

Interactive Notes Chapter 5 – Exploring Numerical Data

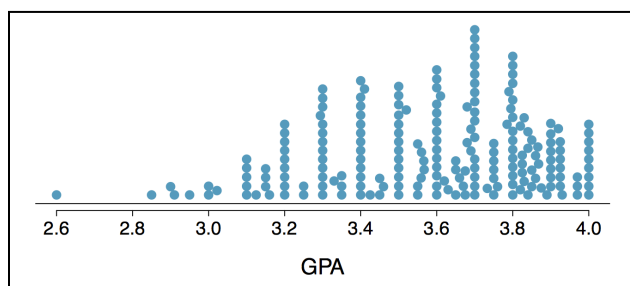
Math 217

Saidi

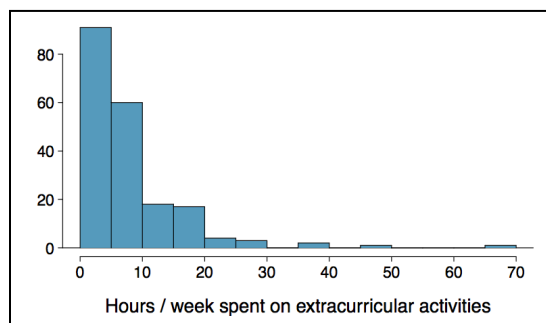
5.1 Single Numerical Variable Characteristics

Quantitative Data Distributions

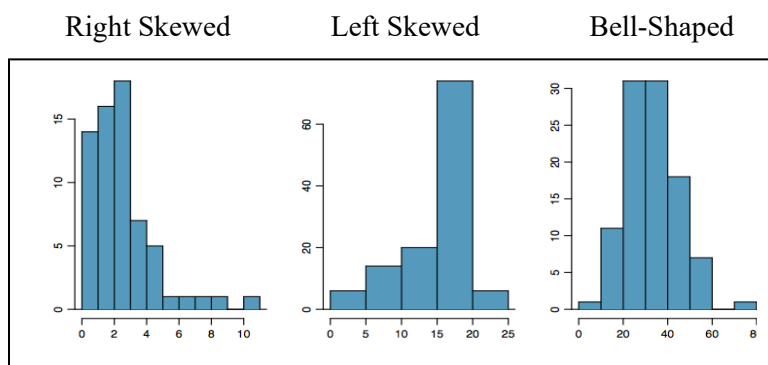
Dotplot – stacks of data points on a horizontal line



Histogram – show data density with bars; vertical axis is “frequency”



The **shape** may be described as **symmetric (bell-shaped)**, **right-skewed**, or **left-skewed**:



An **outlier** is a value that is distinctly far outside the range of the other values in the dataset.

5.4 Variance and standard deviation

How much variability there is in a dataset is an important aspect to study?

Is all the data grouped closely together?

Is the data spread out over a wide range?

Are there a few observations that are FAR from the rest of the group?

Using the mean, \bar{x} as the reference point or center, we find the distance of each point x_i from the mean. These values $(x_i - \bar{x})$ are called the deviations from the mean and can be negative as well as positive. These deviations always mathematically add up to a zero since the negative ones can be shown to exactly cancel out the positive ones.

Σ Summation

Fact: $\Sigma(x_i - \bar{x}) = 0$ for any data set.

$x_i - \bar{x}$ = how far that one observation is from mean
 $\Sigma(x_i - \bar{x})$

Thus, averaging these deviations does not tell us anything useful about the data spread. On the other hand, squaring these deviations makes them all positive and if we then average these, it will result in a useful measure of spread. It turns out that it is better to divide the sum of these squared deviations by $(n - 1)$, rather than n , to give us what we will later call 30 Descriptive Statistics: Graphical and Numerical Summaries an “unbiased” measure of the true variation. This division by $(n - 1)$ may also be justified by the fact that only $(n - 1)$ of these deviations are actually independently determined, since their total is always fixed at zero. The standard deviation is a useful alternate measure of spread:

Standard Deviation

The **standard deviation** for a quantitative variable measures the spread of the data

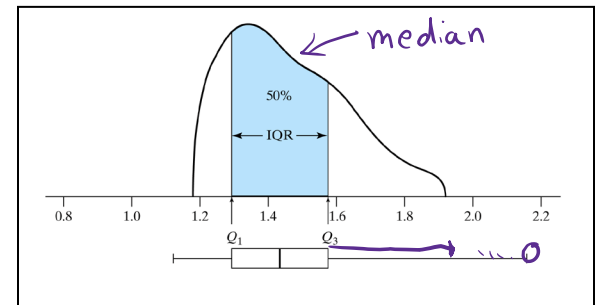
$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

- Sample standard deviation: s
- Population standard deviation: σ (“sigma”)

Table 2.6.1 Illustration of the formula for the sample standard deviation		
Observation (y_i)	Deviation ($y_i - \bar{y}$)	Squared deviation ($(y_i - \bar{y})^2$)
76	3	9
72	-1	1
65	-8	64
70	-3	9
82	9	81
Sum $365 = \sum_{i=1}^n y_i$	0	$164 = \sum_{i=1}^n (y_i - \bar{y})^2$

Visualizing the Standard Deviation

Table 2.6.2 Average daily gain (kg/day) of 39 Charolais bulls							
1.18	1.24	1.29	1.37	1.41	1.51	1.58	1.72
1.20	1.26	1.33	1.37	1.41	1.53	1.59	1.76
1.23	1.27	1.34	1.38	1.44	1.55	1.64	1.83
1.23	1.29	1.36	1.40	1.48	1.57	1.64	1.92
1.23	1.29	1.36	1.41	1.50	1.58	1.65	



For this data set, the graph to the right shows a smoothed histogram and boxplot of 39 daily gain measurements. The shaded area contains about 50% of the observations.

5.5 Box plots, quartiles, and the median

5-Number Summary – Min, Q_1, M, Q_3, Max

A boxplot is better for describing skewed distributions or those w/ strong outliers

Quartile – every 25% of the data

Outlier – shown as upper or lower points above or below the boxplot

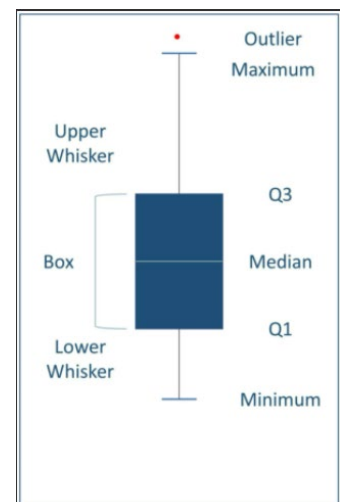
Min – the smallest valued observation

Q1 – the 25th percentile of observations (the lower quartile)

M – the 50th percentile of observations (the median)

Q3 – the 75th percentile of observations (the upper quartile)

Max – the largest valued observation



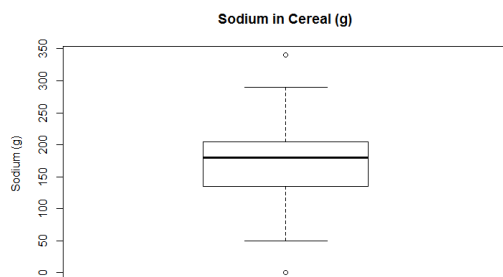
EXAMPLE: Use the data about sodium in boxed cereal:

1. List the sd, mean, and 5-number summary
2. Create a boxplot
3. Determine if there are any potential outliers in the cereal sodium data

0	340	70	140	200	180
210	150	100	130	140	180
190	160	290	50	220	180
200	210				

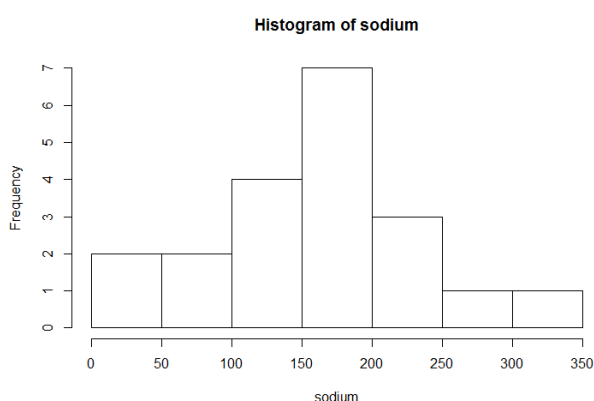
ADD TITLE AND Y-AXIS LABEL

```
boxplot(sodium, main = "Sodium in Cereal (g)", ylab = "Sodium (g)") # If you are interested in putting a title and y-axis label
```



Notice there are both upper and lower outliers. You should try using the formula to prove these are outliers:

CREATE A HISTOGRAM



There is a default number of breaks for a base histogram (function “hist”). If you want to change the number of breaks, you simply write the code:

Try playing with changing the number of breaks to see how that changes the histogram’s shape (i.e. distribution). If you have questions on how “breaks” works, try the code:

```
?hist
```

This will provide more information about how breaks are calculated.

Live Coding: Working with single numerical variables

Use the chapter5notes.qmd file to follow this section:

Numerical variables: the distribution of shape, center, and spread

When you want to know more about a function, use a question mark with the function

Example: learn more about the function, boxplot

```
# ?boxplot (delete the # to run ?boxplot)
```

I. Shape

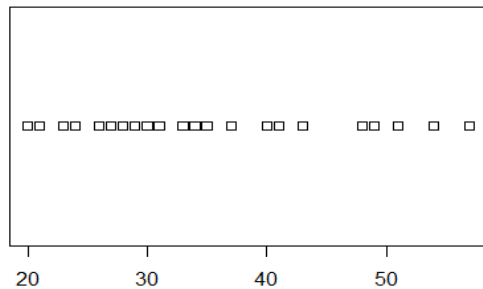
1. Dotplots (not necessarily the best way to show distributions)

```
# load the vector of data for number of segments
```

```
segment <- c(23, 30, 54, 28, 31, 29, 34, 35, 30, 27, 21, 43, 51, 35, 51, 49, 35, 24, 26,  
29, 21, 29, 37, 27, 28, 33, 33, 23, 37, 27, 40, 48, 41, 20, 30, 57)
```

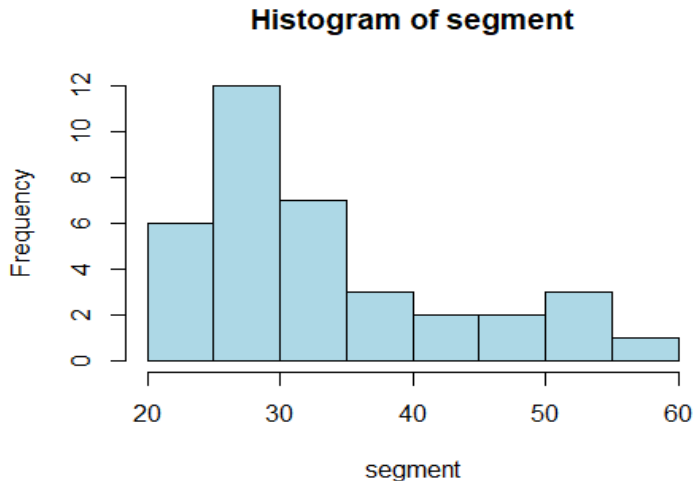
```
# stripchart is a base-R dotplot
```

```
stripchart(segment)
```



2. Histograms are better at displaying a numerical variable's shape

```
hist(segment, col = "lightblue")
```



We can see that the distribution of this variable is right-skewed (the tail is to the right)

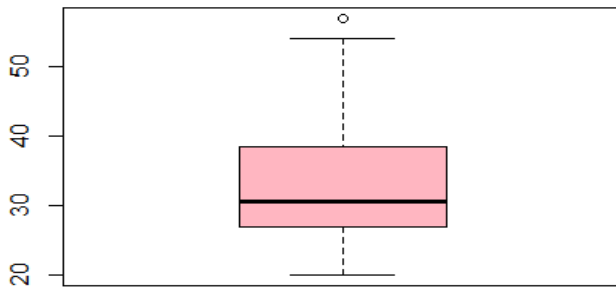
3. Boxplots show OTHER information about the distribution, based on the 5-number summary

```
# 5-number summary + mean
```

```
summary(segment)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20.00	27.00	30.50	33.78	37.75	57.00

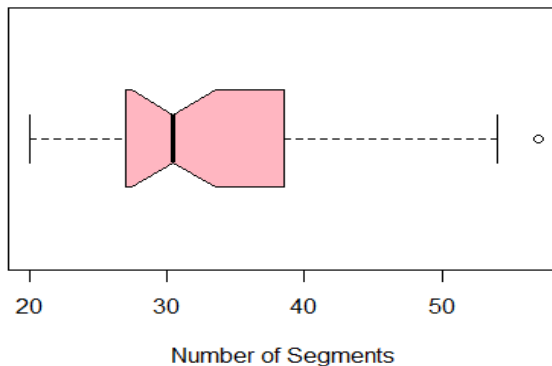
```
# now create the boxplot using these values
boxplot(segment, col = "lightpink")
```



- Notice that there is one upper outlier, which is 57, based on the max value from the summary output.
- The median is 30.5.
- The IQR ($Q3 - Q1 = 37.75 - 27 = 10.75$) is the middle box

```
# Rotate your boxplot horizontally and add a notch at the median.
# Add a title (main =) and x-axis label (xlab =)
boxplot(segment, col = "lightpink", notch = TRUE, horizontal = TRUE, main = "Notched boxplot of number of segments", xlab = "Number of Segments")
```

Notched boxplot of number of segments



II. Center

The two measure of center for a numerical variable are: **mean** and **median**

```
mean(segment)
[1] 33.77778

median(segment)
[1] 30.5
```

When the mean and median are essentially the same value, that indicates a bell shape. When mean is greater than median (as it is in this case), the distribution is right-skewed, since the mean is dragged towards the tail.

III. Spread

Spread describes how wide the values are across a distribution. We can use the following:

- range = max - min
- IQR = Q3 - Q1
- standard deviation - this has a complicated formula (see notes). The function in R is “sd”

```
IQR(segment)
```

```
[1] 10.75
```

```
range(segment)
```

```
[1] 20 57
```

```
sd(segment)
```

```
[1] 9.865122
```

Your Turn!!!

Use the following data: sodium

Sodium is a vector containing the grams of sodium in 20 cereal boxes.

```
sodium <- c(0, 340, 70, 140, 200, 180, 210, 150, 100, 130, 140,
180, 190, 160, 290, 50, 220, 180, 200, 210)
```

Create a chunk for each question below and use appropriate code to answer.

1. Create a dotplot, histogram, and boxplot of the distribution of sodium.

```
# dotplot
```

```
# histogram
```

```
# boxplot
```

2. Find the 5-number summary plus mean for sodium.
3. What is the IQR and sd?
4. Make a histogram or boxplot, but add a title, axis labels, and change the color.

5.6 Robust statistics

Robust statistics are values that are **not heavily influenced by outliers**. The **median and IQR** are called robust statistics because extreme observations have little effect on their values: moving the most extreme value generally has little influence on these statistics. On the other hand, the **mean and standard deviation are more heavily influenced by changes in extreme observations**. The choice of center affects the types of statistical analysis you will need to choose.

5.7 Transforming data

When data are very strongly skewed, we sometimes transform them so they are easier to model.

A scatterplot of the population change from 2010 to 2017 against the population in 2010 is shown in Figure 5.11. In this first scatterplot, it's hard to decipher any interesting patterns because the population variable is so strongly skewed (left plot). However, if we apply a \log_{10} transformation to the population variable, as shown in Figure 5.11, a positive association between the variables is revealed (right plot). In fact, we may be interested in fitting a trend line to the data when we explore methods around fitting regression lines in Chapter 7.

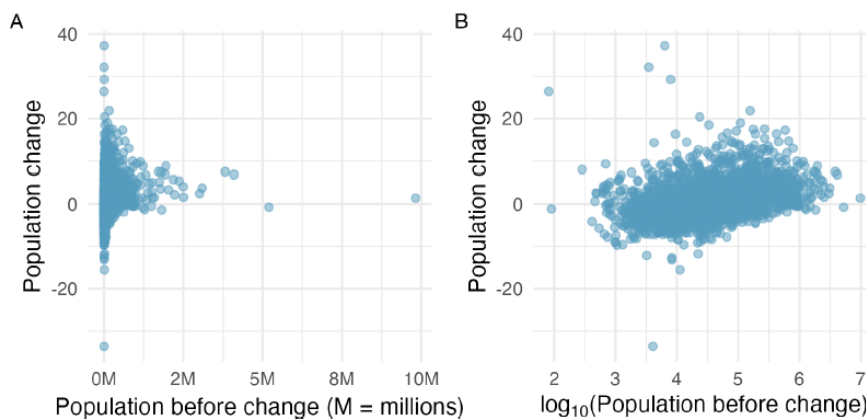


Figure 5.11: Plot A: Scatterplot of population change against the population before the change. Plot B: A scatterplot of the same data but where the population size has been log-transformed.

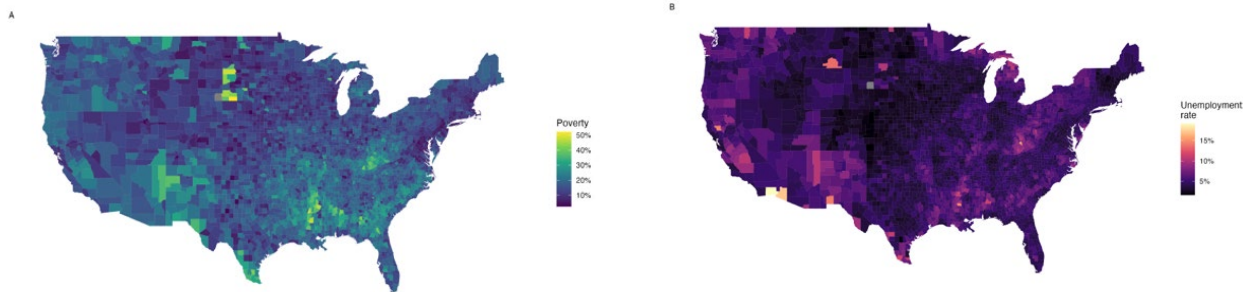
Transformation Options:

1. Log transformations: `log(variable)`
2. Square root: `sqrt(variable)`
3. Inverse: `(1/variable)`

Transformations other than the logarithm can be useful, too. For instance, the square root $\sqrt{\text{original observation}}$ and inverse $\left(\frac{1}{\text{original observation}}\right)$ are commonly used by data scientists. Common goals in transforming data are to see the data structure differently, reduce skew, assist in modeling, or straighten a nonlinear relationship in a scatterplot.

5.8 Mapping data

For **geographic data**, create an **intensity map**, where colors are used to show higher and lower values of a variable. Figures 5.12 and 5.13 show intensity maps for **poverty rate in percent (poverty)** and **unemployment rate in percent (unemployment_rate)**. The color key indicates which colors correspond to which values. The intensity maps are not generally very helpful for getting precise values in any given county, but they are very helpful for seeing geographic trends and generating interesting research questions or hypotheses.



Homework Chapter 5

1. Review section 5.9 (the chapter review)
2. Suggested problems from textbook section 5.10 exercises: 10, 11 – 25 odd only
3. Suggested tutorials: [Tutorial 1 – Lessons 1, 2, and 3](#)
 - [Visualizing categorical data](#)
 - [Visualizing numerical data](#)
 - [Summarizing with statistics](#)