Cogs 9
Discussion Section

FA22 Week 5
Will McCarthy

Upcoming due dates

No reading quiz this week!

Friday, October 28th

- **Assignment 2**
- **Mid Way Team Evaluations extra credit**

This week's content: *data visualization*

Assignment 2 tips

4(a): Evan M. Peck, et al., 2019, Attitudes and Perceptions of Data Visualization

4(b): Hadley Wickham, et al., 2010, Graphical Inference for Infovis

Assignment 2 Tips: data

*Make sure data is in **tidy** format*

1. Each **variable** forms a **column**
2. Each **observation** forms a **row**
3. Each type of observational unit forms a table

Also follow best practices

(Remember units! Unlike these)

Untidy!

id	age	time_1	time_2	time_3	time_4
saawr	19	12	15	12	25
ajojet	21	13	13	14	20
tswar	20	20	15	13	19
serbse	19	15	20	14	19

Tidy!

id	age	time	measurement
saawr	19	1	12
saawr	19	2	15
saawr	19	3	12
saawr	19	4	25
ajojet	21	1	13
ajojet	21	2	13

Assignment 2 Tips: data

Untidy!

id	age_yrs	day	weight_AM_kg	weight_PM_kg
saawr	19	1	88	74
saawr	19	2	87	75
ajojet	21	1	70	74
ajojet	21	2	70	76
tswar	20	1	102	103
tswar	20	2	103	103
serbse	19	1	54	68

Assignment 2 Tips: visualization

Your visualization should:

1. Ask a question of the data
2. Be a true representation of the data

When interpreting your visualization:

> "what you want the viewer to take away from your visualization" **does not mean** "what do you wish to be true of your data"

Don't force an interpretation: a null result is still a result

Cover the basics, a checklist:

[ ] plot title

[ ] axis labels

[ ] axis units

[ ] tick labels

[ ] all data is represented

[ ] legend?

4(a): Evan M. Peck, et al., 2019, Attitudes and Perceptions of Data Visualization

See Shilpa and Mason's lecture slides

4(b): Hadley Wickham, et al., 2010, Graphical Inference for Infovis

What can we extract from the *abstract*?

Infovis used to discover new relationships

Statistics used to prevent spurious relationships from being recorded

Apophenia: human tendency to see patterns in noise/ see meaningful connections when none exist

Two new tools:

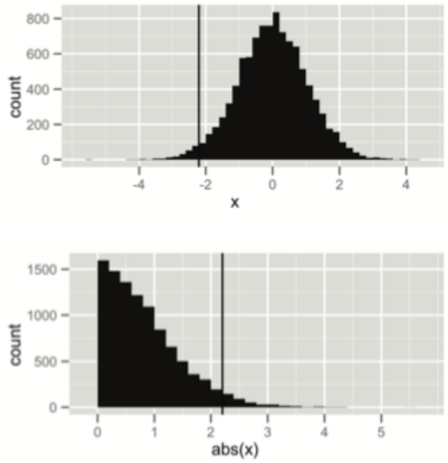
- Rorschach: Helps analysts calibrate their understanding of uncertainty
- Line-up: for assessing significance of visual discoveries

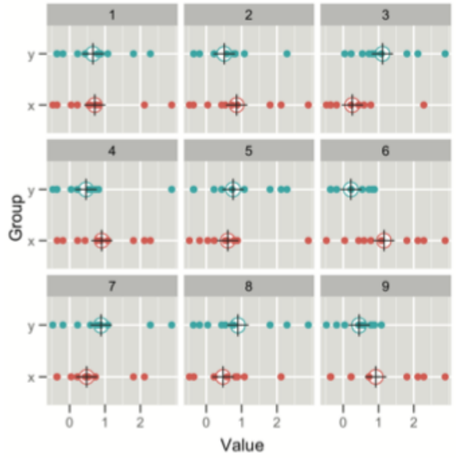
Visual inference: what is and why?

Goal of many *statistical* methods is **inference***:*

Drawing conclusions about the population that sample data came from

Statistics works great with well-behaved data that follows a known distribution

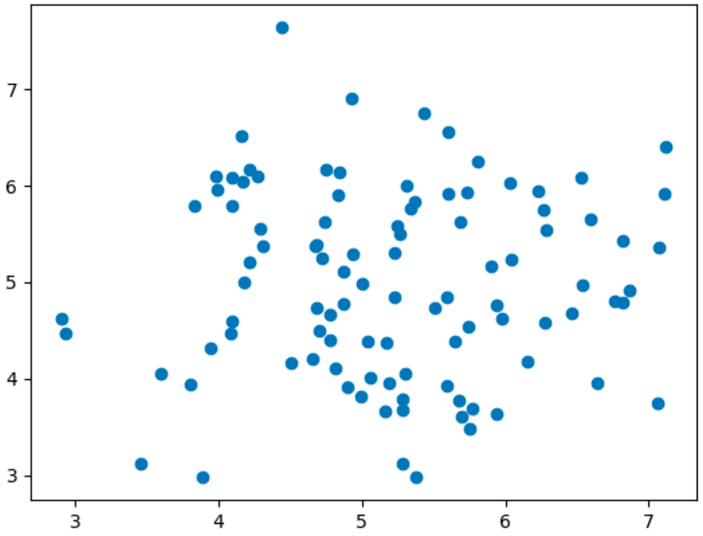




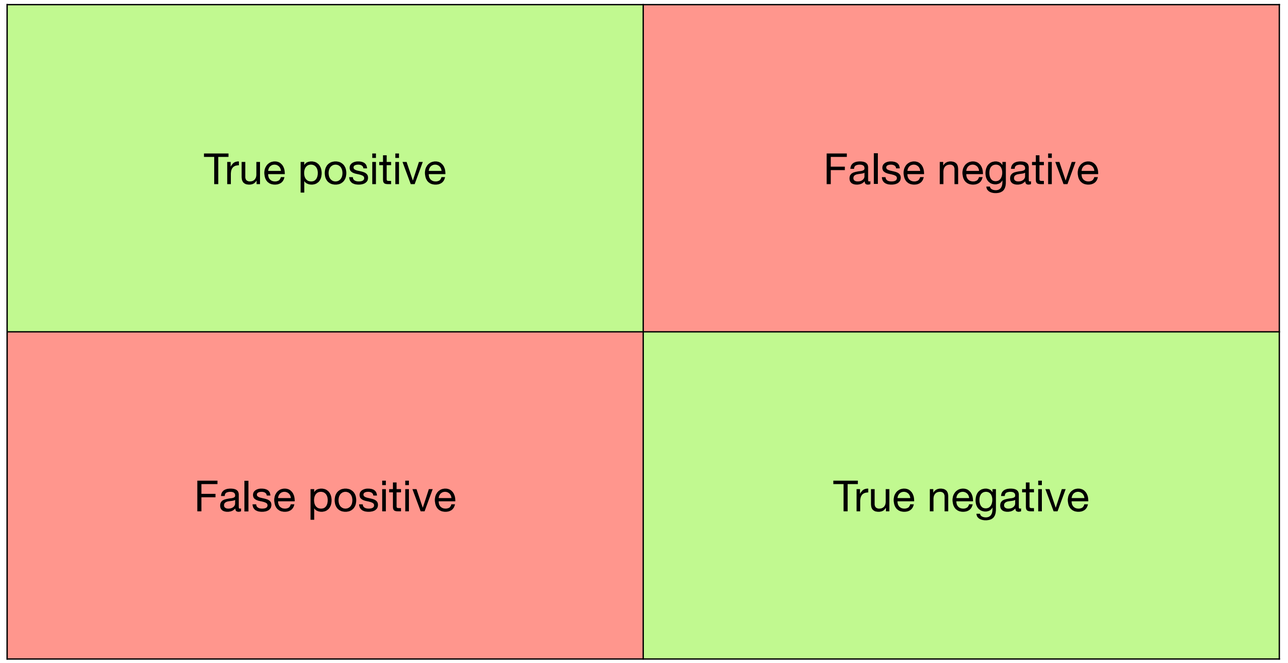
Visual inference: use *vision* to draw conclusions about the population that sample data came from

Visual inference can be used in **complex data analysis settings that do not have corresponding numerical tests** (but not necessarily an easy task!)

Infovis and statistics push in opposite directions for making inferences



Infovis: find as many relationships as possible (curiosity)



$p > 0.05$

Statistics: check to see if relationships are actually true (skepticism)

Two tools for better visual inference:

1. Rorschach Protocol
2. Line-up

Rorschach Protocol

Rorschach Test:

What do you see in random ink blots?

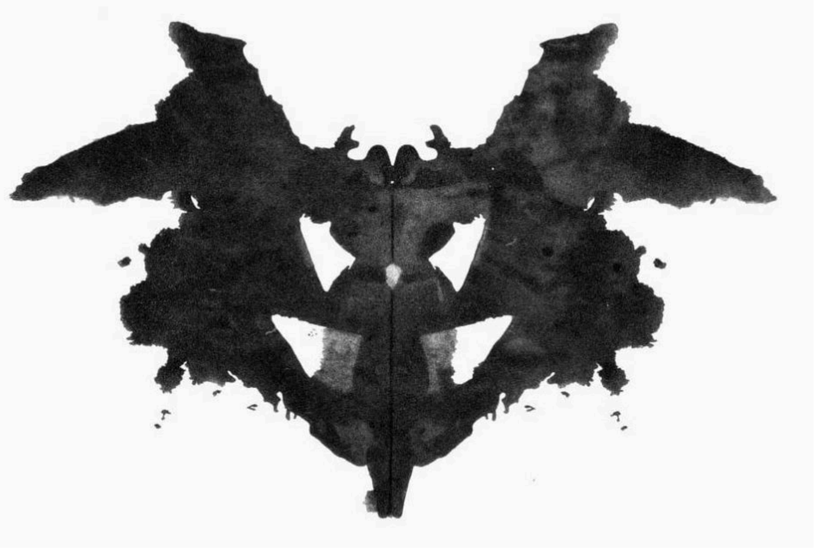
We tend to overestimate meaning in randomness (“apophenia”)

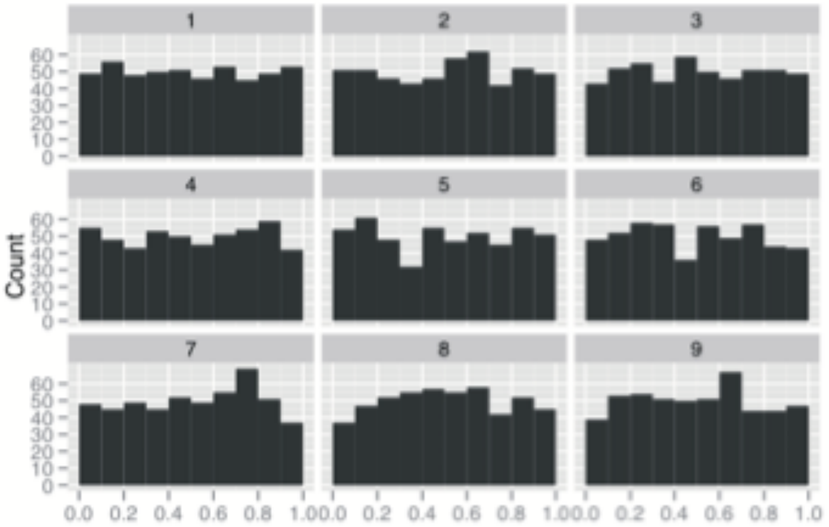
This means we underestimate variability in plots

Rorschach Protocol:

What do you see in null plots?

“Calibrate our vision to the natural variability in plots”





Line-up

Generate a set of “innocents” i.e. null plots

Randomly position actual plot among these null plots

Show to *impartial observer* (ideally)

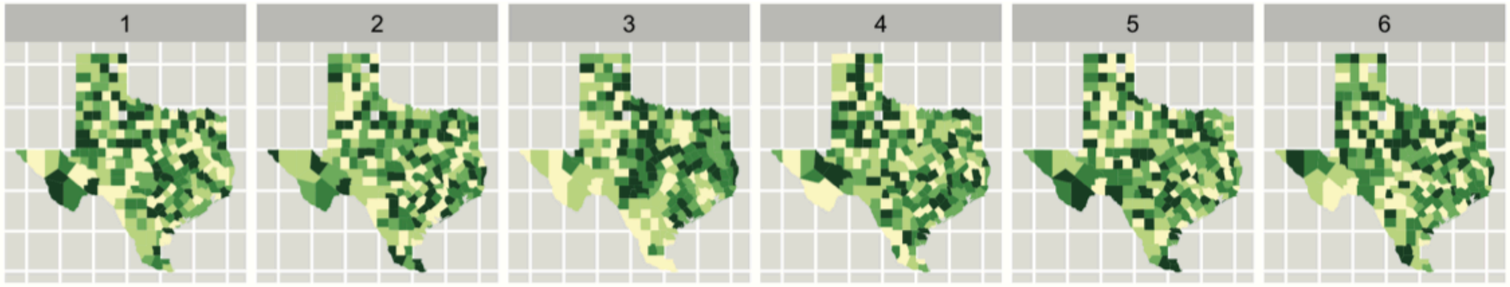


Fig. 1. One of these plots doesn’t belong. These six plots show choropleth maps of cancer deaths in Texas, where darker colors = more deaths. Can you spot which of the six plots is made from a real dataset and not simulated under the null hypothesis of spatial independence? If so, you’ve provided formal statistical evidence that deaths from cancer have spatial dependence. See Section 8 for the answer.

To use the line-up protocol we need to:

- Identify the question the plot is trying to answer.

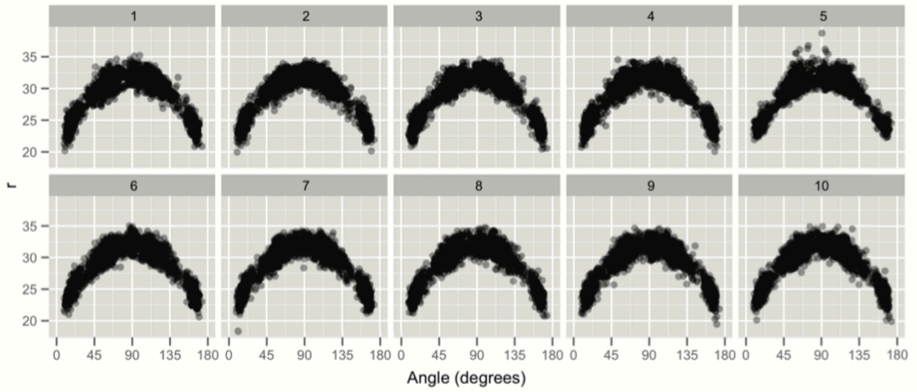
  (Usually determined by the plot.
  e.g. “what is the relationship between x and y)

- Characterize the null-hypothesis (the position of the defense).

  (Usually the least interesting answer to the question.
  e.g. “no relationship”)

- Figure out how to generate null datasets.

  (Resampling: does data come from this distribution
  vs
  Simulation: more specific relationship
  e.g. does x decrease linearly with y?)



Group work / questions

Future Readings

5(a): Nicholas Diakopoulos, 2016, Accountability in Algorithmic Decision Making

5(b): Julia Angwin, et al., 2016, Machine Bias