Centrality and Dispersion

Part 1: Concepts and Calculations

1

Below is a list of voter turnout rates in twelve Wisconsin Counties during the 2020 presidential election. Calculate the mean, median, and mode for the level of voter turnout in these counties.

Wisconsin County	Pct Turnout
Clark	63
Dane	87
Forrest	71
Grant	63
Iowa	78
Iron	82
Jackson	65
Kenosha	71
Marinette	71
Milwaukee	68
Portage	74
Taylor	70

```r
#Calculate the mean:
sum(63+87+71+63+78+82+65+71+71+68+74+70)
```

```
## [1] 863
```

```r
863/12
```

```
## [1] 71.91667
```

```r
#Calculate the median:
#(n+1)/2
(12+1)/2
```

```
## [1] 6.5
```

```r
#sorted variables in order of magnitude:
#63,63,65,68,70,71,71,71,74,78,82,87
#The value associated with the 6.5th obseervation is 71, therefore the median would be 71

#Calculate the mode:
#71 is the number with the highest frequency, appearing 3 times.
```

Mean: 71.92 Median: 71 Mode: 71

I believe that in this case, the median is the most appropriate measure of central tendency because it captures what we could expect to be a potential mid number for Presidential election turnout rate among all twelve counties in Wisconsin. The median observation is 6.5, which lies right in between the mode of the sample, 71. Therefore, both the median and mode of this observation are both 71. However, because we have a high outlier (87% voter turnout), the mean, or average, turns out to be slightly skewed (71.92), thus not appropriately reflecting what the smaller half of the values in this sample are conveying.

2

Fill in the deviation of each observation from the mean and the square of the deviation and calculate the mean absolute deviation and the standard deviation. Interpret these statistics.

Wisconsin County	Pct Turnout	$X_i - \overline{X}$	$(X_i - \overline{X})^2$
Clark	63	-8.92	79.5664
Dane	87	15.08	227.4064
Forrest	71	-0.92	0.8464
Grant	63	-8.92	79.5664
Iowa	78	6.08	36.9664
Iron	82	10.08	101.6064
Jackson	65	-6.92	47.8864
Kenosha	71	-0.92	0.8464
Marinette	71	-0.92	0.8464
Milwaukee	68	-3.92	15.3664
Portage	74	2.08	4.3264
Taylor	70	-1.92	3.6864

```
#Calculate standard deviation from the mean
63-71.92
```

```
## [1] -8.92
```

```
87-71.92
```

```
## [1] 15.08
```

```
71-71.92
```

```
## [1] -0.92
```

```
63-71.92
```

```
## [1] -8.92
```

```
78-71.92
```

```
## [1] 6.08
```

```
82-71.92
```

```
## [1] 10.08
```

```
65-71.92
```

```
## [1] -6.92
```

```
71-71.92
```

```
## [1] -0.92
```

```
71-71.92
```

```
## [1] -0.92
```

```
68-71.92
```

```
## [1] -3.92
```

```
74-71.92
```

```
## [1] 2.08
```

```
70-71.92
```

```
## [1] -1.92
```

```
#Square the deviations
(63-71.92)^2
```

```
## [1] 79.5664
```

```
(87-71.92)^2
```

```
## [1] 227.4064
```

```
(71-71.92)^2
```

```
## [1] 0.8464
```

```
(63-71.92)^2
```

```
## [1] 79.5664
```

```
(78-71.92)^2
```

```
## [1] 36.9664
```

```
(82-71.92)^2
```

```
## [1] 101.6064
```

```
(65-71.92)^2
```

```
## [1] 47.8864
```

```
(71-71.92)^2
```

```
## [1] 0.8464
```

```
(71-71.92)^2
```

```
## [1] 0.8464
```

```
(68-71.92)^2
```

```
## [1] 15.3664
```

```
(74-71.92)^2
```

```
## [1] 4.3264
```

```
(70-71.92)^2
```

```
## [1] 3.6864
```

```
#Calculate the mean absolute deviation
(-8.92+15.08-0.92-8.92+6.08+10.08-6.92-0.92-0.92-3.92+2.08-1.92)/12
```

```
## [1] -0.003333333
```

```
#Calculate standard deviation
(-8.92+15.08-0.92-8.92+6.08+10.08-6.92-0.92-0.92-3.92+2.08-1.92)/11 #variance
```

```
## [1] -0.003636364
```

```
sqrt(-0.003636364)
```

```
## Warning in sqrt(-0.003636364): NaNs produced

## [1] NaN
```

I believe that the easiest statistic to interpret is the standard deviation, because it shows how dispersed the data is as compared to the mean of the sample. Therefore, I think it gives a good overall glimpse into how diverse the sample results are, and they show us if our data looks viable to use or not.

```
#Coefficient of variation
(-0.003636364)/(71.92)
```

```
## [1] -5.056123e-05
```

The coefficient of variation tells us that the results for the Presidential election voter turnout percentage are very closely concentrated around the mean.

3

Across the fifty states, the average cumulative number of COVID-19 cases per 10,000 population in August of 2021 was 1161, and the standard deviation was 274. The cases per 10,000 were 888 in Virginia and 1427 in South Carolina. What percent of states do you estimate have values equal to or less than Virginia's, and what percent do you expect to have values equal to or higher than South Carolina's?

```
#Calculate deviation from the mean
888-1161 #Virginia
```

```
## [1] -273
```

```
1427-1161 #South Carolina
```

```
## [1] 266
```

```
#Express deviation from the mean, relative to standard deviation
-273/274 #Virginia
```

```
## [1] -0.9963504
```

```
266/274 #South Carolina
```

```
## [1] 0.9708029
```

Virgina's value is about one standard deviation below the mean, which means that around 68.26% of the states monitored will have equal or lower values. Similarly, South Carolina's value is one standard deviation above the mean, which also means that around 68.26% of the states monitored will have equal or lower values. In order to calculate this, I subtracted Virginia's and South Carolina's values from the sample's mean value, and then divided the results by the sample's standard deviation.

Part 2: R Problems

1

Use R to report all measures of central tendency that are appropriate for each of the following variables: The feeling thermometer rating for the National Rifle Association (`anes20$V202178`), Latinos as a percent of state populations (`states20$latino`), party identification (`anes20$V201231x`), and region of the country where ANES survey respondents live (`anes20$V203003`). Where appropriate, also discuss skewness.

Feeling thermometer rating for the National Rifle Association (`anes20$V202178`)

```r
load("data/anes20.rda")

#Calculate Median
#Get the median, treating the variable as numeric
median(as.numeric(anes20$V202178), na.rm=T)
```

```
## [1] 50
```

```r
#Calculate Mean
mean(anes20$V202178, na.rm=T)
```

```
## [1] 48.53579
```

Latinos as a percent of state populations (`states20$latino`)

```r
load("data/states20.rda")

library(moments)
library(descr)
library(DescTools)
library(methods)
library(stats)
library(utils)

#Calculate Median
#Get the median, treating the variable as numeric
median(states20$latino)
```

```
## [1] 4.65
```

```r
#Calculate Mean
mean(states20$latino)
```

```
## [1] 7.402
```

```r
#Calculate Skewness
Skew(states20$latino)
```

```
## [1] 2.167047
```

Party identification (`anes20$V201231x`)

```r
#Calculate Mode
Mode(anes20$V201231x, na.rm=T)
```

```
## [1] 1. Strong Democrat
## attr(,"freq")
## [1] 1961
## 7 Levels: 1. Strong Democrat ... 7. Strong Republican
```

```
#Calculate Median
#Get the median, treating the variable as numeric
median(as.numeric(anes20$V201231x), na.rm=T)
```

```
## [1] 4
```

```
#Calculate Mean
mean(as.numeric(anes20$V201231x), na.rm=T)
```

```
## [1] 3.887811
```

```
#Calculate Skewness
Skew(as.numeric(anes20$V201231x), na.rm=T)
```

```
## [1] 0.0766003
```

Region of the country where ANES survey respondents live (`anes20$V203003`)

```
#Calculate Mode
Mode(anes20$V203003, na.rm=T)
```

```
## [1] 3. South
## attr(,"freq")
## [1] 3081
## Levels: 1. Northeast 2. Midwest 3. South 4. West
```

```
#Calculate Median
#Get the median, treating the variable as numeric
median(as.numeric(anes20$V203003), na.rm=T)
```

```
## [1] 3
```

```
#Calculate Mean
mean(as.numeric(anes20$V203003), na.rm=T)
```

```
## [1] 2.639734
```

```
#Calculate Skewness
Skew(as.numeric(anes20$V203003), na.rm=T)
```

```
## [1] -0.2386413
```

2

Create a density plot that includes vertical lines showing the mean and median outcomes.

Density plot for Latinos as a percent of state populations (`states20$latino`)

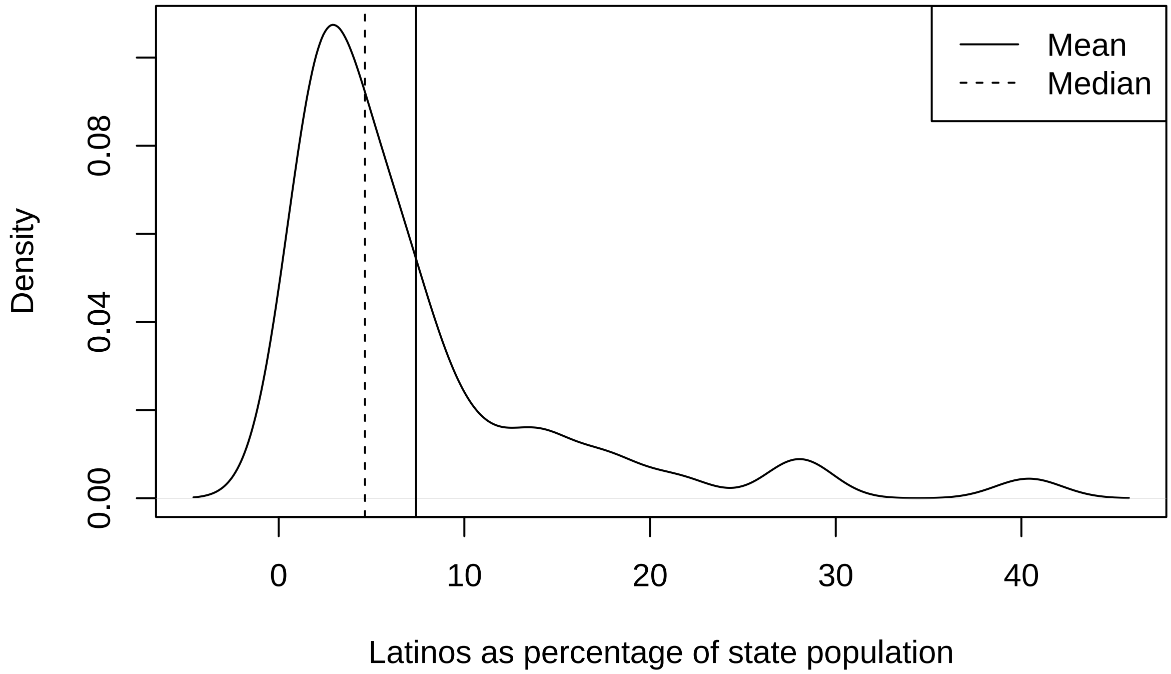
```
median(states20$latino)
```

```
## [1] 4.65
```

```
mean(states20$latino)
```

```
## [1] 7.402
```

```
plot(density(states20$latino),
xlab="Latinos as percentage of state population",
main="")
#Insert vertical lines for mean and median
abline(v=mean(7.402))
abline(v=median(4.65), lty=2)
#Add the legend
legend("topright", legend=c("Mean", "Median"), lty=1:2)
```



As gathered from previously calculating the skewness of this variable, the graph for "Latinos as a percentage of state population" is very skewed. The difference between the mean and the median is 2.752%, a relatively small number if we contextualize it within a state's entire population percentage. In this case, our distribution's mean is somewhat higher than its median, thus making it a positively skewed graph. The skewedness of this graph and its relative variable might be due to the fact that those states with closer proximity to the South American border will definitely see a bigger influx of Latinos citizens trying to migrate or establish themselves permanently in those areas. States with an historical higher presence of Latinos will also have a higher percentage.

3

Estimate the area under the normal curve for each of the following:

- Above Z=1.8

```
#Get area under the curve to the right (or above) z=1.8
pnorm(1.8, lower.tail = F)
```

```
## [1] 0.03593032
```

- Z= -1.3

```
#Get area under the curve to the left of z=-1.3
pnorm(-1.3)
```

```
## [1] 0.09680048
```

- Between Z=-1.3 and Z=1.8

```
#Area between z=-1.3 and z=1.8
0.03593032+0.09680048
```

```
## [1] 0.1327308
```

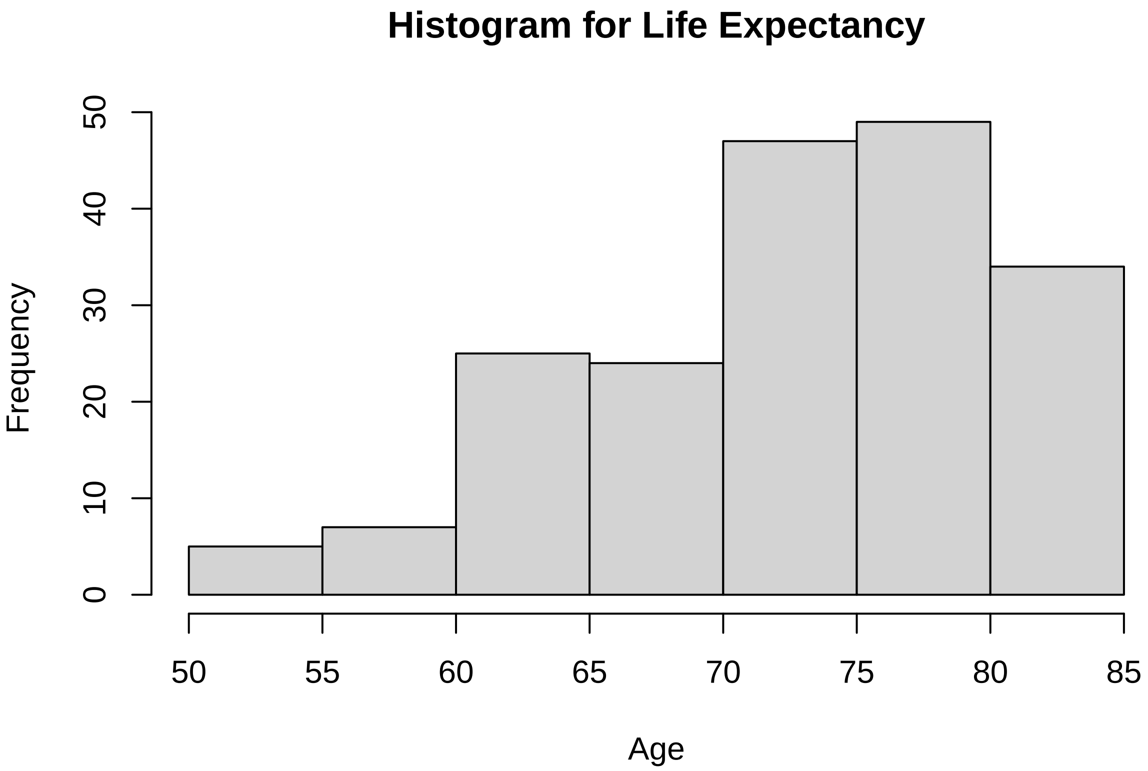
```
100-0.1327308
```

```
## [1] 99.86727
```

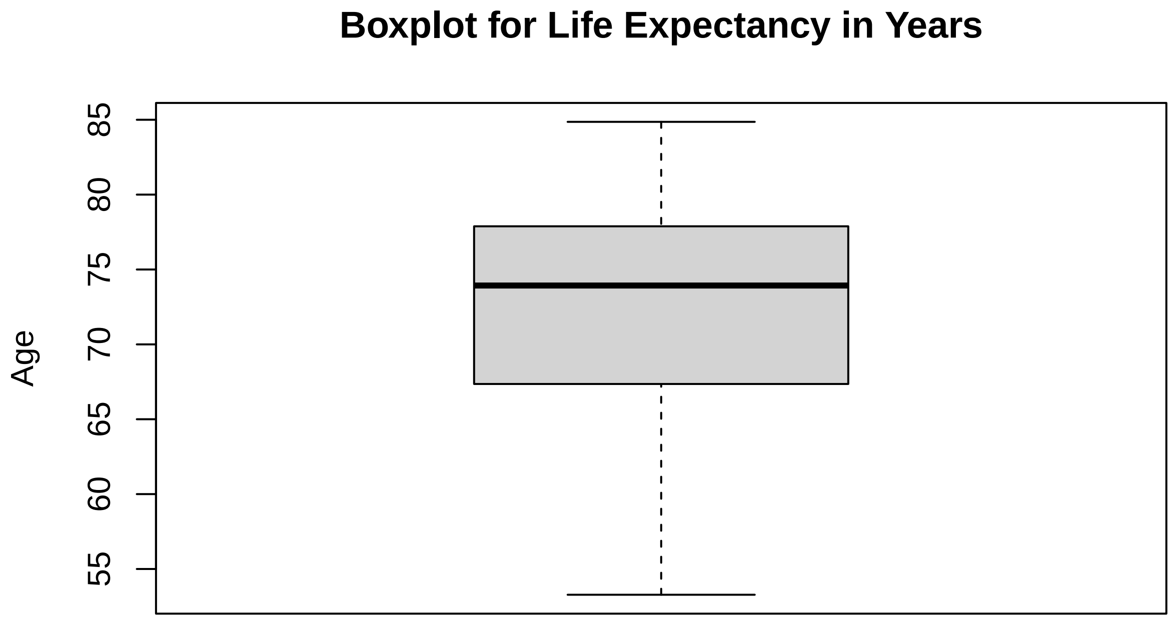
4

Using a histogram and a boxplot, describe the distribution of life expectancy (`countries2$lifexp`). This variable is an estimate of the number of years a typical person born today is expected to live.

```
load("data/countries2.rda")
#Histogram
hist(countries2$lifexp, xlab="Age",
main="Histogram for Life Expectancy")
```



```
#Boxplot
boxplot(countries2$lifexp, main="Boxplot for Life Expectancy in Years", ylab="Age")
```



I believe that using a boxplot in order to estimate the variation of life expectancy is the most useful, as the boxplot itself points out different useful statistics that can give us a more well-rounded picture of our data. However, I believe that if we want to zone in on skewness alone, the histogram might paint a clearer picture right away, because its columns show at a glance which results come up the most often and how the data might be skewed. Both graphs clealry show that the variable for life expectancy is contained within a range of years that goes from 50 years to 85 years. The histogram clearly shows that the data is skewed, since the majority people are expected to die between the ages of 70 to 80 years old, and much fewer people are expected to die younger, from 50 to 70 years of age. On the other hand, the boxplot shows us that the median for this distribution is around 74-75 years, while the middle 50% of the responses is situated between 66-67 years (25th percentile) and 76-77 years (75th percentile). The fact that the median is much closer to the 75th percentile shows us that this distribution is quite skewed. In addition, the boxplot also points out the lowest non-outlier outcome, 50 years, and the highest non-outlier outcome, 85 years.

5

Use the Desc command and describe the amount of variation in life expectancy, focusing on the range, inter-quartile range, and the standard deviation.

```
Desc(countries2$lifexp, plot=F)
```

```
## ------------------------------------------------------------------------------
## countries2$lifexp (numeric)
##
##   length       n     NAs  unique      0s    mean  meanCI'
##      195     191       4     182       0  72.629  71.569
##             97.9%     2.1%              0.0%            73.690
##
##      .05      .10      .25   median      .75      .90      .95
##   59.015   61.580   67.355   73.930   77.885   82.240   82.885
##
##    range       sd    vcoef      mad      IQR     skew     kurt
##   31.580    7.430    0.102    7.146   10.530   -0.550   -0.432
##
## lowest : 53.28, 54.24, 54.33, 54.69, 54.7
## highest: 83.57, 83.62, 83.78, 84.63, 84.86
##
## ' 95%-CI (classic)
```

The range for this variable is 31.580 years, meaning that the difference between the youngest age at which a person can die and its oldest, is around 31.580 years. The interquartile range is 10.530 years, meaning that 50% of the population (or the observations between the 25th percentile and the 75th percentile) falls within 10.530 years of life expectancy. This shows us that most of the countries' life expectancy is contained within a range of ten years. The standard deviation for life expectancy is 7.430 years, meaning that the "typical" deviation from the mean is about seven years.

6

Suppose you want to compare the amount of variation in life expectancy to variation in Gross Domestic Product (GDP) per capita (countries2$gdppc).

```
#Life Expectancy
Desc(countries2$lifexp, plot=F)
```

```
## ------------------------------------------------------------------------------
## countries2$lifexp (numeric)
##
##   length        n      NAs   unique       0s     mean   meanCI'
##      195      191        4      182        0   72.629   71.569
##             97.9%     2.1%              0.0%            73.690
##
##      .05      .10      .25   median      .75      .90      .95
##   59.015   61.580   67.355   73.930   77.885   82.240   82.885
##
##    range       sd    vcoef      mad      IQR     skew     kurt
##   31.580    7.430    0.102    7.146   10.530   -0.550   -0.432
##
## lowest : 53.28, 54.24, 54.33, 54.69, 54.7
## highest: 83.57, 83.62, 83.78, 84.63, 84.86
##
## ' 95%-CI (classic)
```

```
#GDP Per Capita
Desc(countries2$gdp_pc, plot=F)
```

```
## ------------------------------------------------------------------------------
## countries2$gdp_pc (numeric)
##
```

```
##        length            n          NAs       unique           0s        mean'
##           195          184           11          = n            0   20'623.061
##                      94.4%         5.6%                       0.0%
##
##           .05          .10          .25       median          .75          .90
##     1'657.043    2'240.906    5'040.056   13'365.924   30'010.376   49'819.193
##
##         range           sd        vcoef          mad          IQR         skew
##   113'729.870   20'902.473        1.014   14'665.721   24'970.319        1.611
##
##        meanCI
##    17'582.747
##    23'663.375
##
##           .95
##    59'811.432
##
##          kurt
##         2.840
##
## lowest : 751.664, 944.868, 1'059.723, 1'097.949, 1'219.077
## highest: 68'627.829, 86'781.390, 92'651.070, 97'341.469, 114'481.534
##
## ' 95%-CI (classic)
```

In order to make this comparison, I would use the interquartile range, for life expectancy and GDP per capita. The IQR for life expectancy is 10.530 years, meaning that 50% of the population (or the observations between the 25th percentile and the 75th percentile) falls within 10.530 years of life expectancy. The IQR for GDP per capita is 24'970.319, meaning that 50% of the population falls within a difference of 24'970.319 in GDP per capita. However, thse two statistics are nt only measuring totally different things, but they have also been measured in different ways (age against dollars). Therefore, we cannot make an accurate comparison between the two.