Worksheet Activity:
Displaying and Summarizing Univariate Quantitative Data

Here we review common graphical tools and statistics to summarize univariate quantitative variables. In today's activity we will focus specifically on histograms, stem-and-leaf plots, means and medians.

Learning Outcomes:
By the end of the lecture, you will be expected to:

- Interpret methods for summarizing and comparing data sets: the common graphical tools of histograms and stem-and-leaf plots and the summary statistics, mean and median.
- Assess which methods for summarizing a data set are most appropriate to highlight interesting features of the data.

We will use a dataset containing the top 300 grossing movies (up to Nov 2017). This data was taken from Box Office Mojo (https://www.boxofficemojo.com/), an online database that contains information about movies, television shows, and video games.

In this activity, we will be focussing on the Adjusted Gross Earnings column, which represents the domestic box-office gross earnings adjusted for inflation. In other words, it is the box-office amount that would have been earned from ticket sales at 2017 prices (in a million dollars). The top 10 movies, according to adjusted gross earnings, are provided below.

Rank	Title	Studio	Release Year	Adjusted Gross Earnings (million dollars)	Unadjusted Gross Earnings (million dollars)
1	Gone with the Wind	MGM	1939	1,804	199
2	Star Wars	Fox	1977	1,591	461
3	The Sound of Music	Fox	1965	1,272	159
4	E.T.: The Extra-Terrestrial	Uni.	1982	1,267	435
5	Titanic	Par.	1997	1,210	659
6	The Ten Commandments	Par.	1956	1,170	66
7	Jaws	Uni.	1975	1,144	260
8	Doctor Zhivago	MGM	1965	1,109	112
9	The Exorcist	WB	1973	988	233
10	Snow White and the Seven Dwarfs	Dis.	1937	973	185
⋮	⋮	⋮	⋮	⋮	⋮

1. What type of variable (quantitative vs. categorical) is **adjusted gross earnings**?

1 Stem-and-Leaf Plots and Histograms

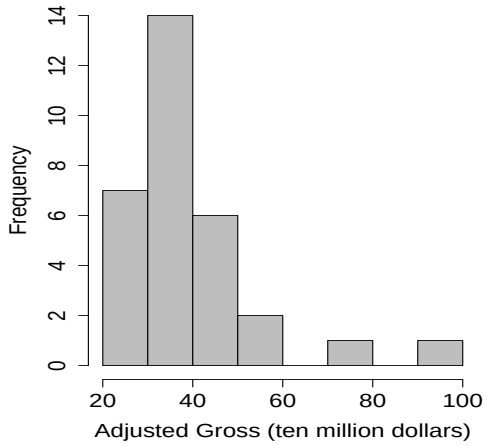
2. Let's focus on a subset of the dataset. The following are the adjusted gross earnings (rounded to ten million dollars) of 31 movies released from 2013 - 2017:

27 28 28 28 29 29 29 30 31 34 34 35 36 36 37 37
37 38 38 39 39 42 42 44 44 46 48 50 54 71 97

(a) Complete the pattern using the adjusted gross earnings (ten million dollars):

```
2 | 7 8 8 8 9 9 9
3 | 0 1 4 4 5 6 6 7 7 7 8 8 9 9
4 | 2 2 4 4 6 8
5 |
6 |
7 |
8 |
9 |
```

(b) Compare and contrast your stem-and-leaf plots above with the histogram of the adjusted gross earnings (given below).



2 Comparing the Mean and the Median

Recall:
Mean: It is the arithmetic average of the observations.

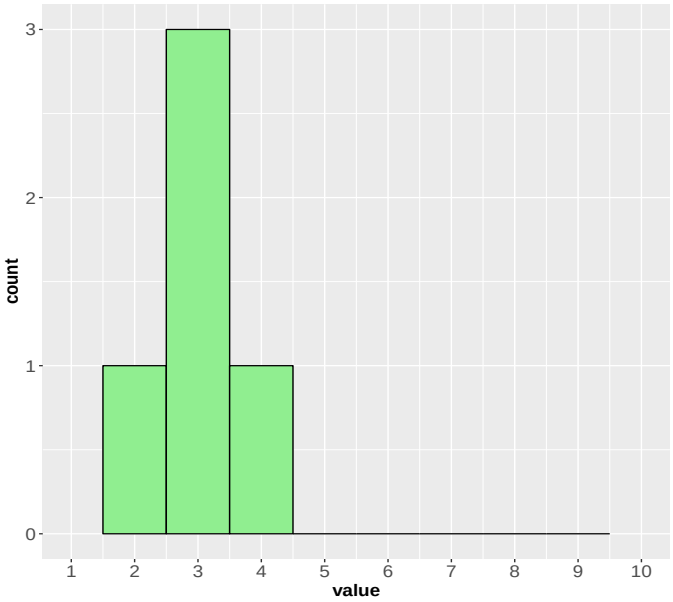
Mean $(\bar{y}) = \frac{\text{sum of values of all observations}}{\text{number of observations}}$
Formula: $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1+y_2+\ldots+y_n}{n}$
Notation: y_i is the ith observation, and n is the #observations

Median: It is the middle number of the observations. It divides the data set into two equal parts.

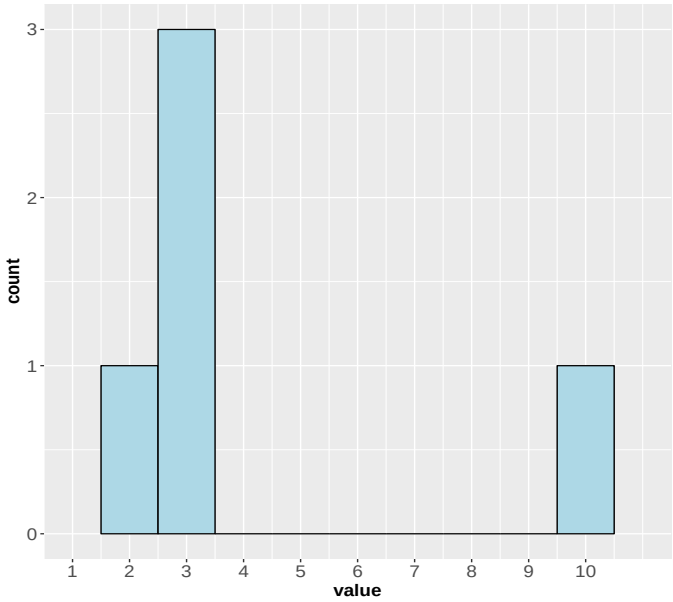
Calculating the median:
1. Arrange the data in ascending order
2. If the number of observations (n) is
odd: median = $(\frac{n+1}{2})th$ observation
even: median = average of the $(\frac{n}{2})th$ and $(\frac{n}{2}+1)th$ observations

3. Let's consider two example datasets to get a deeper understanding of the differences between two measures of center, the mean and the median.

Dataset A: 2, 3, 3, 3, **4**



Dataset B: 2, 3, 3, 3, **10**



(a) Describe the shape of the distribution for

Dataset A:

Dataset B:

(b) Calculate the mean and median for dataset B. Compare your results to the mean and median for dataset A calculated below.

Dataset A:

$mean_A = \frac{2+3+3+3+4}{5} = 3$ $\qquad$ $median_A = \frac{5+1}{2} = 3rd\ observation = 3$

Dataset B:

(c) Mark on the histograms on the previous page where the respective means and medians are for each dataset. Which numeric measure of center, the mean or the median, would you use to summarize dataset A? Dataset B? Why?