descriptive statistics 1-1: relationships: summarizing more than one variable: crosstabs and correlation, (Wheelan, 2013, ch3,4)

Adam Okulicz-Kozaryn

adam.okulicz.kozaryn@gmail.com

this version: Tuesday 17th September, 2024 16:52

ps1 fa24 general comments (specific on canvas)

- be clear about u/a
- for quant study need $>$ 20 or 30, if n=1 or 2, say just Camden, can do case study–talk to dr Robinson (and in this class just do lit rev)
- dont forget to put your topic in google scholar, eg: `https://scholar.google.com/scholar?hl=en&as_sdt=0%2C31&q=big+brothers+big+sisters&btnG=`
- check out `https://www.statepolicyindex.com/data/` and `http://ippsr.msu.edu/public-policy/correlates-state-policy`

ps1 fa24

- think about mechanism–whats the lag between the predictor and outcome
- again, try not to collect the data; especially that pretty much any data you may need has already been collected

ps1 fa21 (old comments)

- start by putting your research topic into goog scholar!
  (and lit will also tell you about data)
- remember that in social science many variables play a role!
- no single outcome can be predicted by just one variable!
- again, ideally do something you're familiar with, eg
  job-related or been researching earlier in other classes

ps1 fa21

- great ideas, but do get actual data! just google it; and produce some stats
- `https://www.policymap.com` may be useful
- for many of you i commented, do before-after
  - discuss quickly; and for elaboration in few weeks
    `https://theaok.github.io/res/res_des.pdf`
- check out redlining maps

  `https://dsl.richmond.edu/panorama/redlining`

ps1 fa21: great ideas, but start working on it asap!

- just start writing!
- (can do free-writing, dump all the ideas on paper, worry later about organization)
- few discussed specific data and literature
- narrow down, be focused on sth specific,
- be specific, eg how would you measure 'fairness'
- measurment is the key! email me about finding data!
- use tools from class on your data asap!
- great to kill 2 birds with one stone: internship, etc
- and study something you are passionate about!

howto describe data?

- depends on lev of measurement! cat v num q&a
- numbers
- graphs (always better unless very few ua, say <5)
  - humans recognize patterns in graphs better and faster
- break it up into subsets/subsamples! dig deeper!
  - say see hist/tab for males and females separately
  - eg may rise and fall, eg swb and place size in china
- whiteboard

few categories / categorical

- use contingency tab / cross-tab (bc you cross-tab dat)
- use percents, not counts: usually clearer
- so what's the relationship: age and being a student?

What is your age?	Are you a student?			Total
	Yes - Full Time	Yes - Part Time	No	
15 and under	88%	12%	-	8
16 - 18	95%	-	5%	42
19 - 23	68%	12%	20%	205
24 - 29	16%	10%	74%	353
30 - 35	5%	9%	86%	192
36 - 45	4%	8%	88%	165
over 45	1%	7%	92%	129

- http://www.custominsight.com/articles/crosstab-sample.asp

crosstabs: row percents v col percents

ort: Cols ▾ Rows ▾ Count All % **Row %** Col %

Number of Employees at Company

ob Satisfaction	1-25	26-100	101-999	1,000-3,000	> 3000	Total
Hate my job	24.4%	14.1%	26.9%	12.8%	21.8%	100%
'm not happy in my job	31.6%	21.3%	19.2%	6.3%	21.5%	100%
t's a paycheck	27.6%	20.4%	22.6%	7.7%	21.8%	100%
enjoy going to work	32.3%	21.8%	21.3%	7.0%	17.6%	100%
Love my job	47.8%	17.2%	17.0%	5.0%	13.0%	100%

ort: Cols ▾ Rows ▾ Count All % Row % **Col %**

Number of Employees at Company

ob Satisfaction	1-25	26-100	101-999	1,000-3,000	> 3000
Hate my job	0.8%	0.8%	1.5%	2.2%	1.5%
'm not happy in my job	6.6%	7.9%	7.1%	7.2%	9.3%
t's a paycheck	12.6%	16.4%	18.1%	18.9%	20.4%
enjoy going to work	43.3%	51.6%	50.3%	50.8%	48.4%
Love my job	36.7%	23.2%	23.0%	20.9%	20.5%
Total	100%	100%	100%	100%	100%

crosstabs covid example

- https://www.cdc.gov/mmwr/volumes/70/wr/mm7037e1.htm#T1_down

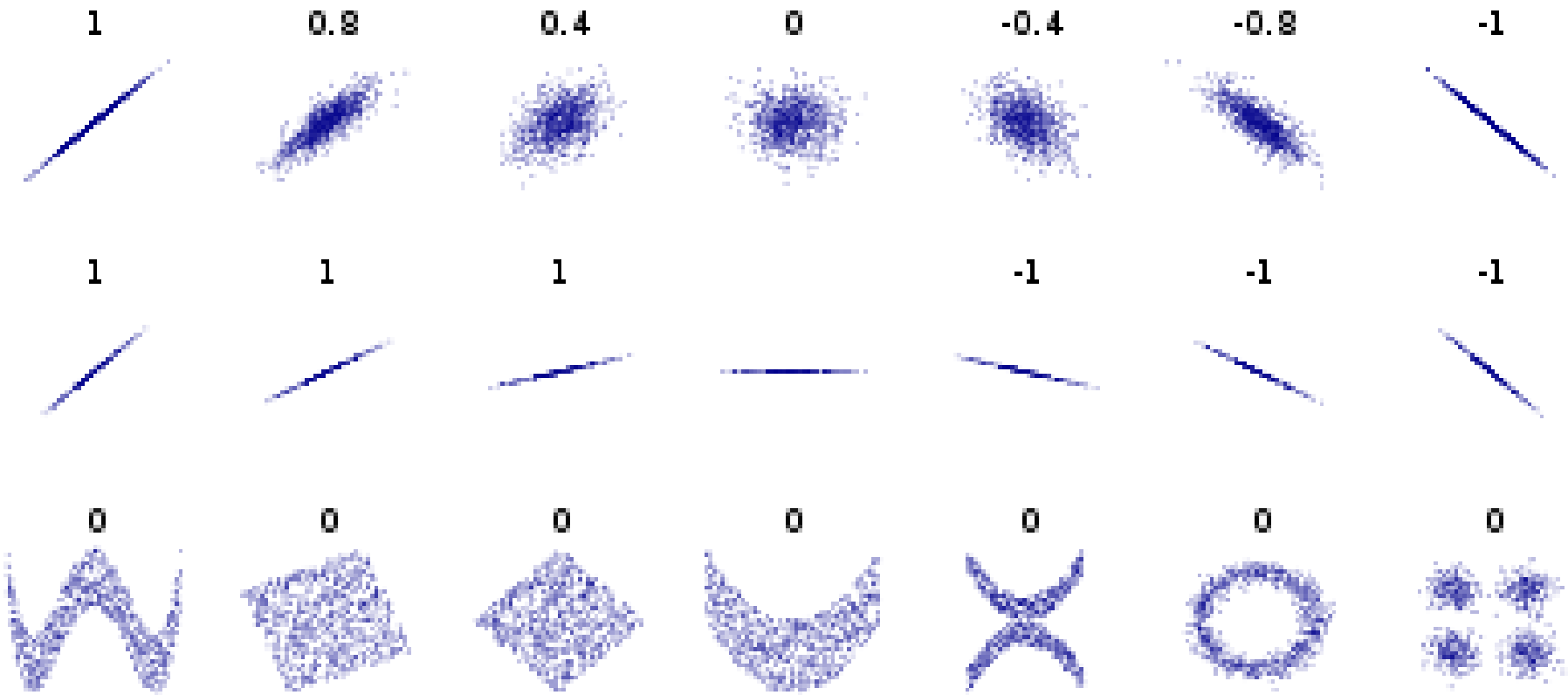
percentage change v percentage point change

- good school's dropout rate increases from 2% to 4%
  - percentage point increase is $4 - 2 = 2$
  - percentage increase is $(\frac{4-2}{2}) * 100 = 100$
  -
- bad school's dropout rate increases from 50% to 75%
  - percentage point increase is $75 - 50 = 25$
  - percentage increase is $(\frac{75-50}{50}) * 100 = 50$
  -
- if you start from low base (eg 2), then small percentage point increase is huge percent increase!
  - it matters! eg anti-immigrant sentiment in Scandinavia

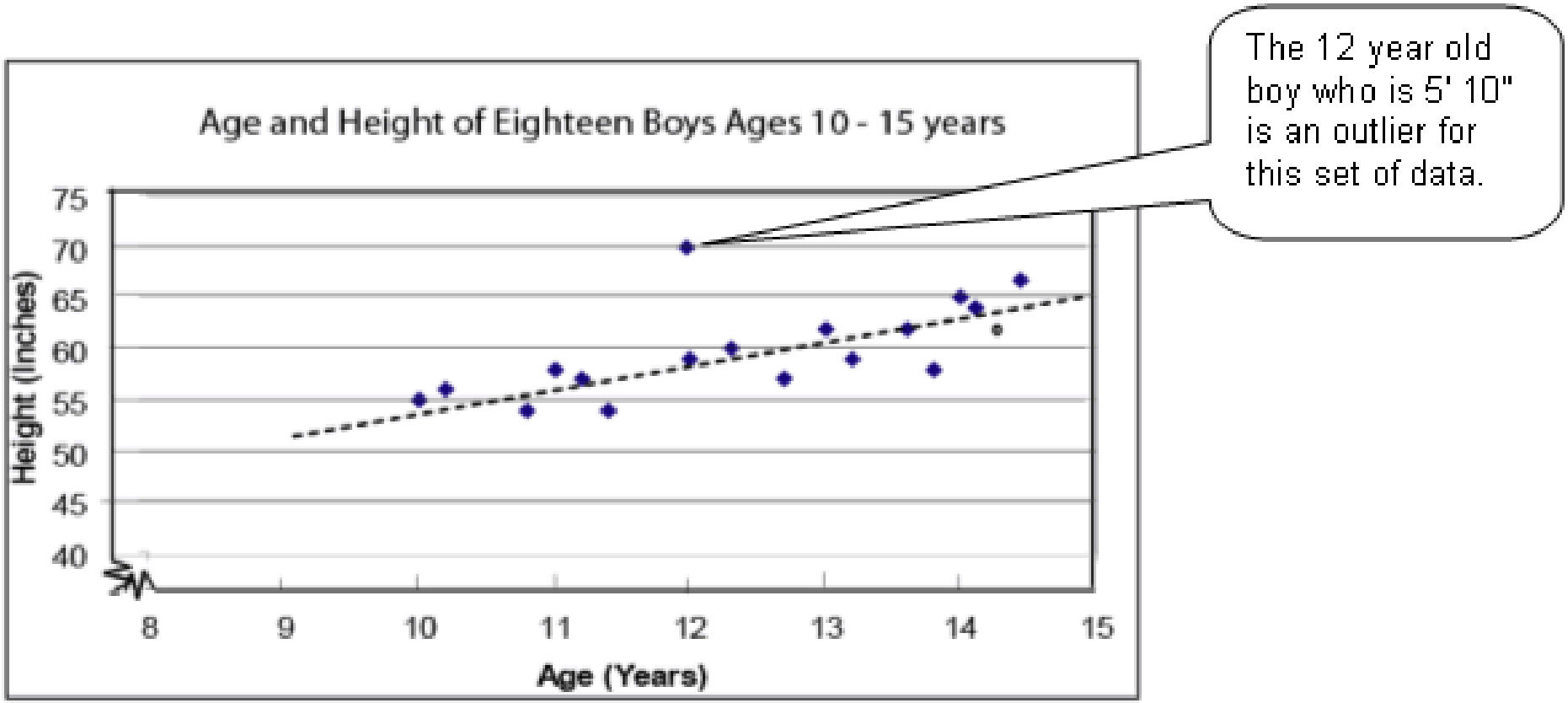
many cat / continuous data: corr and scatterplots

- just plot data in scatterplot; identify outliers!
- ex: outliers cops/1k and crime; marriage and divorce
- correlation range: -1 to 1
- $< |4|$ low
- $|.4 - .6|$ moderate
- $> |.7|$ strong
- again, keep in mind causation v correlation

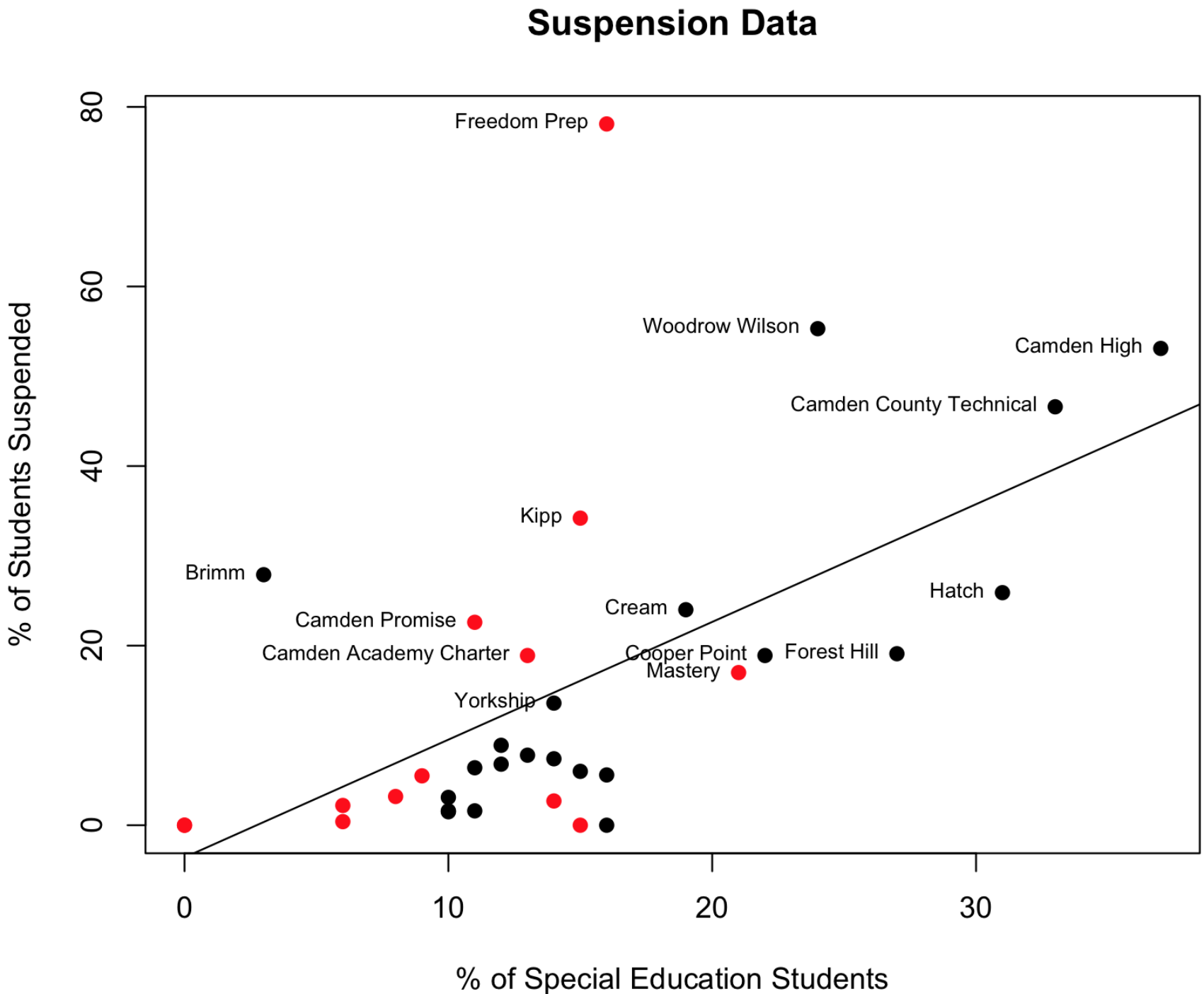
correlations for different scenarios



scatterplot



-
  - height and self-esteem:
    `http://www.socialresearchmethods.net/kb/statcorr.php`
-
- next slide: `https://danley.camden.rutgers.edu/2017/04/13/who-suspends-the-highest-percentage-of-camden-students-freedom-prep/`
  - red: charter/renaissance; black: Camden schools


Suspension Data
% of Students Suspended
% of Special Education Students
0
20
40
60
80
0
10
20
30
Freedom Prep
Woodrow Wilson
Camden High
Camden County Technical
Kipp
Brimm
Hatch
Cream
Camden Promise
Camden Academy Charter
Cooper Point
Forest Hill
Mastery
Yorkship

do scatterplots

- very useful to produce a scatterplot
- you'd see outliers–
- and whether the relationship is due to them
- blackboard: relationships biased due to outliers
- say marriage rate and divorce rate and that one state where really a lot of people get divorced (and married)

calculate it!

- there are formulas in wheelan and trochim
- but can just calc with software :)
- lets see trochim's example of height and weight from syl
- `https://conjointly.com/kb/correlation-statistic`

Wheelan in ch11 mentions Whitehall studies

- high status causes better health!
  - great book 'Status Syndrome' by m marmot
- eg nobel or oscar boosts one's health and longevity
  - these successful folks live longer and in better health
  - than exact same people (income, lifestyle, etc) but without status
- `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2566175/`
- Table 2A: correlations
  - esp 'Decision latitude' (scroll down)
  - conclusions?

Nudge by Thaler

- use bus tools for pub pol and adm
- ...

wrap-up

- end every class discussing what we covered and quick look at next week
- end with a review Q&A,
- give some examples (essp in pub pol and pub adm) for concepts covered
- students will discuss concepts from the class
-
- quick look at next class

bibliography I

Wheelan, C. (2013): Naked statistics: stripping the dread from the data, WW Norton & Company.