Final Projects

kmg

Overview

Overview

Your final projects will be comprised of two distinct parts:

- Code & Written report
- Presentation

All components are group efforts, and only one report needs to be turned in per group. Group sizes can be from 3-5 people.

Roles

It is up to your group to divide up roles. For some groups, one person writes, another codes, and another presents. In another group, it might make more sense for all members to contribute to all aspects of the project.

Required pieces

1) At least one function you created
2) Data manipulation
3) Statistical analysis
4) Plots
5) Introduce at least one function from CRAN that hasn't been used in class
6) Rmarkdown file (this is your final report)

While it is not required, it is strongly recommended that you use GitHub to collaborate on your code.

Projects

1. **Median differences**: examine state anxiety on and off caffeine
2. **Multiple Testing**: crime stats by region/county across time within NC
3. **Lying with Statistics/plots*: data we've used in class

Projects

The constant theme throughout projects is: (1) doing correct analysis and fair depiction of results and (2) doing misleading or incorrect analysis and depiction of results

You can use any data you would like.

Median Differences

Median Differences

- One option is to use the data set *sai* that is in the package *psych*.
- Across 11 studies, people were given caffeine or a placebo.
- Anxiety was the primary outcome of interest.
- The variables are described here: https://rdrr.io/cran/psychTools/man/sai.html

Median Differences: Project

- Examine just variables at time 1 so that we are doing between-subject analyses.
- Find a non-normally distributed variable.
- Identify if people differ in their medians between placebo and drug on this variable.

Median Differences: Details

1. Data “sai” can be obtained from the “psych” package.
2. Conduct and interpret results obtained from bootstrapping to test median differences.
3. Conduct and interpret results obtained from using t-test without bootstrapping.
4. Create figures to depict results and distributions.

Multiple testing

Multiple testing

- Data collected across time and across geographic areas provides a lot of opportunity for interesting graphics
- There's also a lot of tests that can be done:
  - mean differences between two years for all regions
  - compare all regions' values at a given time to the average to identify those significantly above/below
  - across all time points, calculate the correlation between two variables
- When we do a lot of tests, we have to correct for them! Some will be significant by chance, so we need to ensure we make accurate inferences.

Multiple testing: Project

- The data are crime data from NC across various counties and years: https://rdrr.io/cran/plm/man/Crime.html
- The primary goal of this project is to identify differences or correlations that are significant by looking across a lot of tests.
- Options are:
  - to look across time: which counties/regions had significant changes between two time points?
  - to look across counties/regions at the same time point: which ones are signficantly different than the average? - You can look across time for one region/county, or across counties/regions at one time, or both.

Multiple testing: Figures

- You can create figures to depict output, data or both.
- Inspiration: https://r-graph-gallery.com/

Multiple testing: Details

1. Data "Crime" can be obtained from the "plm" package.
2. Conduct your tests, and correct for multiple testing in some way. Justify your approach.
3. Discuss inferences and how they differ or are the same from results where you did *not* correct for multiple testing.

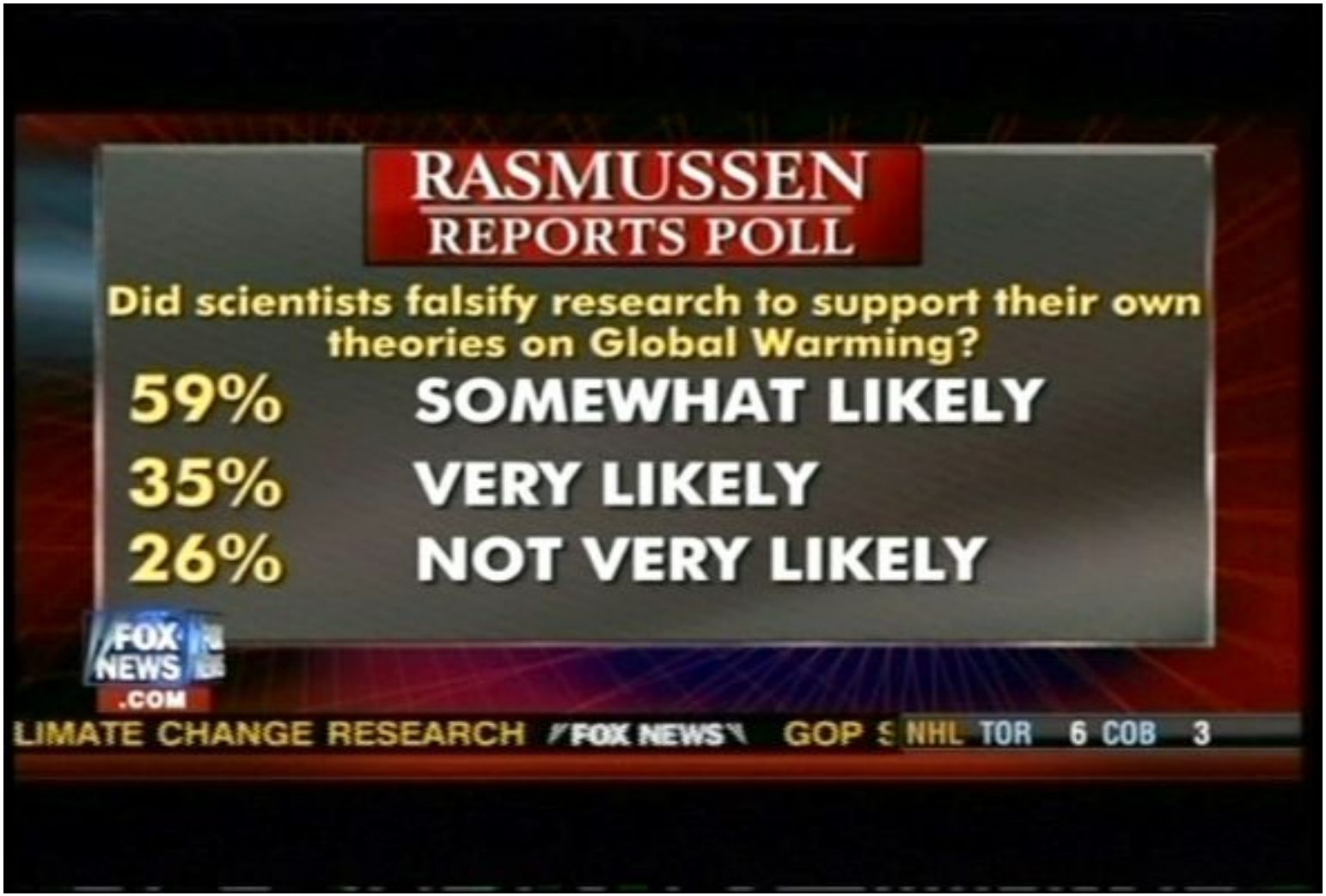
Lying with statistics/plots

Lying with statistics/plots

- When discussing how to create effective figures, we talked about how easy it can be to depict information in a misleading way.
- People can also do this with statistics:
  - p-hacking: estimating multiple models to find one that is significant and supports their hypothesis
  - recode data in a way that achieves a significant result
  - keep outliers that greatly influence results
  - remove covariates that render results null
  - cherry pick the sample

Lying with statistics/plots

Sometimes it's really obvious. . .

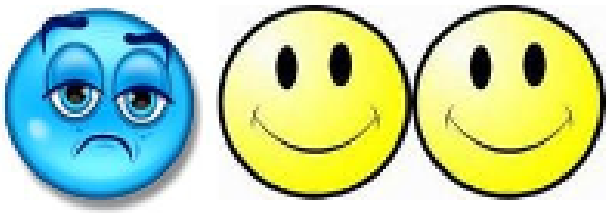


Lying with statistics/plots

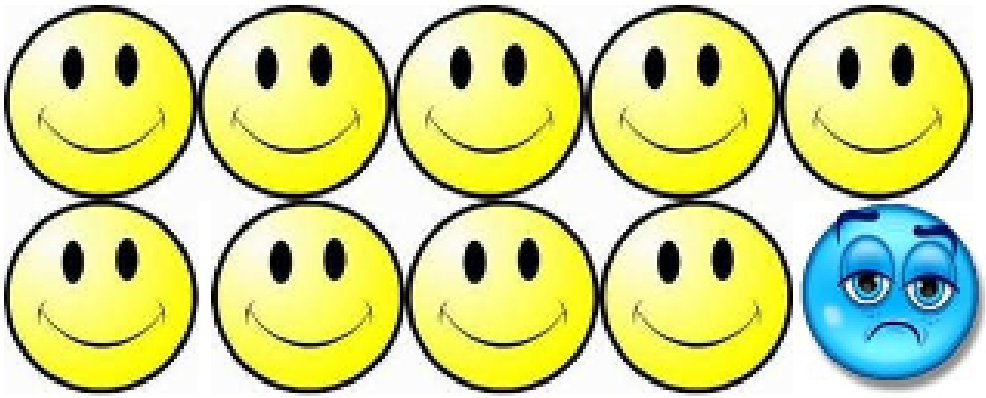
. . . but usually it's more subtle.

For instance, NPR reported that, "1/3 of McCain ads has been negative. . . 9/10 Obama ads have been positive."

A more precise way to say that: 66% of McCain ads have been positive whereas 90% of Obama ads were positive".



McCain ads (by NPR's analogy)



Obama ads
(by NPR's analogy)

Lying with statistics/plots

Obama ads
(by NPR's analogy)

McCain ads
(depicted like
Obama ads)

Lying with statistcs: Project

- You can use any data you'd like.
- Provide a fair figure depicting data or results and also conduct and interpret a statistical test of your choosing.
- Create a figure and/or analysis that provides a distorted version of what we actually would find in the data.