{n}x_iy_i-b\sum_{i=1}^{n}x_i^2 &= 0 \\
\sum_{i=1}^{n}x_iy_i &= b\sum_{i=1}^{n}x_i^2 \\
b^* &= \frac{\sum_{i=1}^{n}x_iy_i}{\sum_{i=1}^{n}x_i^2} = \frac{S_{xy}}{S_x^2}
\end{aligned}
$$

   This is our usual $\hat{\beta}_1$.

8. (a) You are in charge of the scatterplot.

   (b) $r=0.9852$

   (c) $\hat{\beta}_0=-0.9625$ and $\hat{\beta}_1=1.0014$