Gov 50: 8. Measurement: Summarizing Bivariate Relationships

Matthew Blackwell

Harvard University

Fall 2018

1. Today's agenda

2. Investigating fraud

3. Bivariate relationships

1/ Today's agenda

Logistics

- Problem set 2:
  - due Thursday by midnight.
  - remember to turn in Rmd and compiled pdf!
  - this time we start to take points off for Rmd files that don't compile.
- Midterm 1:
  - Next Tuesday.
  - Covers material through today.
  - Review session on Thursday.
- Mike's Monday section rescheduled to this Thursday (10/4) at 12pm in CGIS S020.
- Midterm course evaluations after the midterm.

Where are we? Where are going?

- Talked about survey sampling, its problems
- How to summarize a single variable? Mean, median, range, SD.
- Now: how to summarize relationship *between* variables.

- Review 3.5–3.6
- Revisit the gay-marriage experiment:
  - LaCour and Green (2015). “When contact changes minds: An experiment of transmission of support for gay equality.” *Science*, Vol. 346, No. 6215 pp. 1366–1369.
  - Broockman, Kalla, Aronow (2015). “Irregularities in LaCour (2014)”

2/ Investigating fraud

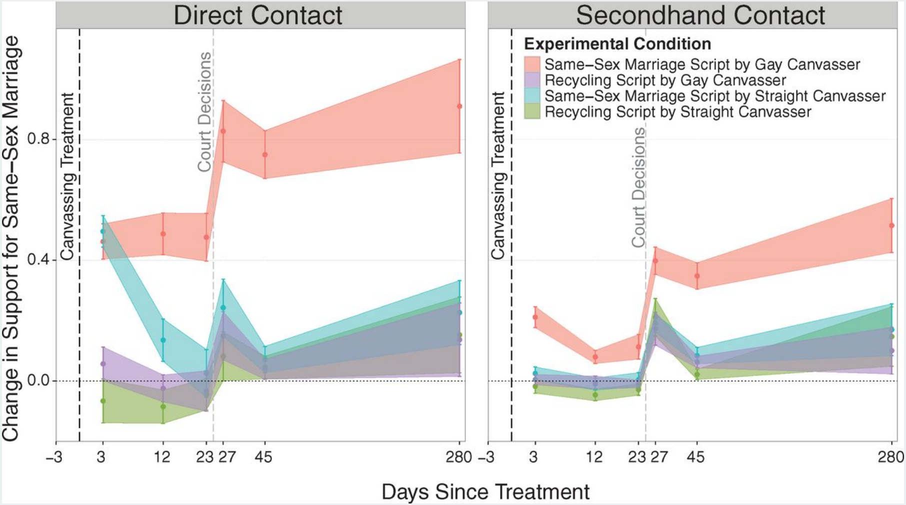
Changing minds on gay marriage

- Question: Can we effectively persuade people to change their minds?
- **Contact Hypothesis**: outgroup hostility diminishes when people from different groups interact with one another.
- Two randomized control trials in Los Angeles
- **Target population**: voters in Los Angeles.
- **Sampling frame**: registered voter list.
  - invited randomly selected voters to participate in an online baseline survey.
  - asked them to refer their friends and families with compensation.
  - those friends and family are also invited to participate in the online baseline survey.
  - panel data: baseline plus 6 waves.

Study design

- Randomized treatment:
  - gay vs. straight canvassers with similar characteristics
  - same-sex marriage vs. recycling scripts (*placebo*)
  - control group: no canvassing
- Persuasion scripts are the same except on important difference:
  - gay canvassers: they would like to get married but law prohibits it.
  - straight canvassers: their gay child, friend, or relative would to get married but the law prohibits it.
- Outcome measures:
  - support for same-sex marriage.
  - feeling toward gay people.

Big and lasting effects of persuasion



Reshaped data

Name	Description
study	Which study is the data from (1 = Study1, 2 = Study2)
treatment	Five possible treatment assignment options
therm1	Survey thermometer rating of feeling towards gay couples in waves 1 (0–100)
therm2	Survey thermometer rating of feeling towards gay couples in waves 2 (0–100)
therm3	Survey thermometer rating of feeling towards gay couples in waves 3 (0–100)
therm4	Survey thermometer rating of feeling towards gay couples in waves 4 (0–100)

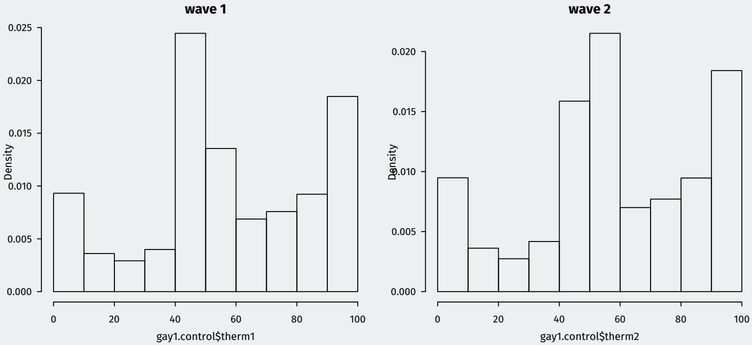
```r
gay.reshaped <- read.csv("data/gayreshaped.csv")
names(gay.reshaped)
```

```
## [1] "study"     "treatment" "therm1"    "therm2"
## [5] "therm3"    "therm4"
```

Comparison of gay thermometer across waves

- Compare between waves 1 and 2 for the control group in Study 1:

```
gay1.control <- subset(gay.reshaped, (study == 1) &
                            (treatment == "No Contact"))
hist(gay1.control$therm1, freq = FALSE, main = "wave 1")
hist(gay1.control$therm2, freq = FALSE, main = "wave 2")
```



3/ Bivariate relationships

Scatterplot

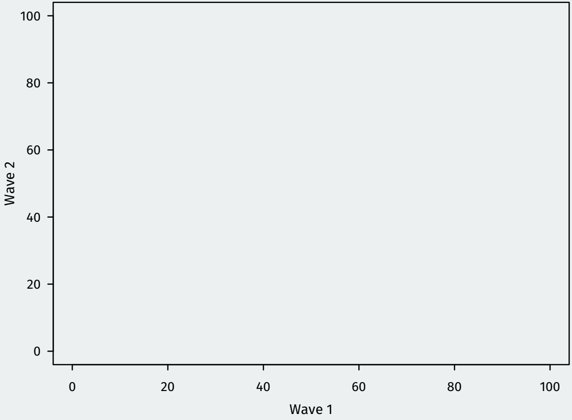
- Direct graphical comparison of two variables.
- Each point on the scatterplot (x_i, y_i)
- Use the `plot()` function

```
plot(x = gay1.control$therm1, y = gay1.control$therm2,
     xlab = "Wave 1", ylab = "Wave 2")
```

Scatterplot

```
gay1.control[1, c("therm1", "therm2")]
```

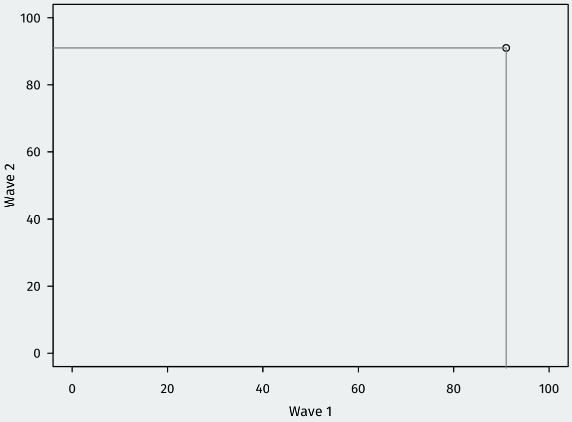
```
##   therm1 therm2
## 1     91     91
```



Scatterplot

```
gay1.control[1, c("therm1", "therm2")]
```

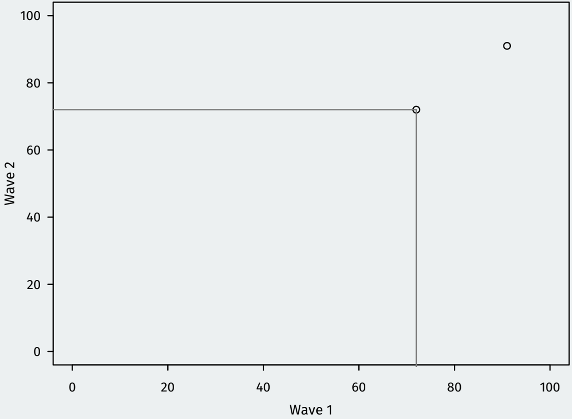
```
##   therm1 therm2
## 1     91     91
```



Scatterplot

```r
gay1.control[2, c("therm1", "therm2")]
```

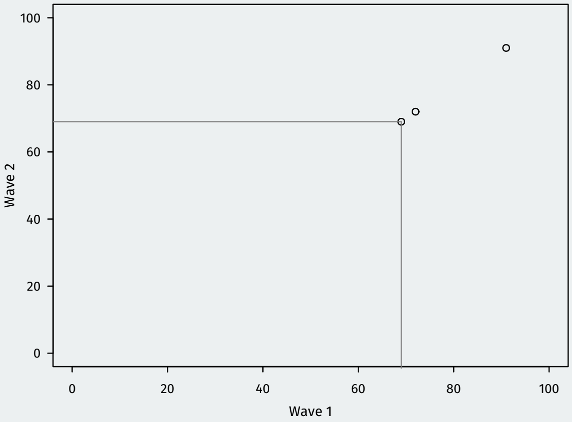
```
##   therm1 therm2
## 2     72     72
```



Scatterplot

```
gay1.control[3, c("therm1", "therm2")]
```

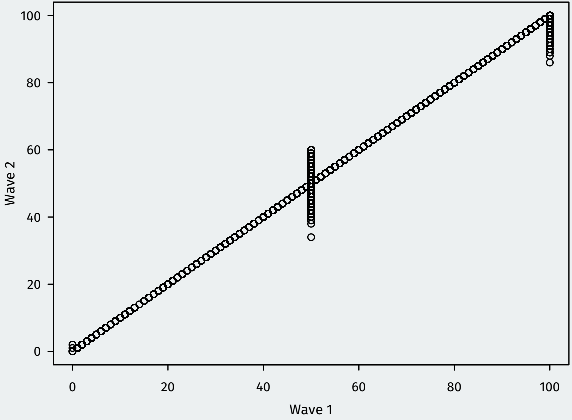
```
##   therm1 therm2
## 3     69     69
```



Scatterplot

```
gay1.control[1,c("therm1", "therm2")]
```

```
##   therm1 therm2
## 1     91     91
```



How big is big?

- Variables can be on different scales: makes it difficult to assess how well they "go together"
- Need a way to put any variable on common units.
- **z-score**:

$$\text{z-score of } x_i = \frac{x_i - \text{mean of } x}{\text{standard deviation of } x}$$

- z-scores don't depend on units:

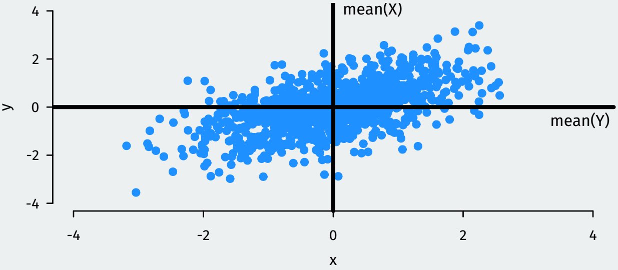
$$\text{z-score of } (ax_i + b) = \text{z-score of } x_i$$

Correlation

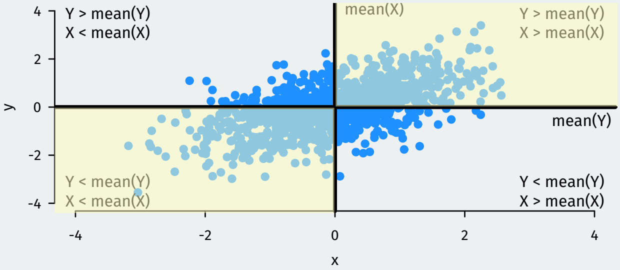
- How do variables move together on average?
- If I know one variable is big, does that tell me anything about how big the other variable is?
  - Positive correlation: when x is big, y is also big
  - Negative correlation: when x is big, y is small
  - High correlation: data cluster tightly around a line.
- The technical definition of the **correlation coefficient**:

$$\frac{1}{n-1}\sum_{i=1}^{n}\left[(\text{z-score for } x_i) \times (\text{z-score for } y_i)\right]$$

Correlation intuition

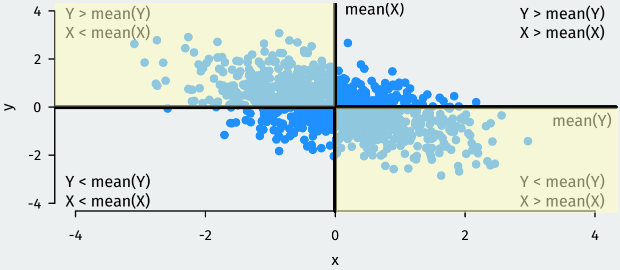


Correlation intuition



- Large values of X tend to occur with large values of Y:
  - (z-score for x_i) $\times$ (z-score for y_i) $=$ (pos. num.) $\times$ (pos. num) $= +$
- Small values of X tend to occur with small values of Y:
  - (z-score for x_i) $\times$ (z-score for y_i) $=$ (neg. num.) $\times$ (neg. num) $= +$
- If these dominate $\rightsquigarrow$ positive correlation.

Correlation intuition



- Large values of X tend to occur with small values of Y:
  - $(\text{z-score for } x_i) \times (\text{z-score for } y_i) = (\text{pos. num.}) \times (\text{neg. num}) = -$
- Small values of X tend to occur with large values of Y:
  - $(\text{z-score for } x_i) \times (\text{z-score for } y_i) = (\text{neg. num.}) \times (\text{pos. num}) = -$
- If these dominate $\rightsquigarrow$ negative correlation.

Properties of correlation coefficient

- Correlation measures **linear** association.
- Interpretation:
  - Correlation is between -1 and 1
  - Correlation of 0 means no linear association.
  - Positive correlations $\rightsquigarrow$ positive associations.
  - Negative correlations $\rightsquigarrow$ negative associations.
  - Closer to -1 or 1 means stronger association.
- Order doesn't matter: `cor(x,y) = cor(y,x)`
- Not affected by changes of scale:
  - `cor(x,y) = cor(ax+b, cy+d)`
  - Celsius vs. Fahreneheit; dollars vs. pesos; cm vs. in.

Correlation in R

- Use the `cor()` function
- Missing values: set the `use = "pairwise"` $\rightsquigarrow$ available case analysis

```
cor(gay1.control$therm1, gay1.control$therm2,
    use = "pairwise")
```

```
## [1] 0.998
```

- Extremely high correlation!

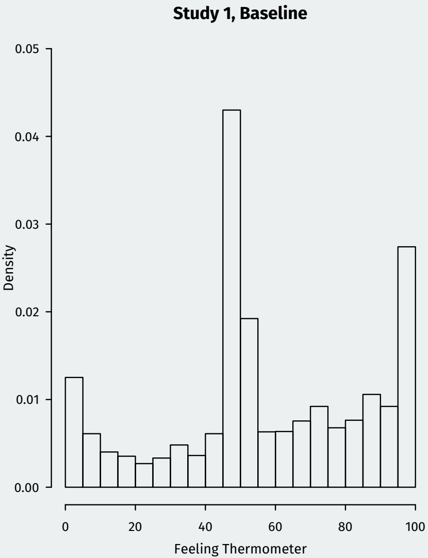
Comparisons between studies

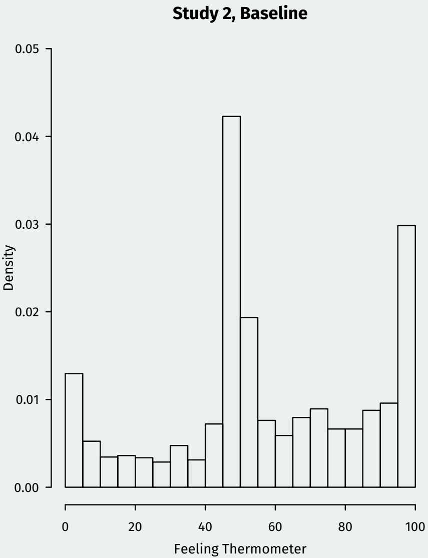
- Cannot use `plot()` or `cor()`. Why?
- Different studies have different respondents.
- Start with histograms:

```
gay1 <- subset(gay.reshaped, (study == 1))
gay2 <- subset(gay.reshaped, (study == 2))
```

```
hist(gay1$therm1, freq = FALSE, breaks = 20,
 ylim = c(0, 0.05), xlab = "Feeling Thermometer",
 main = "Study 1, Baseline")
hist(gay2$therm1, freq = FALSE, breaks = 20,
  ylim = c(0, 0.05), xlab = "Feeling Thermometer",
  main = "Study 2, Baseline")
```

Very similar!!



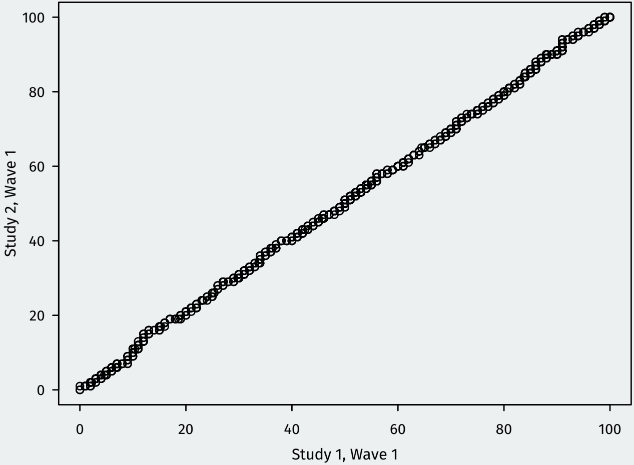


Quantile-Quantile Plot

- **Quantile-quantile plot (qq-plot)**: Plot the **quantiles** of each distribution against each other.
- Example points:
  - (min of X, min of Y)
  - (median of X, median of Y)
  - (25th percentile of X, 25th percentile of Y)
- 45 degree line indicates quality of the two distributions.

QQ-plot example

```
qqplot(gay1$therm1, gay2$therm1, xlab = "Study 1, Wave 1",
       ylab = "Study 2, Wave 1")
```



What is going on?!!?

- Question wording of thermometer score attributed to 2012 Cooperative Campaign Analysis Project (CCAP):

Name	Description
`caseid`	unique respondent ID
`gaytherm`	Survey thermometer rating (0-100) of feeling towards gay couples

- CCAP has some missing data:

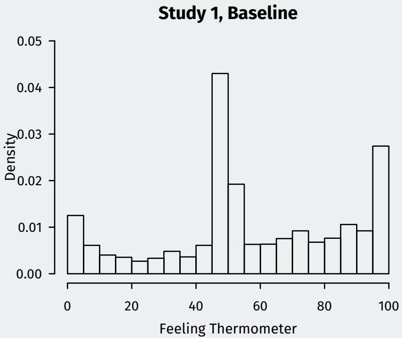
```
ccap <- read.csv("data/ccap2012.csv")
mean(is.na(ccap$gaytherm))
```

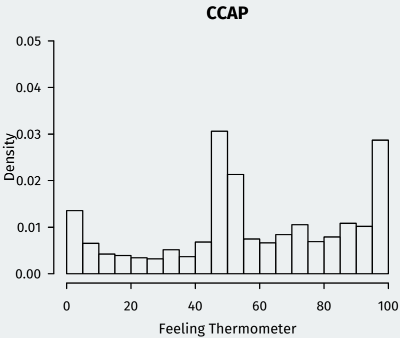
```
## [1] 0.0704
```

```
mean(is.na(gay1$therm1))
```

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

Comparison of CCAP and Study 1

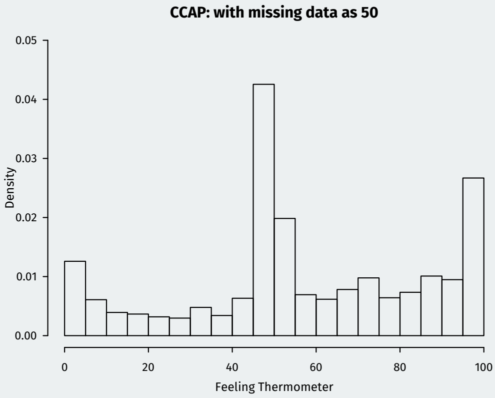




- Suspiciously similar!
- What's the difference?

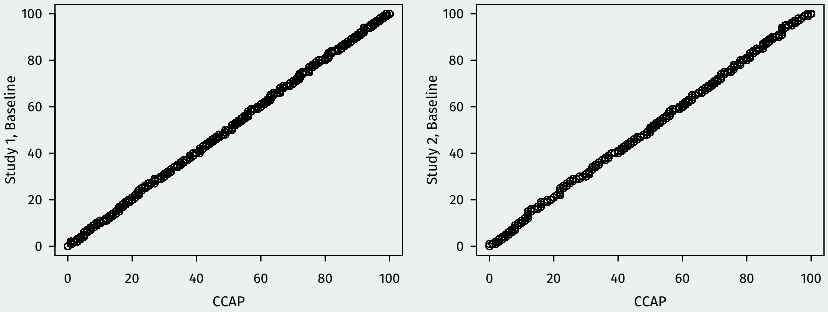
Recoding missing as 50s

```
ccap$gaytherm[is.na(ccap$gaytherm)] <- 50
hist(ccap$gaytherm, freq = FALSE,
     ylim = c(0, 0.05), xlab = "Feeling Thermometer",
     main = "CCAP: with missing data as 50")
```



QQ plots reveal extreme similarity

```
qqplot(ccap$gaytherm, gay1$therm1, xlab = "CCAP",
       ylab = "Study 1, Baseline")
qqplot(ccap$gaytherm, gay2$therm1, xlab = "CCAP",
       ylab = "Study 2, Baseline")
```



Retraction

SCIENCE   359 COMMENTS

Doubts About Study of Gay Canvassers Rattle the Field

By BENEDICT CAREY and PAM BELLUCK MAY 25, 2015

Donald P. Green, left, a co-author of a challenged study by Michael LaCour, right, from Mr. LaCour's Facebook page.

Wrapping up

- Scatterplots, correlation, and QQ-plots all help us visualize relationships between variables.
- With gay-marriage study, helped us detect fraud.
- After midterm: prediction!