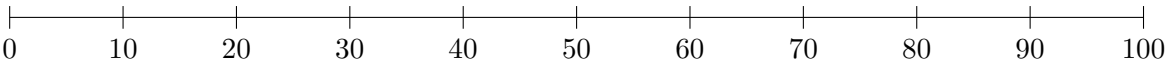
Boxplots (cont.) and Transformations – Solutions
STAT-UB.0103 – Statistics for Business Control and Regression Models

Boxplots

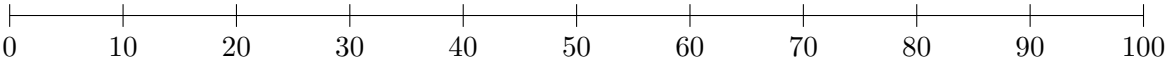
1. Here are the 35 reported expected starting salaries for the male survey respondents (in $1K per year). Make a boxplot of the data.

50, 50, 50, 50, 60, 60, 60, 60, 60, 60, 60, 60, 60, 62.4 65, 65, 65, 70, 70, 70, 75, 76, 80, 80, 80, 80, 80, 80, 80, 85, 90, 90,100,250,300



2. Here are the 18 reported expected starting salaries for the female survey respondents. Make a boxplot of the data.

40, 45, 54, 60, 60, 60, 60, 60, 65, 67, 70, 70, 70, 70, 80, 80, 85, 100



Scaling Values

3. In the online class survey, 61 students report their social media usage in a typical week, in hours. The mean and sample standard deviation of the reported values are:

$$\bar{x} = 10.6 \text{ hours}$$
$$s = 10 \text{ hours}.$$

If we were to convert the reported social media usages from hours to minutes, what would be the new mean and sample standard deviation?

Solution: We convert from minutes to hours by multiplying by 60. Thus, the new mean and sample standard deviation would be

$$\text{mean} = 60 \times 10.6 = 636 \text{ minutes}$$

and

$$\text{std. dev.} = |60| \times 10 = 600 \text{ minutes}.$$

4. Here is dataset X:

$$103.0,\ 98.0,\ 102.0,\ 101.0,\ 102.5,\ 110.0,\ 101.5,\ 100.0,\ 108.0,\ 98.0.$$

The mean and standard deviation of this dataset are

$$\bar{x} = 102.4$$
$$s_X = 3.9.$$

Suppose we construct another dataset, Y, by multiplying every item in X by 5:

$$515.0,\ 490.0,\ 510.0,\ 505.0,\ 512.5,\ 550.0,\ 507.5,\ 500.0,\ 540.0,\ 490.0.$$

That is, $y_i = 5x_i$.

(a) What is the mean of dataset Y?

Solution: We have that $y_i = 5x_i$. Thus,

$$\bar{y} = 5\bar{x} = 5 \cdot 102.4 = 512.$$

(b) What is the sample standard deviation of dataset Y?

Solution:

$$s_Y = |5| \cdot s_X = 5 \cdot 3.9 = 19.5.$$

Shifting Values

5. Students filled out the online class survey between 17:18:28 ET on September 3 and 00:23:16 ET on September 4. The mean and standard deviation of the timestamps were

$$\bar{x} = \text{ 17:18:28 ET on September 3,}$$
$$s = 4.3 \text{ hours}$$

If we convert the times to Pacific Time (PT) by subtracting 3 hours from each value, what will be the mean and sample standard deviation?

Solution: The mean will be shifted by -3 hours: 14:18:28 PT on September 3; the standard deviation will be unchanged: 4.3 hours.

6. Consider a dataset X with $n = 10$ items:

$$3.0,\ -2.0,\ 2.0,\ 1.0,\ 2.5,\ 10.0,\ 1.5,\ 0.0,\ 8.0,\ -2.0.$$

The mean and sample standard deviation of dataset X are

$$\bar{x} = 2.4,$$
$$s_X = 3.9.$$

Suppose we construct a new dataset, Y, by adding 100 to every item in X:

$$103.0,\ 98.0,\ 102.0,\ 101.0,\ 102.5,\ 110.0,\ 101.5,\ 100.0,\ 108.0,\ 98.0.$$

That is, $y_i = x_i + 100$.

(a) What is the mean of dataset Y?

Solution: Shifted by 100: $\bar{y} = 2.4 + 100 = 102.4$.

(b) What is the sample standard deviation of dataset Y?

Solution: Unchanged: $s_Y = 3.9$.

Affine Transformations

7. You have a dataset with $n = 500$ values: $x_1, x_2, \ldots, x_{500}$. The mean value is $\bar{x} = 25$ and the sample standard deviation is $s_X = 4$. You construct a new dataset $y_1, y_2, \ldots, y_{500}$, where

$$y_i = 3x_i + 7.$$

(a) What is the mean of the new dataset?

Solution:

$$\bar{y} = 3\bar{x} + 7 = 3 \cdot 25 + 7 = 82.$$

(b) What is the sample standard deviation of the new dataset?

Solution:

$$s_Y = |3| \cdot s_X = 3 \cdot 4 = 12.$$

8. Consider again the dataset from question 7, consisting of $x_1, x_2, \ldots, x_{500}$ with $\bar{x} = 25$ and $s_X = 4$. You construct a new dataset $z_1, z_2, \ldots, z_{500}$, where

$$z_i = \frac{x_i - \bar{x}}{s_X} = \frac{x_i - 25}{4}.$$

What are the mean and the sample standard deviation of the new dataset?
Hint: $z_i = \frac{1}{4}x_i - \frac{25}{4}$.

Solution:

$$\bar{z} = \tfrac{1}{4}\bar{x} - \tfrac{25}{4} = 0,$$
$$s_Z = \left|\tfrac{1}{4}\right| \cdot s_X = 1.$$

General Transformations

9. Consider the dataset $x_1, x_2, \ldots, x_{25}$ with mean $\bar{x} = 3.2$, median $M = 3$, sample standard deviation $s = 1$, and inter-quartile range IQR $= 2$. Suppose you construct a new dataset $w_1, w_2, \ldots w_{25}$, where

$$w_i = \log x_i$$

(assume that all x_i values are positive, so w_i is well-defined).

Which of the following can you compute for the w_i values using only the information provided in the problem: mean, median, sample standard devation, inter-quartile range?

Solution: The only quantity we can reliably compute is the median. Taking logarithms preserves the order of the values, so the median of the w_i values is equal to $\log M = \log 3$.