Measures of Central Tendency and Dispersion

Simon J. Kiss

17/03/2020

Learning Outcomes

1. students should be able to demonstrate an understanding of:

    - *mean*
    - *median*
    - *mode*
    - *variance*
    - *standard deviation*
    - *Normal distribution*

Goals of the presentation

2. Demonstrate concepts using the Labour Force Survey

3. R code provided is meant to:

   - *reinforce previous familiarity with R*
   - *demonstrate concepts*

Central Tendency

- One of the goals is to describe the world with numbers.
- describing where the centre of a set of data is is pretty useful

Mean

- most common measure of central tendency

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Mean

Example

Load data from the Labour Force Survey

- These data are stored in a file-format called `sav` which is really common in the social sciences
- loading the `haven` library provides the `read_sav()` command to read it in
- loading the `labelled` library lets us search through variables quickly

Mean

Example

```r
#install.packages(c('haven', 'labelled'))
library(haven)
lfs<-
      read_sav('https://github.com/sjkiss/DMJN328/raw/master/Lecture_Notes/mar_11/data/lfs.sav')
```

```r
library(labelled)
look_for(lfs, "wages")
```

```
##    variable                                     label
## 36 HRLYEARN Usual hourly wages, employees only
```

Mean

Example

```
look_for(lfs, "education")
```

```
##    variable                             label
## 11     EDUC Highest educational attainment
```

Mean

Example

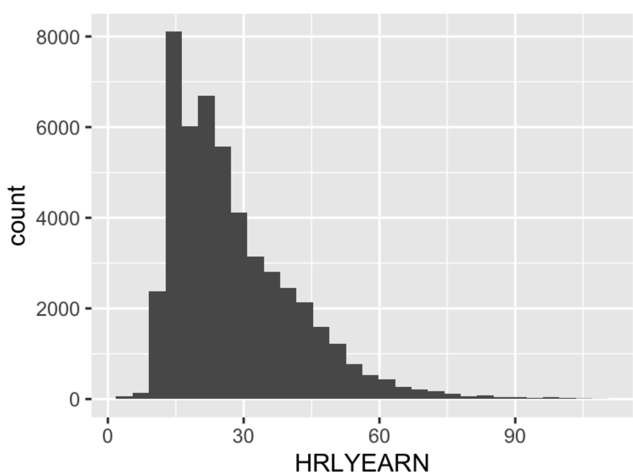
```
look_for(lfs, "sex")
```

```
##   variable              label
## 9      SEX Sex of respondent
```

Mean

Example

```
library(tidyverse)
lfs %>%
  ggplot(., aes(x=HRLYEARN))+geom_histogram()
```



Mean

Calculate mean

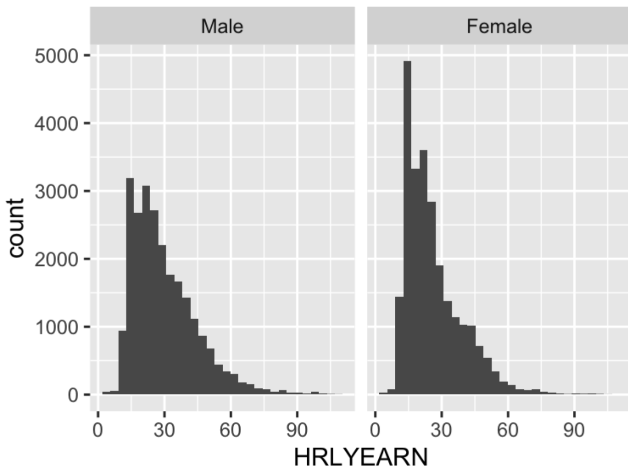
```
mean(lfs$HRLYEARN, na.rm=T)
```

```
## [1] 27.88198
```

Mean

Means by group

```
lfs %>%
ggplot(., aes(x=HRLYEARN))+geom_histogram()+facet_wrap(~as_factor(SEX))
```



Mean

Means by group

```
lfs %>%
group_by(SEX) %>%
summarize(avg=mean(HRLYEARN, na.rm=T))
```

```
## # A tibble: 2 x 2
##           SEX   avg
##     <dbl+lbl> <dbl>
## 1 1 [Male]    29.8
## 2 2 [Female]  26.0
```

Median

- means are vulnerable to outliers

```
vector1<-c(1,2,3,4,5,6,7,8,9,10)
mean(vector1)
```

```
## [1] 5.5
```

```
vector2<-c(1,2,3,4,5,6,7,8,9,10, 1000000)
mean(vector2)
```

```
## [1] 90914.09
```

Median

- Median is a different measure of central tendency
- The value at which half of a variable is above, half is below.

```
lfs %>%
  group_by(SEX) %>%
  summarize(median=median(HRLYEARN, na.rm=T))
```

```
## # A tibble: 2 x 2
##           SEX median
##     <dbl+lbl>  <dbl>
## 1 1 [Male]      26.4
## 2 2 [Female]    22.3
```

Median

- median is immune to outliers

```
median(vector1)
```

```
## [1] 5.5
```

```
median(vector2)
```

```
## [1] 6
```

Mode

- Mode is the most frequently occurring variable in the data
- useful for categorical data

Var1	**Freq**
0 to 8 years	4841
Some high school	11951
High school graduate	20006
Some postsecondary	6207
Postsecondary certificate or diploma	33779
Bachelor’s degree	15166
Above bachelor’s degree	7225

Mode

- Not usually measured for numeric data

Measures of Dispersion

Variance

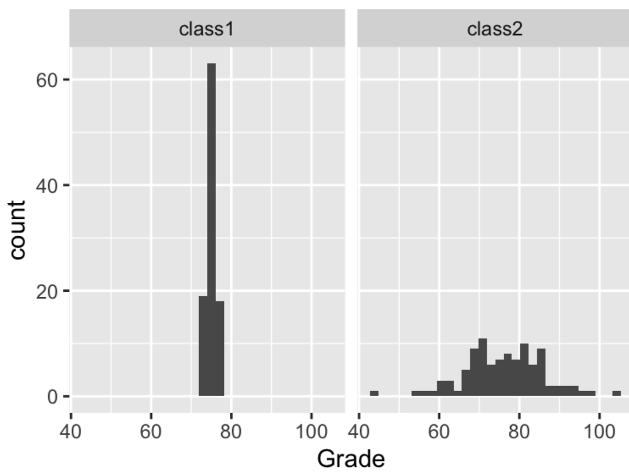
- Often not just interested in the center of the data, but the distribution

```
#make one class of fake data, average =75, standard deviation =1
class1<-rnorm(100, mean=75, sd=1)
#make a second class of fake data, average-=75, standard deviation = 10
class2<-rnorm(100, mean=75, sd=10)
#combine into a dataframe
df<-data.frame(class1, class2)
```

Measures of Dispersion

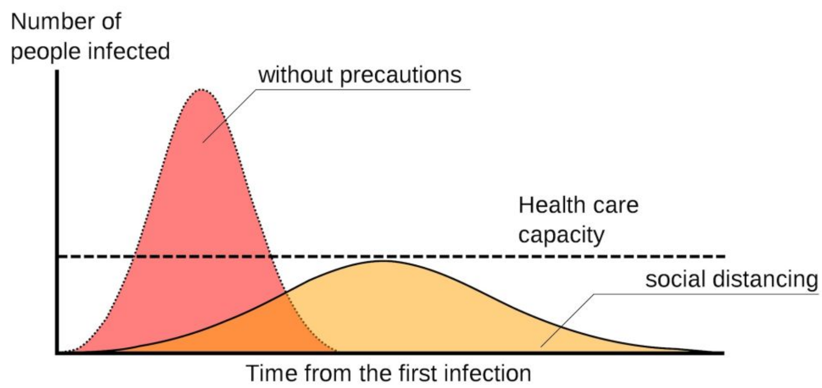
Variance

```
df %>%
  #gather into Class and Grade
  gather(Class, Grade) %>%
  #Graph and facet
  ggplot(., aes(x=Grade))+geom_histogram()+facet_wrap(~Class)
```



Measures of Dispersion

- different spreads have very different real-life consequences



Measures of Dispersion

Variance

1. Subtract the average from each value
2. Square it to get rid of the negatives
3. Sum everything up
4. divide by the sample size.

Measures of Dispersion

Variance

$$s^2 = \text{sample variance}$$

$$\sum = \text{sum everything from the bottom to the top}$$

$$\bar{x} = \text{sample average}$$

$$N = \text{sample size}$$

$$s^2 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^2}{N - 1}$$

Variance

```
#Step 1 is to subtract hte average from each value
step1<-lfs$HRLYEARN-mean(lfs$HRLYEARN, na.rm=T)
#step 2 is to square to get rid of the means
step2<-step1^2
#step 3 is to sum all of the variances, na.rm=T to remove missing values
step3<-sum(step2, na.rm=T)
#step 4 is to divide by the sample size (excluding missing values)
variance<-step3/(length(na.omit(lfs$HRLYEARN))-1)
#compare with the base
print(variance)
```

```
## [1] 194.1875
```

```
var(lfs$HRLYEARN, na.rm=T)
```

```
## [1] 194.1875
```

Standard Deviation

- variance is expressed in units squared
- unsquare it gives us a standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N - 1}}$$

- Commonly we talk about something being 1 standard deviation away, or two standard deviations away.
- Standard deviation is a number that describes the averge distance from the average in the units that variable is taken.

Distributions

- Data come in different types (categorical, numeric)
- there are different processes in the universe that generate data
- data follow different distributions
    - *Recall from ** The Joy of Stats** the importance of a distribution*

Distributions

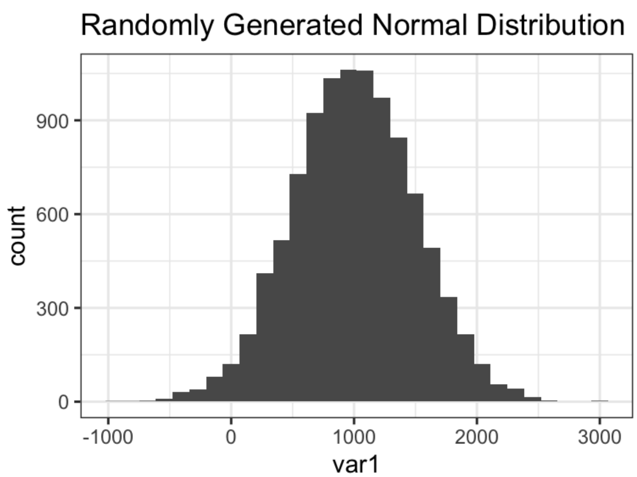
1. Normal distribution

    - *mean = median = mode*
    - *contstantly reoccurring pattern in nature*

Distributions

1. Normal distribution

```
#use rnorm to generate 10000 random numbers according to the normal distribution
#mean of 1000 and standard deviation of 500
var1<-rnorm(10000, mean=1000, sd=500)
#make into a data frame
df<-data.frame(var1)
#graph a histogram
ggplot(df, aes(x=var1))+geom_histogram()+theme_bw()+labs(title="Randomly Generated Normal
        Distribution")
```




```
### Go back upand play with the mean and sd.
```

Distributions

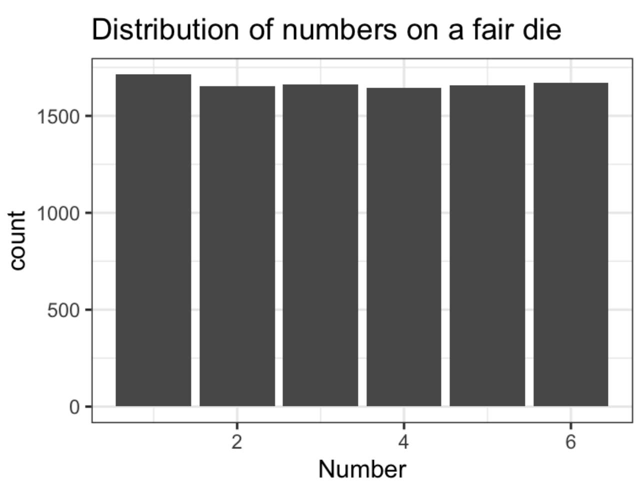
2. Uniform Distribution

- *numbers on a die*

Distributions

2. Uniform Distribution

```r
#Sample a number from 1 to 6 (like on a die roll), 10000 times
var1<-sample(1:6, 10000, replace=T)
#turn into a dataframe (die)
die<-data.frame(var1)
#Graph as a
ggplot(die, aes(x=var1))+
  #as a barplot, counting the numbr of times each number occurs
  geom_bar(stat="count")+
  #turn it black and white
  theme_bw()+
  #give some labels
  labs(title="Distribution of numbers on a fair die", x="Number")
```



Distributions

Normal Distribution

```
mean(df$var1, na.rm=T)
```

```
## [1] 1001.175
```

```
median(df$var1, na.rm=T)
```

```
## [1] 997.6412
```

Normal Distribution

- an absolute key feature of the normal distribution is that approximately:
    - *68% of all cases lie within one standard deviation of the mean;*
    - *95% of cases lie within two standard deviations of the mean and*
    - *99% of cases lie within three standard deviations from the mean.*

Normal Distribution