Data Visualization

Cohen Chapter 2

EDUC/PSY 6600

Always plot your data first!

"Always." - Severus Snape

Always plot your data first!

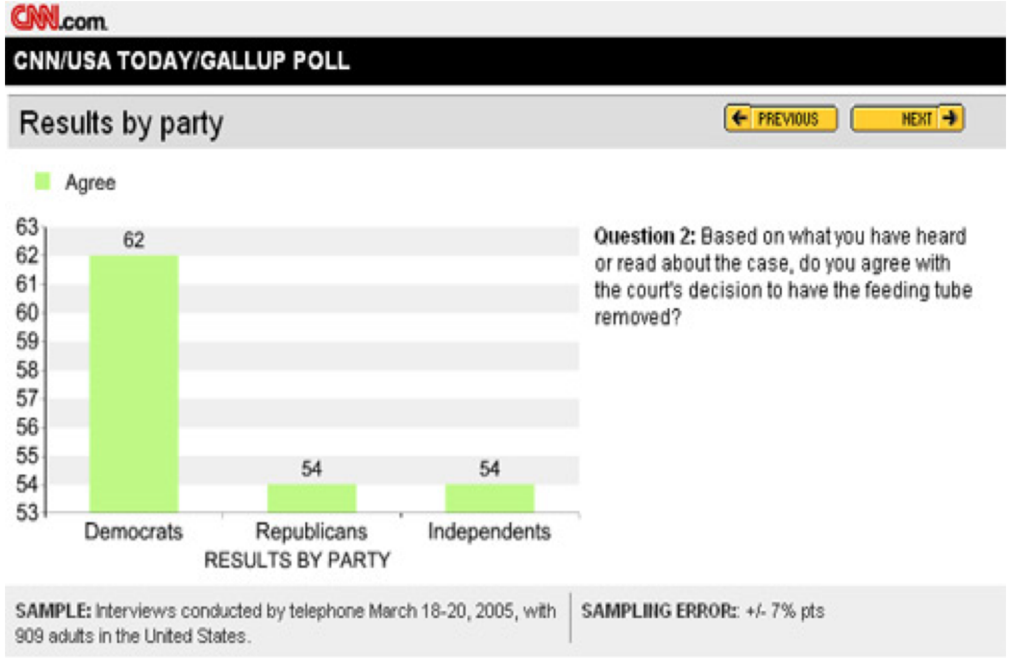
"Always." - Severus Snape

Why?

- Outliers and impossible values
- Determine correct statistical approach
- Assumptions and diagnostics
- Discover new relationships

The Visualization Paradox

- Often the most informative aspect of analysis
- Communicates the "data story" the best
- Most abused area of quantitative science
- Figures can be *very* misleading



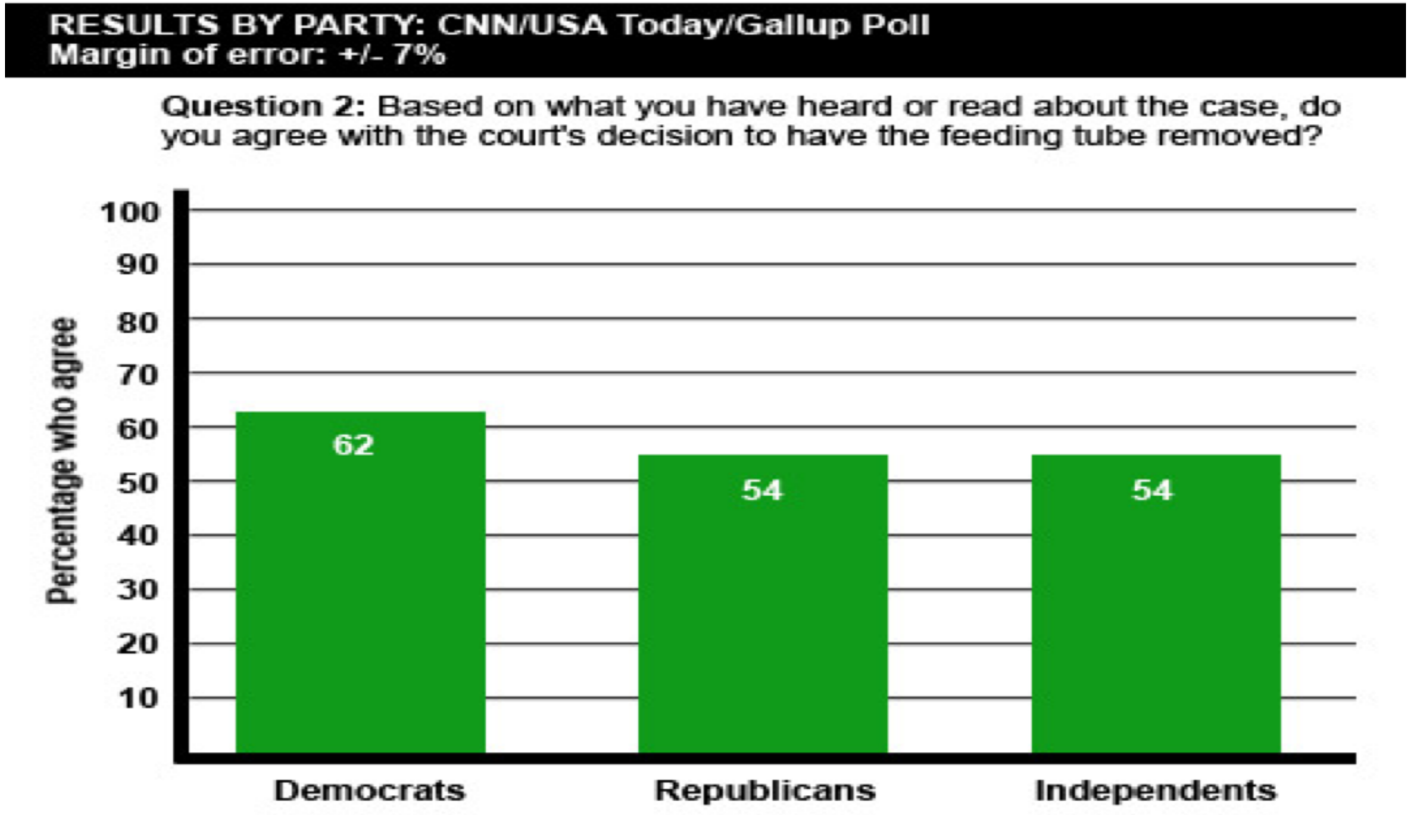
Question 2: Based on what you have heard or read about the case, do you agree with the court's decision to have the feeding tube removed?

SAMPLE: Interviews conducted by telephone March 18-20, 2005, with 909 adults in the United States.

SAMPLING ERROR: +/- 7% pts

Misleading Graphs

Much better


RESULTS BY PARTY: CNN/USA Today/Gallup Poll
Margin of error: +/- 7%
Question 2: Based on what you have heard or read about the case, do you agree with the court's decision to have the feeding tube removed?
Percentage who agree
100
90
80
70
60
50
40
30
20
10
62
54
54
Democrats
Republicans
Independents

Keys to Good Viz's

- Graphical method should match level of measurement
- Label all axes and include figure caption
- Simplicity and clarity
- Avoid of ‘chartjunk’

Keys to Good Viz's

- Graphical method should match level of measurement
- Label all axes and include figure caption
- Simplicity and clarity
- Avoid of 'chartjunk'
- Unless there are 3 or more variables, avoid 3D figures (and even then, avoid it)
- Black & white, grayscale/pattern fine for most simple figures

Data Visualizations

Takes practice -- try a bunch of stuff

Data Visualizations

Takes practice -- try a bunch of stuff

Resources

- Edward Tufte's books
- "R for Data Science" by Grolemund and Wickham
- "Data Visualization for Social Science" by Healy

Frequency Distributions

Counting the number of occurrences of unique events

- Categorical or continuous
- just like with `tableF()` and `table1()`

Can see central tendency (continuous data) or most common value (categorical data)

Can see range and extremes

x	Freq	CumFreq	Percent	CumPerc	Valid	CumValid
1	233	233	23.30%	23.30%	23.92%	23.92%
2	265	498	26.50%	49.80%	27.21%	51.13%
3	232	730	23.20%	73.00%	23.82%	74.95%
4	244	974	24.40%	97.40%	25.05%	100.00%
Missing	26	1000	2.60%	100.00%		

Frequencies and Viz's Together ❤️

Bar Graph

Frequencies and Viz's Together ❤️

Bar Graph

Histogram

What does DISTRIBUTION mean?

The way that the data points are scattered

What does DISTRIBUTION mean?

The way that the data points are scattered

For Continuous

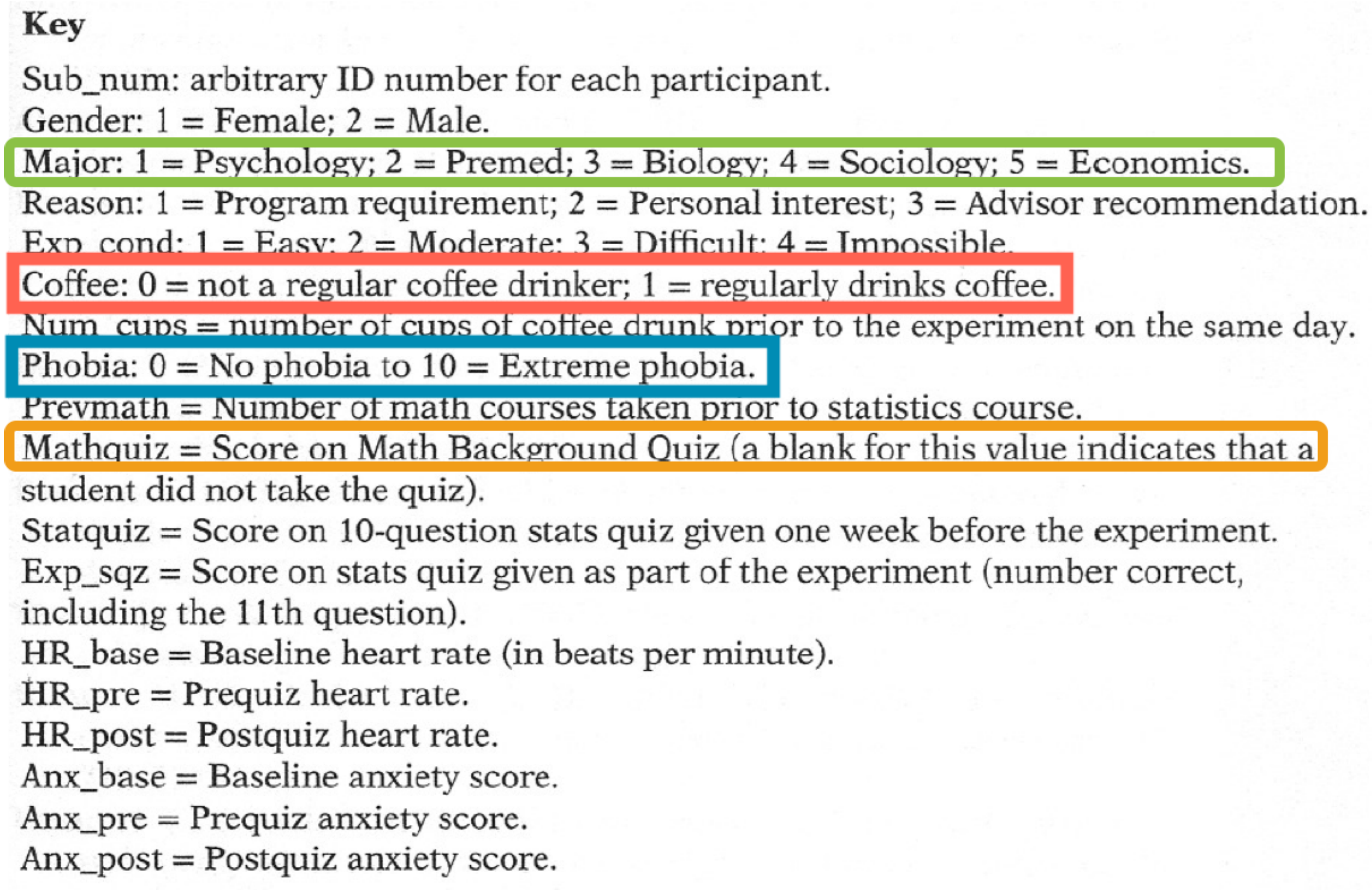
- General shape
- Exceptions (outliers)
- Modes (peaks)
- Center & spread (chap 3)
- Histogram

For Categorical

- Counts of each
- Percent or Rate (adjusts for an ‘out of’ to compare)
- Bar chart
- Pie chart - avoid!

Let's Apply This To the Inho Dataset

Reminder



Key

Sub_num: arbitrary ID number for each participant.
Gender: 1 = Female; 2 = Male.
Major: 1 = Psychology; 2 = Premed; 3 = Biology; 4 = Sociology; 5 = Economics.
Reason: 1 = Program requirement; 2 = Personal interest; 3 = Advisor recommendation.
Exp_cond: 1 = Easy; 2 = Moderate; 3 = Difficult; 4 = Impossible.
Coffee: 0 = not a regular coffee drinker; 1 = regularly drinks coffee.
Num_cups = number of cups of coffee drunk prior to the experiment on the same day.
Phobia: 0 = No phobia to 10 = Extreme phobia.
Prevmath = Number of math courses taken prior to statistics course.
Mathquiz = Score on Math Background Quiz (a blank for this value indicates that a student did not take the quiz).
Statquiz = Score on 10-question stats quiz given one week before the experiment.
Exp_sqz = Score on stats quiz given as part of the experiment (number correct, including the 11th question).
HR_base = Baseline heart rate (in beats per minute).
HR_pre = Prequiz heart rate.
HR_post = Postquiz heart rate.
Anx_base = Baseline anxiety score.
Anx_pre = Prequiz anxiety score.
Anx_post = Postquiz anxiety score.

Read in the Data

```
library(tidyverse)   # the easy button
library(readxl)      # read in Excel files
library(furniture)   # nice tables

data_raw <- readxl::read_excel("Ihno_dataset.xls") %>%
  dplyr::rename_all(tolower)                   # converts all variable names to lower case
```

Read in the Data

```
library(tidyverse)   # the easy button
library(readxl)      # read in Excel files
library(furniture)   # nice tables

data_raw <- readxl::read_excel("Ihno_dataset.xls") %>%
  dplyr::rename_all(tolower)                    # converts all variable names to lower case
```

And Clean It

```
data_clean <- data_raw %>%
  dplyr::mutate(majorF = factor(major,
                                levels= c(1, 2, 3, 4, 5),
                                labels = c("Psychology", "Premed",
                                           "Biology", "Sociology",
                                           "Economics"))) %>%
  dplyr::mutate(coffeeF = factor(coffee,
                                 levels = c(0, 1),
                                 labels = c("Not a regular coffee drinker",
                                            "Regularly drinks coffee")))
```

Frequency Distrubutions

```
data_clean %>%
  furniture::tableF(major)
```

```
##
## ────────────────────────────────────
##  major Freq CumFreq Percent CumPerc
##  1     29   29      29.00%  29.00%
##  2     25   54      25.00%  54.00%
##  3     21   75      21.00%  75.00%
##  4     15   90      15.00%  90.00%
##  5     10   100     10.00%  100.00%
## ────────────────────────────────────
```

```
data_clean %>%
  furniture::tableF(phobia)
```

```
##
## ─────────────────────────────────────
##  phobia Freq CumFreq Percent CumPerc
##  0      12   12      12.00%  12.00%
##  1      15   27      15.00%  27.00%
##  2      12   39      12.00%  39.00%
##  3      16   55      16.00%  55.00%
##  4      21   76      21.00%  76.00%
##  5      11   87      11.00%  87.00%
##  6      1    88      1.00%   88.00%
##  7      4    92      4.00%   92.00%
##  8      4    96      4.00%   96.00%
##  9      1    97      1.00%   97.00%
##  10     3    100     3.00%   100.00%
## ─────────────────────────────────────
```

Frequency Viz's

For viz's, we will use `ggplot2`

This provides the most powerful, beautiful framework for data visualizations

Frequency Viz's

For viz's, we will use `ggplot2`

This provides the most powerful, beautiful framework for data visualizations

- It is built on making layers
- Each plot has a "geom" function
  - e.g. `geom_bar()` for bar charts, `geom_histogram()` for histograms, etc.

Bar Charts

```
data_clean %>%
  ggplot(aes(major)) +
  geom_bar()
```

Bar Charts

```
data_clean %>%
  ggplot(aes(coffee)) +
  geom_bar()
```

Histograms

```
data_clean %>%
  ggplot(aes(phobia)) +
  geom_histogram()
```

Histograms (change number of bins)

```
data_clean %>%
  ggplot(aes(phobia)) +
  geom_histogram(bins = 8)
```

Histograms (change bins to size 5)

```
data_clean %>%
  ggplot(aes(phobia)) +
  geom_histogram(binwidth = 5)
```

Histograms

```
data_clean %>%
  ggplot(aes(mathquiz)) +
  geom_histogram(binwidth = 4)
```

Histograms -by- a Factor (columns)

```
data_clean %>%
  ggplot(aes(mathquiz)) +
  geom_histogram(binwidth = 4) +
  facet_grid(. ~ coffeeF)
```

Histograms -by- a Factor (rows)

```
data_clean %>%
  ggplot(aes(mathquiz)) +
  geom_histogram(binwidth = 4) +
  facet_grid(coffeeF ~ .)
```

Deciles (break into 10% chunks)

```r
data_clean %>%
  dplyr::pull(statquiz) %>%
  quantile(probs = c(.10, .20, .30, .40, .50, .60, .70, .80, .90))
```

```
## 10% 20% 30% 40% 50% 60% 70% 80% 90%
## 4.0 6.0 6.0 7.0 7.0 8.0 8.0 8.0 8.1
```

Deciles - with missing values

```
data_clean %>%
  dplyr::pull(mathquiz) %>%
  quantile(probs = c(.10, .20, .30, .40, .50, .60, .70, .80, .90))
```

```
Error in quantile.default(., probs = c(0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, : missing
values and NaN's not allowed if 'na.rm' is FALSE
```

Deciles - `na.rm = TRUE`

```
data_clean %>%
  dplyr::pull(mathquiz) %>%
  quantile(probs = c(.10, .20, .30, .40, .50, .60, .70, .80, .90),
           na.rm =TRUE)
```

```
##  10%  20%  30%  40%  50%  60%  70%  80%  90% 
## 15.0 21.0 25.2 28.0 30.0 32.0 33.8 37.2 41.0
```

Quartiles (break into 4 chunks)

```
data_clean %>%
  dplyr::pull(statquiz) %>%
  quantile(probs = c(0, .25, .50, .75, 1))
```

```
##   0%  25%  50%  75% 100% 
##    1    6    7    8   10
```

Percentiles

```
data_clean %>%
  dplyr::pull(statquiz) %>%
  quantile(probs = c(.01, .05, .173, .90))
```

```
##    1%    5% 17.3%   90%
##  2.98  3.00  5.00  8.10
```

Questions?

Next Topic

Center and Spread