BISC 305: Statistics for the Biological Sciences
Discussion #2
TA: Saket Choudhary (skchoudh@usc.edu)
9^{th} September, 2019
Topics covered: Description of Samples and Populations. Chapter 2, 40-59 Description of Samples and Populations. Chapter 2, 59-67

1. Prof. Peter has been given weights of your class mates in pounds (lbs) while your TA has them in kilograms (kgs). They need to make sure that the entries are almost similar. Reporting the entire five point summary will require them to match up 5 numbers, so they just rely on matching just the measure of dispersion. What measure of dispersion would you suggest them to use? Keep in mind, converting any units to other also involves a lot of work.

2. Trojans' score for the last season are as follows: 43, 3, 14, 39, 24, 31, 28, 35, 38, 14, 27, 17 Draw a boxplot manually and verify using R.

Review Sheet

1. **Median ($\tilde{y}$):** The value that most nearly lies in the middle of the sample. If the number of observations are odd, it is the middle most value. If the number of observations is even, it is the mean of the middle most two values.
   Mean ($\bar{y}$): sum of observation divided by number of observations
   Deviation: The difference between each data point and the mean of the entire data set deviation $=$ (observation $- \bar{y}$)

2. **Quartiles:** Divides the distribution intow four equal parts:

   - **First Quartile (Q_1)**: Median of values in the lower half of distribution
   - **Second Quartile (Q_2)**: Median of all the values in the distribution
   - **Third Quartile (Q_3)**: Median of values in the upper half of distribution

3. **Interquartile range (IQR)**: Difference between third and first quartile. $IQR = Q_3 - Q_1$

4. **Outlier**: A data point that differs *so much* from the rest of the data that it doesn't seem to belong with the other data. Outliers can be helpful in pinpointing a problem with the experimental protocol.

   Lower fence: $Q_1 - 1.5 \times$ IQR

   Upper fence: $Q_3 + 1.5 \times$ IQR

   Outlier lies out of the bounds of both the fences. That is a data point is an outlier if:

   $$\text{data point} < Q_1 - 1.5 \times \text{IQR}$$

   or

   $$\text{data point} > Q_3 + 1.5 \times \text{IQR}$$

5. **Boxplot**: A visual representation of the five-number summary (minimum, Q_1, median, Q_2 and maximum).

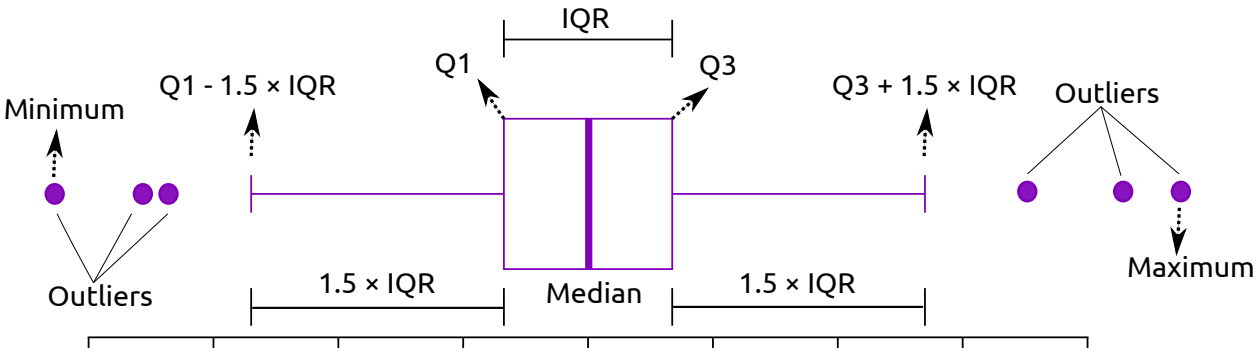

Figure 1: Boxplot with outliers. Note that if there are no outliers, the whiskers ($Q_1 - 1.5 \times$ IQR and $Q_3 + 1.5 \times$ IQR) represent the minimum and the maximum values in the data respectively.

6. **Univariate summary**: A graphical or numerical summary of a single variables.
   Bivariate summary: A graphical or numerical summary of the relationship between pairs of variables:

   - **Bivariate frequency table**: Used to understand the relationship between two categorical variables.
   - **Stacked bar chart**: A visualization of the bivariate frequency table

- **Stacked relative frequency chart**: A visualization of the bivariate frequency table such that the total counts of each category have been normalized.
- **Side-by-side boxplot**: Used to compare the center, spread, skewness and outliers of a dataset across different groups.
- **Scatterplot**: Used to examine the relationship between two numeric variables X and Y. It plots each observed pair (x, y) as a dot on the $x - y$ plane.

7. **Measures of dispersion**:

   - **Range:** The difference between the maximum and minimum values in the data (max $-$ min).
   - **Interquartile range (IQR):** Difference between third and first quartile. $Q_3 - Q_1$
   - **Standard deviation (SD):** Often denoted by s and defined as :

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

   $\sum_{i=1}^{n}(y_i - \bar{y})^2$ = Sum of squared deviations. The sum of squared deviation can also be written as $(\sum_{i=1}^{N} y_i^2) - \bar{y}^2$. $n - 1$ is called the degrees of freedom. A simple (but incomplete) explaination of the denominator being $n - 1$ is that it would otherwise not hold for $n = 1$, resulting in $\frac{0}{0}$.
   - **Coefficient of Variation**: Ratio of Standard deviation to mean (might be expresssed as percentage). It is unitless and is not affected by change in scale (unlike the above measures of dispersion) and hence is useful for comparing disperion of two or more variables that have been measured on different scales.

Measure of disperion	**Formula**	**Robust**	**Units**	**Effect of multiplicative transformation:** $a \times Y$	**Effect of additive transorfmation:** $Y + c$
Range (R)	max $-$ min	No	Same as data	Scales with the multiplicative factor: $a \times R$	Remains Same: R
Inter Quartile Range (IQR)	$Q_3 - Q_1$	Yes	Same as data	Scales with the multiplicative factor: $a \times IQR$	Remains Same: IQR
Standard Deviation (s)	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$	No	Same as data	Scales with the multiplicative factor: $a \times s$	Remains Same: s
Coefficient of Variation (cv)	$\frac{\text{standard deviation}}{\text{mean}} = \frac{s}{\bar{y}}$	No	Unitless	Remains same: cv	Remains same: cv

Table 1: Measures of dispersion

8. **Linear Transformation**: A variable Y can be transformed *linearly* to Y'. If the graph of Y' agains Y is a straight line, such a transformation is linear.

$$Y' = aY \qquad \text{Multiplicative transformation}$$
$$Y' = Y + c \qquad \text{Additive transformation}$$

Statistic	**Effect of multiplicative transformation:** $a \times Y$	**Effect of additive transorfmation:** $Y + c$
Mean ($\bar{y}$)	Scales with multiplicative factor: $a \times \bar{y}$	Shifts by constant factor: $\bar{y} + c$
Median (y_{median})	Scales with multiplicative factor: $a \times y_{\text{median}}$	Shifts by constant factor: $y_{\text{median}} + c$
Mode (y_{mode})	Scales with multiplicative factor: $a \times y_{\text{mode}}$	Shifts by constant factor: $y_{\text{mode}} + c$

Table 2: Effect of linear transformation on various descriptive statistics