POLS0083: Quantitative Data Analysis

Lecture 3: Description & Measurement

Michal Ovádek & Indraneel Sircar

Today's Lecture

This course is about making arguments with quantitative evidence.

If you have a collection of data, you cannot simply *show* people the data.

To communicate effectively, you need to

a. effectively *summarize* the data with words, numbers and/or visually and

b. be *explicit* about what the data represents/measures (and doesn't).

Lecture Outline

Descriptive Statistics for and Visualising One Variable

Bivariate Descriptive Statistics and Visualisation

Measurement in Social Science

Introduction to RMarkdown

Conclusion

Motivating Example

Legislative Representation of Women

The proportion of women in legislatures around the world varies substantially. A few legislatures have women represented in proportion to their share of the population, most have fewer or far fewer women than in the population. What are the features of countries that are associated with having more women in (the lower house of) their national legislature?

Descriptive Statistics for and Visualising One Variable

Data set

```r
head(WL_Data)
```

```
##                    Name Code Women_Lower_House GDP_PC_PPP
## 1                 Aruba  ABW                NA         NA
## 2           Afghanistan  AFG              27.7   1951.559
## 3                Angola  AGO              30.5   6440.976
## 4               Albania  ALB              27.9  13325.555
## 5               Andorra  AND              32.1         NA
## 6  United Arab Emirates  ARE              22.5  74942.721
##                         Region         Income_Group
## 1  Latin America & Caribbean          High income
## 2                 South Asia           Low income
## 3         Sub-Saharan Africa Lower middle income
## 4     Europe & Central Asia Upper middle income
## 5     Europe & Central Asia          High income
## 6 Middle East & North Africa          High income
```

Missing Data

As we just saw when we used the `head()` command to look at this data set, there are missing data for some variables for some countries, denoted `NA`

Missing data is (sadly) very common in social science data!

Types of Missing Data

The most commonly used terminology is from survey research:

- Unit non-response
  - A person who refused to participate in your survey
  - In our example, a country whose information is absent from the data set
- Item non-response
  - A single survey question that a particular person did not respond to
  - In our example, a country for which we are missing a particular variable

R takes missing data seriously...

To calculate a mean or standard deviation on a variable with missing data, you have to use the ,na.rm=TRUE option to tell R to remove/ignore the missing values.

```
mean(WL_Data$Women_Lower_House)
```

```
## [1] NA
```

```
mean(WL_Data$Women_Lower_House,na.rm=TRUE)
```

```
## [1] 21.78601
```

How many missing values are there in these data?

```
table(is.na(WL_Data$Women_Lower_House))
```

```
##
## FALSE  TRUE
##   193    24
```

```
table(is.na(WL_Data$GDP_PC_PPP))
```

```
##
## FALSE  TRUE
##   183    34
```

Which are the missing values?

```
WL_Data$Name[is.na(WL_Data$Women_Lower_House)]
```

```
##  [1] Aruba                     American Samoa
##  [3] Bermuda                   Channel Islands
##  [5] Curacao                   Cayman Islands
##  [7] Faroe Islands             Gibraltar
##  [9] Greenland                 Guam
## [11] Hong Kong SAR, China      Isle of Man
## [13] Macao SAR, China          St. Martin (French part)
## [15] Northern Mariana Islands  New Caledonia
## [17] Puerto Rico               West Bank and Gaza
## [19] French Polynesia          Sint Maarten (Dutch part)
## [21] Turks and Caicos Islands  British Virgin Islands
## [23] Virgin Islands (U.S.)     Kosovo
## 217 Levels: Afghanistan Albania Algeria American Samoa Andorra ... Z
```

Which are the missing values?

```
WL_Data$Name[is.na(WL_Data$GDP_PC_PPP)]
```

```
##  [1] Aruba                   Andorra
##  [3] American Samoa          Bahamas, The
##  [5] Bermuda                 Barbados
##  [7] Channel Islands         Cuba
##  [9] Curacao                 Cayman Islands
## [11] Cyprus                  Djibouti
## [13] Eritrea                 Faroe Islands
## [15] Gibraltar               Greenland
## [17] Guam                    Isle of Man
## [19] Iran, Islamic Rep.      Liechtenstein
## [21] St. Martin (French part) Monaco
## [23] Northern Mariana Islands New Caledonia
## [25] Korea, Dem. People's Rep. French Polynesia
## [27] San Marino              Somalia
## [29] South Sudan             Sint Maarten (Dutch part)
## [31] Syrian Arab Republic    Venezuela, RB
## [33] British Virgin Islands  Virgin Islands (U.S.)
```

Strategies for working with missing data

- Deletion
  - Partial deletion: drop units only when missing a variable needed in a specific analysis
  - Listwise deletion: drop units that have any variable missing from all analyses
- There are a variety of more advanced approaches involving some form of "imputation"
  - Advantage: more observations, does not redefine the sample/population under study to those without missing data.
  - Disadvantage: requires "making up" the missing data

We will proceed from here using listwise deletion

```r
dim(WL_Data)
```

```
## [1] 217   6
```

```r
# remove observations with any missing data
WL_Data <- na.omit(WL_Data)

dim(WL_Data)
```

```
## [1] 177   6
```

Measures of Central Tendency Revisited

```r
mean(WL_Data$Women_Lower_House)
```

```
## [1] 21.68362
```

Measures of Central Tendency Revisited

```
mean(WL_Data$Women_Lower_House)
```

[1] 21.68362

The average percentage women in lower houses in countries in the data set is 21.7%.

```
median(WL_Data$Women_Lower_House)
```

[1] 20

Measures of Central Tendency Revisited

```r
mean(WL_Data$Women_Lower_House)
```

```
## [1] 21.68362
```

The average percentage women in lower houses in countries in the data set is 21.7%.

```r
median(WL_Data$Women_Lower_House)
```

```
## [1] 20
```

Half of countries in the data set have more than 20% women in their lower house.

Beyond the Central Tendency

What about the rest of the data?

This is not the only potential summary we might care about.

We may be interested in how the data is *distributed around* the measures of central tendency.

Variable Summaries

R will provide generic summaries of any kind of variable

```r
summary(WL_Data$Women_Lower_House) # numeric variable
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.30   20.00   21.68   30.50   61.30
```

```r
summary(WL_Data$Income_Group) # factor variable
```

```
##           Low income Lower middle income Upper middle income
##                   26                  45                  55
##          High income
##                   51
```

You can also get summaries of all of the variables in a data frame at once with summary(WL_Data)

Quantiles

- The `summary()` command provided the 0th, 25th, 50th (median), 75th and 100th percentile.
    - The Xth "percentile" (0-100) is the value below which X percent of observations in a group of observations fall.
    - The Xth "quantile" (0-1) is the value below which X proportion of observations in a group of observations fall.
- You can request any specific quantile of the distribution you want
    - The 5th, 50th and 95th percentiles are the 0.05, 0.5 and 0.95 quantiles, respectively:

```
quantile(WL_Data$Women_Lower_House,c(0.05,0.5,0.95))
```

```
##    5%   50%   95%
##  5.18 20.00 42.06
```

Etc

```r
range(WL_Data$Women_Lower_House)
```

```
## [1]  0.0 61.3
```

```r
max(WL_Data$Women_Lower_House)
```

```
## [1] 61.3
```

```r
min(WL_Data$Women_Lower_House)
```

```
## [1] 0
```

Summarising Dispersion

How do we summarise *variation* or *dispersion* in a data set?

How do we describe the difference between these distributions, all of which have the same mean/median?

- 20, 20, 20, 20, 20
- 15, 18, 20, 22, 25
- 10, 15, 20, 25, 30
- 0, 10, 20, 30, 40

Range

One measure of dispersion is the range, the difference between the
0th and 100th percentile (or equivalently, the 0.00 and 1.00
quantile)

- **20**, 20, 20, 20, **20** $\rightarrow$ 0
- **15**, 18, 20, 22, **25** $\rightarrow$ 10
- **10**, 15, 20, 25, **30** $\rightarrow$ 20
- **0**, 10, 20, 30, **40** $\rightarrow$ 40

This turns out to be a poor choice in practice, because it is very
sensitive to changes in a single extreme value.

Interquartile Range

A better measure of dispersion is the interquartile range, the difference between the 25th and 75th percentile (or equivalently, the 0.25th and 0.75th quantile)

- 20, **20**, 20, **20**, 20 $\rightarrow$ 0
- 15, **18**, 20, **22**, 25 $\rightarrow$ 4
- 10, **15**, 20, **25**, 30 $\rightarrow$ 10
- 0, **10**, 20, **30**, 40 $\rightarrow$ 20

Interquartile Range

```r
q <- quantile(WL_Data$Women_Lower_House,c(0.25,0.75))
q
```

```
##  25%  75%
## 12.3 30.5
```

```r
q[2] - q[1]
```

```
##  75%
## 18.2
```

Standard deviation

While the interquartile range is easy to understand, far more frequently we use the *standard deviation* as a measure of dispersion.

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

Standard deviation

While the interquartile range is easy to understand, far more frequently we use the *standard deviation* as a measure of dispersion.

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

Building the standard deviation: Start with the difference between each observation and the mean:

$$(x_i - \overline{x})$$

Standard deviation

While the interquartile range is easy to understand, far more frequently we use the *standard deviation* as a measure of dispersion.

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

Building the standard deviation: Square these so that bigger deviations from the mean (whether positive or negative) are more positive numbers

$$(x_i - \overline{x})^2$$

Standard deviation

While the interquartile range is easy to understand, far more frequently we use the *standard deviation* as a measure of dispersion.

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

Building the standard deviation: Add them all up

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

Standard deviation

While the interquartile range is easy to understand, far more frequently we use the *standard deviation* as a measure of dispersion.

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

Building the standard deviation: Divide by the number of things you added up (minus one, for technical reasons that need not concern us here).

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

Standard deviation

While the interquartile range is easy to understand, far more frequently we use the *standard deviation* as a measure of dispersion.

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

Building the standard deviation: Finally, take a square root to "undo" the square, so that the units of the standard deviation are on the original scale of x.

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

Standard deviation

While the interquartile range is easy to understand, far more
frequently we use the *standard deviation* as a measure of
dispersion.

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

Interpreting the standard deviation:

- Larger standard deviations mean that the data are more
  dispersed / variable.
- Smaller standard deviations mean that the data are less
  dispersed / variable.
  - A standard deviation of 0 occurs only if all the observations
    have the same value.

Example of Standard Deviation

```r
sd(WL_Data$Women_Lower_House)
```

```
## [1] 11.98296
```

Or you could use R to implement the formula

```r
x <- WL_Data$Women_Lower_House
N <- length(x)

sqrt(((1)/(N-1))*sum((x - mean(x))^2))
```

```
## [1] 11.98296
```

Dispersion in data
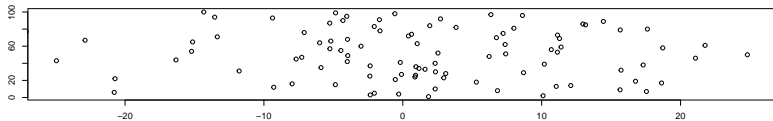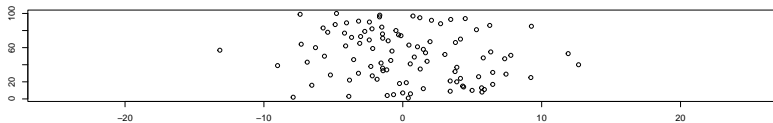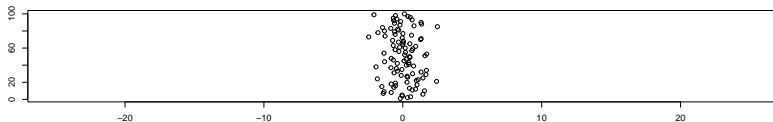
Sometimes it is useful to "standardize" variables in terms of their standard deviation.

- This is typically done by calculating "z-scores"
- This preserves the relative values of x, but the new variable z then has a mean of exactly 0 and a standard deviation of exactly 1

$$z_i = \frac{x_i - mean(x)}{sd(x)}$$

When to use different types of statistics

- Means and standard deviations only make sense if the magnitudes of differences are meaningful.
- Medians, quantiles, ranges only make sense if the different values of a variable have a meaningful ordering.

When to use different types of statistics

- Means and standard deviations only make sense if the magnitudes of differences are meaningful.
- Medians, quantiles, ranges only make sense if the different values of a variable have a meaningful ordering.

As a consequence, which we should use depends on the level of measurement of the variables we are describing:

- **Continuous/Interval**: describe distributions of such variables using any of the statistics
- **Ordinal**: Means and standard deviations are potentially misleading, because variables lack meaningful information about magnitudes of differences.

Visualising Data

Eyes are Powerful Tools!

- Our goal is to communicate quantitative information to humans.
- Humans have eyes/brains that are good at processing information about size, distance and colour.
- Surely we can use this?

This requires you to remember and compare numbers

Var1	Freq
Low income	26
Lower middle income	45
Upper middle income	55
High income	51

Barplot

```
barplot(table(WL_Data$Income_Group))
```
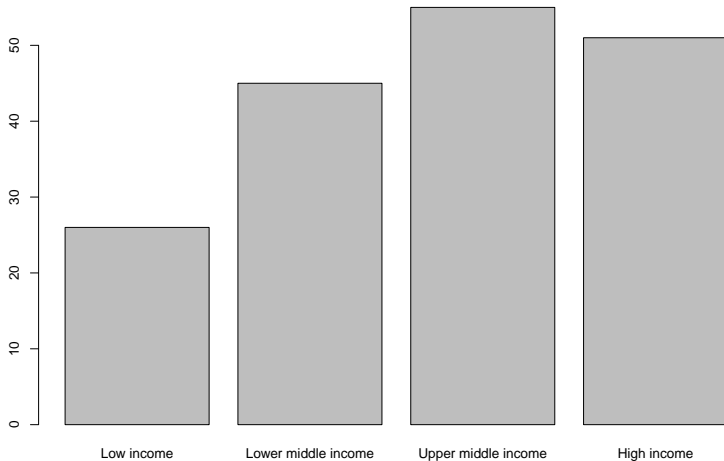
Numerical values for a continuous variable are difficult to read

```
##   [1]  0.0  0.0  0.0  0.0  1.2  2.5  3.1  4.0  4.7  5.3  5.6
##  [12]  5.8  5.9  6.1  6.5  6.7  7.2  7.2  7.4  7.5  8.5  8.6
##  [23]  8.8  8.9  9.1  9.1  9.4  9.5  9.8 10.0 10.1 10.2 10.3
##  [34] 10.5 10.6 10.7 11.0 11.1 11.3 11.6 11.8 11.9 12.3 12.3
##  [45] 12.3 12.5 12.6 12.7 12.7 13.0 13.3 13.7 13.9 14.9 15.0
##  [56] 15.3 15.4 15.8 16.0 16.0 16.0 16.0 16.0 16.7 16.7 16.8
##  [67] 17.0 17.0 17.1 17.1 17.4 17.5 17.6 18.0 18.1 18.1 18.2
##  [78] 18.3 18.5 18.7 19.0 19.2 19.2 19.6 19.8 19.9 20.0 20.0
##  [89] 20.0 20.2 20.3 20.3 20.5 20.6 20.7 21.1 21.2 21.3 21.4
## [100] 21.8 21.9 22.0 22.1 22.2 22.5 22.6 22.8 23.0 23.5 23.6
## [111] 23.8 24.4 24.8 24.9 25.0 25.5 25.5 25.8 26.7 26.7 26.8
## [122] 27.0 27.1 27.5 27.5 27.7 27.7 27.9 28.0 28.3 28.7 29.5
## [133] 30.5 30.5 30.7 31.0 31.0 31.1 31.3 31.5 31.9 32.2 32.5
## [144] 32.7 33.8 34.3 34.4 34.4 34.5 34.8 35.7 36.0 36.4 37.2
## [155] 37.4 38.0 38.0 38.1 38.3 38.3 38.8 38.9 39.1 39.6 39.6
## [166] 41.4 41.8 42.0 42.3 45.6 45.7 46.1 46.2 46.7 48.2 53.1
## [177] 61.3
```

Histogram

```r
hist(WL_Data$Women_Lower_House,
     main="", #title of the plot
     xlab="Percent Women in Lower House") # x-axis label
```
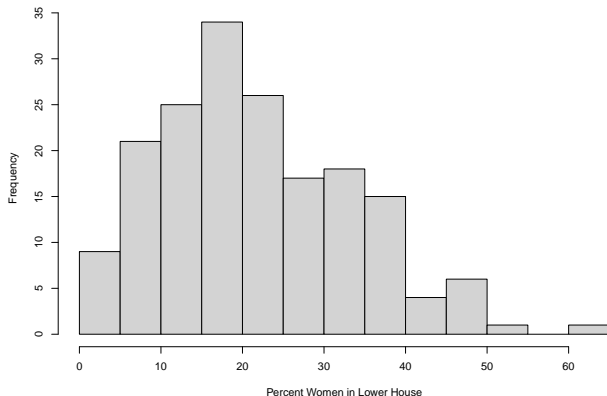
Histogram of z-scores

```r
hist(scale(WL_Data$Women_Lower_House), main="",
     xlab="Z-Scores for Percent Women in Lower House")
```


Z–Scores for Percent Women in Lower House
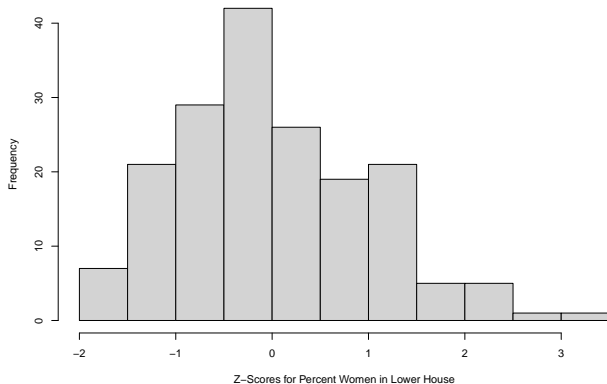
Barplots and Histograms

Barplots (for categorical variables) and Histograms (for continuous variables) indicate the relative frequency of different levels of a variable.

They are effective visualizations because they map a quantity (relative frequency) onto something visual (the size of the bar) and then faciliate comparisons

(Barplots and histograms **must** have y-axes that start at zero to avoid being visually misleading. This is the only way to make the size of the bar proportionate to the numeric value of the frequency.)

Misleading Barcharts
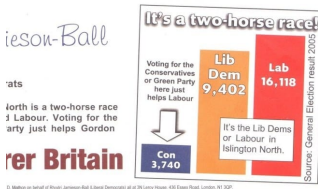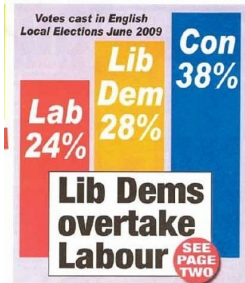
Boxplots

```
boxplot(WL_Data$Women_Lower_House,
        ylab="Percent Women in Lower House") # y-axis label
```
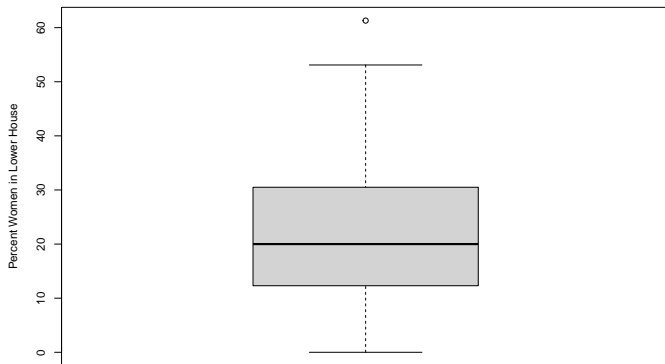
A boxplot displays...

- The median (horizontal line in middle of "box")
- The interquartile range (the top and bottom of the box)
- The range (top and bottom of the tails/"whiskers")
- Conventionally, points beyond 1.5 IQRs from the median are displayed individually.

Comparison
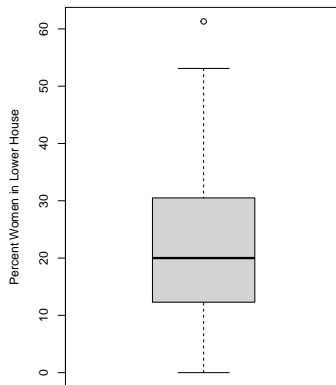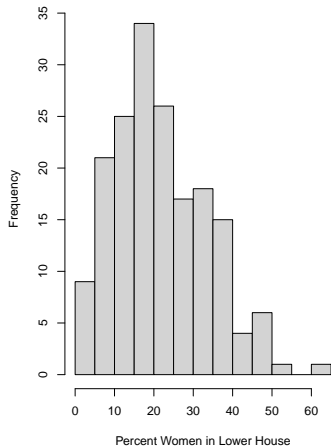
What information is better communicated by the histogram?

Break

Bivariate Descriptive Statistics and Visualisation

Beyond one variable

Very few social science questions can be answered by looking at just one variable

We almost always want to make some kind of comparisons.

Visualisation can be a way to do this, but you have to think carefully about what comparisons you want to facilitate…

Boxplots for comparing multiple groups

```
boxplot(Women_Lower_House~Income_Group,data=WL_Data,
        ylab="Percent Women in Lower House")
```
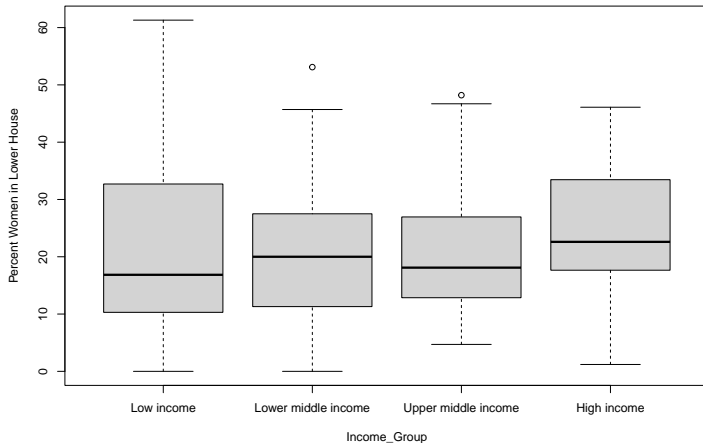
Comparisons with histograms are more difficult to make

Low income
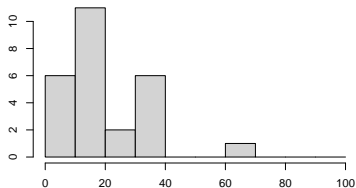
Percent Women in Lower House

Lower middle income
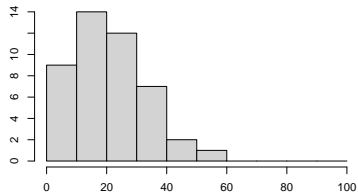
Percent Women in Lower House

Upper middle income
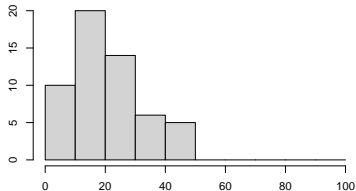
Percent Women in Lower House

High income
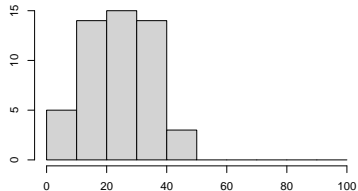
Percent Women in Lower House

Scatterplots

```
plot(WL_Data$GDP_PC_PPP,WL_Data$Women_Lower_House,
     xlab="Per Capita GDP", # x-axis label
     ylab="Percent Women in Lower House") # y-axis label
```
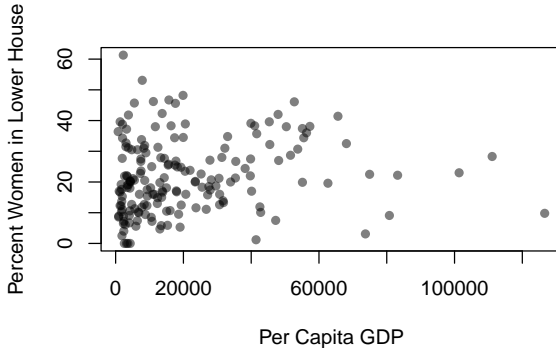
Per capita GDP is highly skewed

```
hist(WL_Data$GDP_PC_PPP, xlab="Per Capita GDP",main="")
```
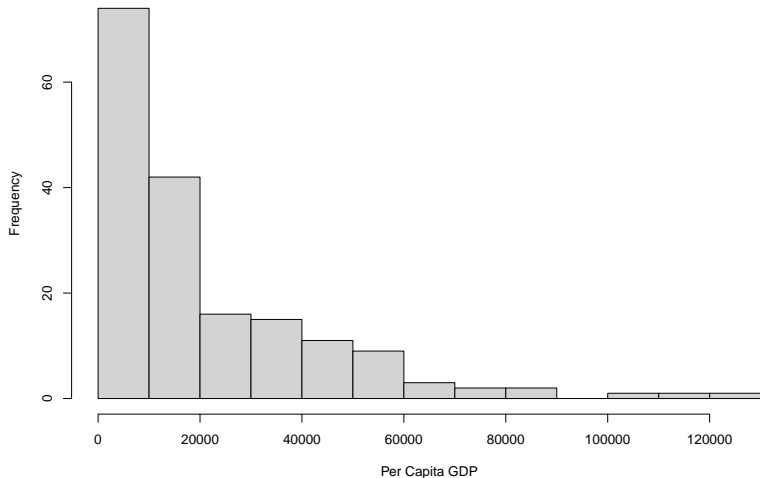
Try to choose meaningful ways to present comparisons

Visual problem: the per capita GDP variable is highly skewed, lots of very small values, a few very large ones.

Substantive problem: it is also the case that a $1000 difference in per capita GDP is a lot more important between $1000 and $2000 than it is between $10,000 and $11,000

It is perhaps more sensible to think about, and visually present, per capita GDP in a way that reflects ratios rather than differences.

- That is, perhaps a per capita GDP of $1000 and $2000 should be *visually* as far apart as $10,000 and $20,000.

Log transformation of variable

The way to translate common ratios ($2000 vs $1000 and $20,000 vs $10,000) into common differences is via a log transformation:

```
log(2000) - log(1000)
```

```
## [1] 0.6931472
```

```
log(20000) - log(10000)
```

```
## [1] 0.6931472
```

Note: these are "natural" logs, also known as log base e, frequently denoted "ln" rather than "log" on calculators. There is an extra slide at the end with some mathematical properties of logarithms.

Scatterplot after log transformation of per capita GDP

```
plot(WL_Data$GDP_PC_PPP,
     WL_Data$Women_Lower_House,
     xlab="Per Capita GDP",
     ylab="% Women in Lower House")
```
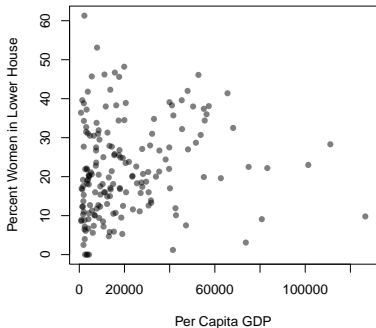
```
plot(log(WL_Data$GDP_PC_PPP),
     WL_Data$Women_Lower_House,
     xlab="Log Per Capita GDP",
     ylab="% Women in Lower House")
```
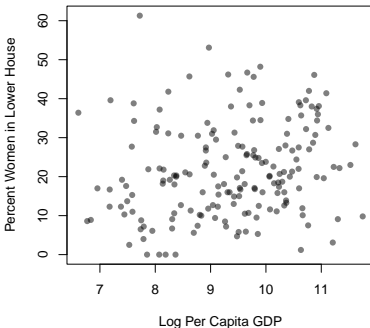
Correlation coefficient

The most commonly used bivariate descriptive statistic is the *correlation coefficient*:

$$\rho = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{S_x} \times \frac{y_i - \bar{y}}{S_y} \right)$$

The correlation coefficient describes the strength of association between two continuous variables.

Note: ρ is pronounced "row", and spelled "rho".

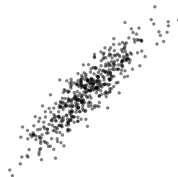# Examples of correlation coefficients
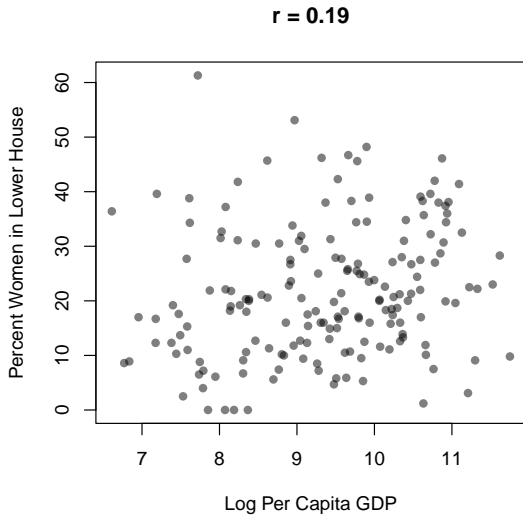

rho = -1   rho = 0   rho = 1

rho = 0.5   rho = 0.75   rho = 0.9

In Our Example

r = 0.19

Percent Women in Lower House vs. Log Per Capita GDP
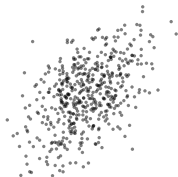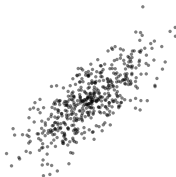
What is a large correlation?

- 0.19 is definitely not a "large" correlation, but what would be?
    - $\rho = 1$ is a perfect positive association
    - $\rho = 0$ is no association
    - $\rho = -1$ is a perfect negative association.
- There is no fixed translation of numbers to words that applies generally
    - In some contexts $\rho = 0.5$ is "strong", in others it is "moderate", in others it is "weak".

Measurement in Social Science

On the difficulties of social measurement

In social science, many of the concepts we are interested in are not directly measurable (e.g. 'social capital', 'democracy' etc).

We choose our variables because we assume they tell us something about (i.e. measure) a thing we are interested in.

But, crucially, the measures we choose are more often than not *imperfect* representations of the concepts we are interested in.

For example

- With the *share of women in the (lower house of the) legislature*, we may be trying to measure the amount of **political power women have** in a given country.
- With *GDP per capita*, we may be trying to measure the level of **economic development** of a given country.

For example

- With the *share of women in the (lower house of the) legislature*, we may be trying to measure the amount of **political power women have** in a given country.
- With *GDP per capita*, we may be trying to measure the level of **economic development** of a given country.

Have a think and discuss: are there reasons to believe that theses measures are imperfect representations of the target concepts?

Measurement inference

Last week we saw that *causal inference* is hard. *Measurement inference* is also hard.

Measurement inference

Last week we saw that *causal inference* is hard. *Measurement inference* is also hard.

Causal inference

= Conclusions drawn from observed data about what we would have observed for *the same units given counterfactual circumstances*.

Measurement inference

Last week we saw that *causal inference* is hard. *Measurement inference* is also hard.

Causal inference

= Conclusions drawn from observed data about what we would have observed for *the same units given counterfactual circumstances*.

Measurement inference

= Conclusions drawn from observed data about *unmeasured, underlying features that characterise the same units*.

A pragmatic approach

- *Accept* that (social science) measures are most of the time just a partial approximation of the concepts of interest.

- *Reflect* on the ways in which a given measure is better or worse than an alternative and on the ways in which it could be improved.

- When making claims based on observed data, *be explicit* about the advantages and disadvantages of a chosen measure.

Don't say things like this

Opinion: The media treats Biden as badly as — or worse than — Trump. Here's proof.

President Biden speaks at the White House on Dec. 3. (Demetrius Freeman/The Washington Post)

Don't say things like this

Artificial intelligence can now measure the negativity with precision. At my request, Forge.ai, a data analytics unit of the information company FiscalNote, combed through more than 200,000 articles — tens of millions of words — from 65 news websites (newspapers, network and cable news, political publications, news wires and more) to do a "sentiment analysis" of coverage. Using algorithms that give weight to certain adjectives based on their placement in the story, it rated the coverage Biden received in the first 11 months of 2021 and the coverage President Donald Trump got in the first 11 months of 2020.

The findings, painstakingly assembled by FiscalNote vice president Bill Frischling, confirmed my fear: My colleagues in the media are serving as accessories to the murder of democracy.

Introduction to RMarkdown

Why R Markdown?

- It allows you to create formatted documents directly from RStudio

- It allows you to include (R) code, plots and tables produced with R easily (without having to copy-paste!)

- More specifically, from this week onward, we are asking you to submit your homework as a (well) *formatted word or pdf document* with typed out answers - RMarkdown makes that easier
    - Note though that you don't **have** to learn RMarkdown, you could also just write your answers in word and copy paste your *code* in there.

- And we would lie if we didn't say it looks incredibly cool
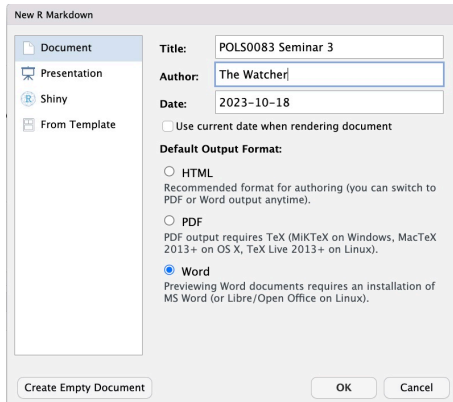
Getting started

1. If not already done, install the R Markdown package (as with any other package in R). In the console, type: install.packages("rmarkdown").

2. Once this is all installed, you can create a new R Markdown file from the "file" menu, shown here:
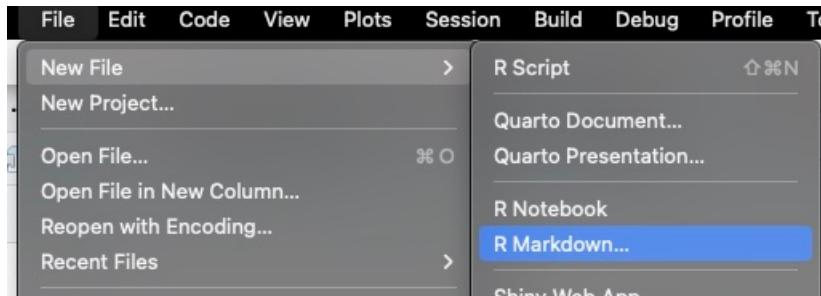
Getting started

3. Enter a title – for instance, "Seminar week 3" – and your name, as shown below.
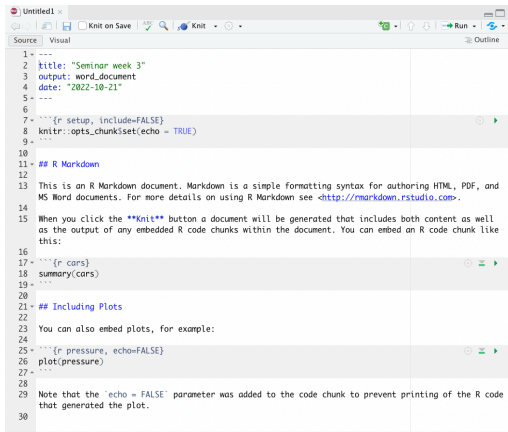4. Decide whether we want to produce a HTML, PDF, or Word file. In this class, we ask that you select the Word option.

Getting started

5. A new document will open in the top left pane of RStudio, with some default text introducing R markdown functionalities (that you can/should delete!)

Getting started

6. Finally, to compile a (Word) document from your markdown file, hit "Knit" at the top of your screen.
   - Note that, if this is the first time you knit the document, your computer will first ask you where you want to save it and under what name.

Writing in R Markdown

When using R Markdown, the interface very usefully integrates together words and our R code.

You can use the "white" spaces to write as you would in any word document.

Although there is some coding syntax needed for formatting text. For instance, this is how you create headers:

```
# Header 1
## Header 2
### Header 3
etc
```

More R Markdown Syntax

```
*some stuff in italics*
**some stuff in bold**
```

some stuff in italics
some stuff in bold

```
- a bullet point
    - bullet sub-point
1. a numbered point
    a. lettered sub-point
```

- a bullet point
    - bullet sub-point
1. a numbered point
    a. lettered sub-point

```
Something said[^footnote]
[A link](https://youtu.be/LLCh_b8Lu20)

[^footnote]: This is fun!
```

Something said[a]
A link (click it!)

---
[a]This is fun!

Coding in R Markdown

There are two ways to include code into your document.

1. Inline

```
Put some code in here `r ` and then continue typing
```

2. As a 'chunk'

```
```{r}
x<-5
x+2
```
```

Conclusion

What have we learned about women's representation in legislatures?

By exploring this data set, we have learned that…

- The typical proportion of women in national legislatures around the world is low (mean = 21.7%, median 20%)

- There is substantial variation across countries (sd = 12 percentage points)

- There are only a very few countries at or near 50% women (eg Rwanda, Bolivia, Mexico, Sweden)

- Wealthier/more developed countries, *as measured by GDP per capita*, have slightly higher proportions of women on average, but differences by income are small compared to variation within income levels across countries.

What have we learned about descriptive statistics?

- We have just covered the most commonly used "descriptive" statistics
    - central tendancy: mean, median, mode
    - dispersion: standard deviation, quantiles / interquartile range
    - association: correlation coefficient
- There are many more statistics of these types, for various types of data
- There are many other features of data that people have developed statistics to describe

What have we learned about measurement?

- Measurement is about representing a feature / characteristic of a unit.
- Especially within the social science, the features / characteristics (i.e. concepts) we are interested in, are hard to summarise in simple measures.
- Virtually all measures we could choose are imperfect and have pros and cons - which doesn't mean we should give up!
- It is important to be *aware* and *explicit* about the pitfalls of measurement.

Descriptive statistics and visualisation are for communication

- You know not to write badly,…
- …don't summarise your data badly…
  - Try to think about which features of the data are relevant to the substantive question that motivated your analysis, and communicate relevant statistics for those.
  - Try not to provide irrelevant or misleading information.

Descriptive statistics and visualisation are for communication

- You know not to write badly,…

- …don't summarise your data badly…
  - Try to think about which features of the data are relevant to the substantive question that motivated your analysis, and communicate relevant statistics for those.
  - Try not to provide irrelevant or misleading information.

- …and don't plot badly either.
  - Try to think about what you are trying to communicate, and make sure your visualisation facilitates the relevant comparison: choose easily distinguishable colors, nice labels, etc.
  - Try not to provide irrelevant or misleading visualisation.

Seminar

In seminars this week, you will learn about …

1. … calculating variable summaries, tables and correlations
2. … working with (more) missing data
3. … scatterplots, boxplots, histograms

Extra: Properties of logarithms and exponential functions

- $e = 2.71828\ldots$ (the second most famous transcendental number)
- $\log(e) = 1$
- $\log(1) = 0$
- $\log(x^r) = r\ \log(x)$
- $\log(e^A) = A$
- $e^{\log A} = A$
- $\log(AB) = \log A + \log B$
- $\log(A/B) = \log A - \log B$
- $e^{AB} = \left(e^A\right)^B$
- $e^{A+B} = e^A e^B$
- $e^{A-B} = e^A / e^B$