

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: Thursday 10th September, 2020 10:40

howto describe data?

- ◇ numbers
- ◇ graphs (always better unless very few data, 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
 - say corr or crosstab for low and hi val separately
that's a quick way to see nonlinear relationship!
eg it may first rise and then fall
- ◇ googSheet or xournal

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://help.statwing.com/knowledge_base/topics/how-do-i-interpret-crosstabs
- ◇ <http://www.custominsight.com/articles/crosstab-sample.asp>

crosstabs: row percents v col percents

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

Number of Employees at Company

Job 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%
I'm not happy in my job	31.6%	21.3%	19.2%	6.3%	21.5%	100%
It's a paycheck	⌵ 27.6%	20.4%	22.6%	7.7%	⌶ 21.8%	100%
I 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%

Sort: Cols ▾ Rows ▾ Count All % **Col %**

Number of Employees at Company

Job 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%
I'm not happy in my job	6.6%	7.9%	7.1%	7.2%	9.3%
It's a paycheck	⌵ 12.6%	16.4%	18.1%	18.9%	⌶ 20.4%
I 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%

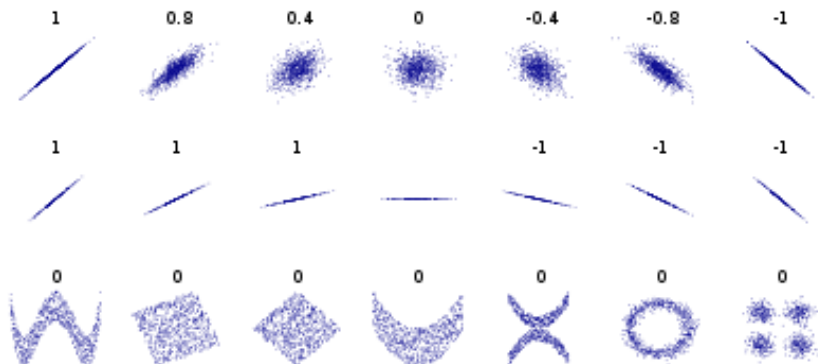
percentage change v percentage point change

- ◇ say good school's dropout rate increases from 2% to 4%
 - percentage point increase is $4 - 2 = 2$
 - percentage increase is $(\frac{4-2}{2}) * 100 = 100$
- ◇
- ◇ say 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!

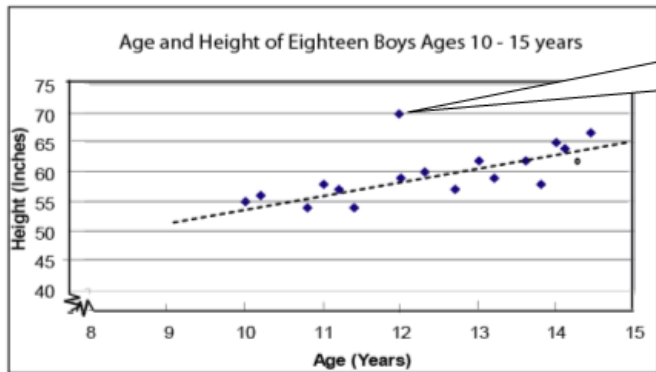
many categories / continuous data

- ◇ use correlation and scatterplots
 - just plot them in scatterplot; identify outliers!
 - **xournal: draw ex with outliers**
 - correlation ranges between -1 and 1
 - $< |.4|$ low
 - $|.4 - .6|$ moderate
 - $> |.7|$ strong
- ◇ again, keep in mind causation v correlation

correlations for different scenarios



scatterplot



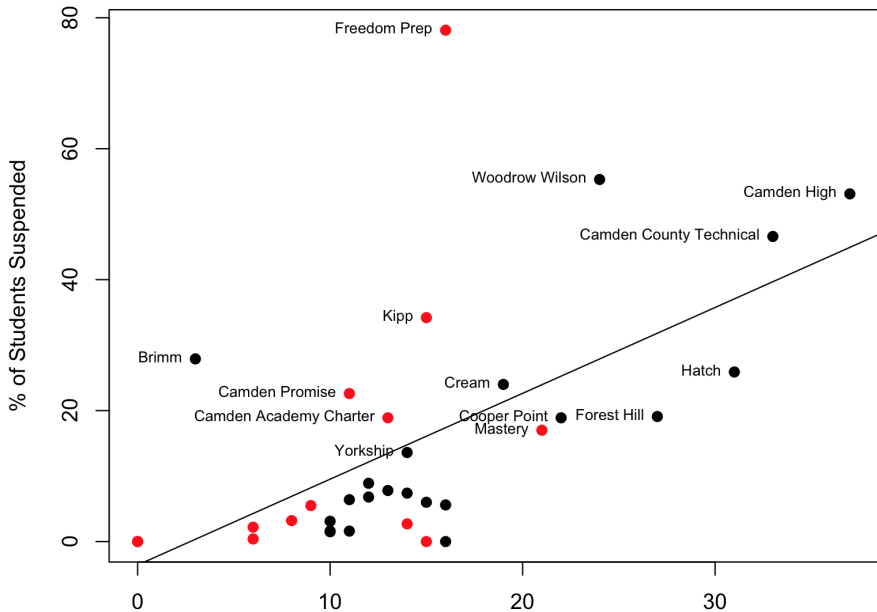
The 12 year old boy who is 5' 10" is an outlier for this set of data.

- also see <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/>

Suspension Data



do scatterplots

- ◇ it is 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 Nevada

calculate it!

- ◇ there are formulas in wheelan and trochim
 - but can just calc with software :)
 - can do it excel or google sheets etc
 - but it's 21st century, so lets do it in Python :)
 - see `des.py`

Wheelan in ch11 mentions Whitehall studies

- ◇ fascinating stuff!
- ◇ high status causes better health!
 - great book 'Status Syndrome' <http://a.co/jaUuwT7>
- ◇ say nobel prize 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

closer look at status syndrome

- ◇ <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2566175/>
- ◇ see Table 2A for correlations
 - especially 'Decision latitude'
 - conclusions? extra credit

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.