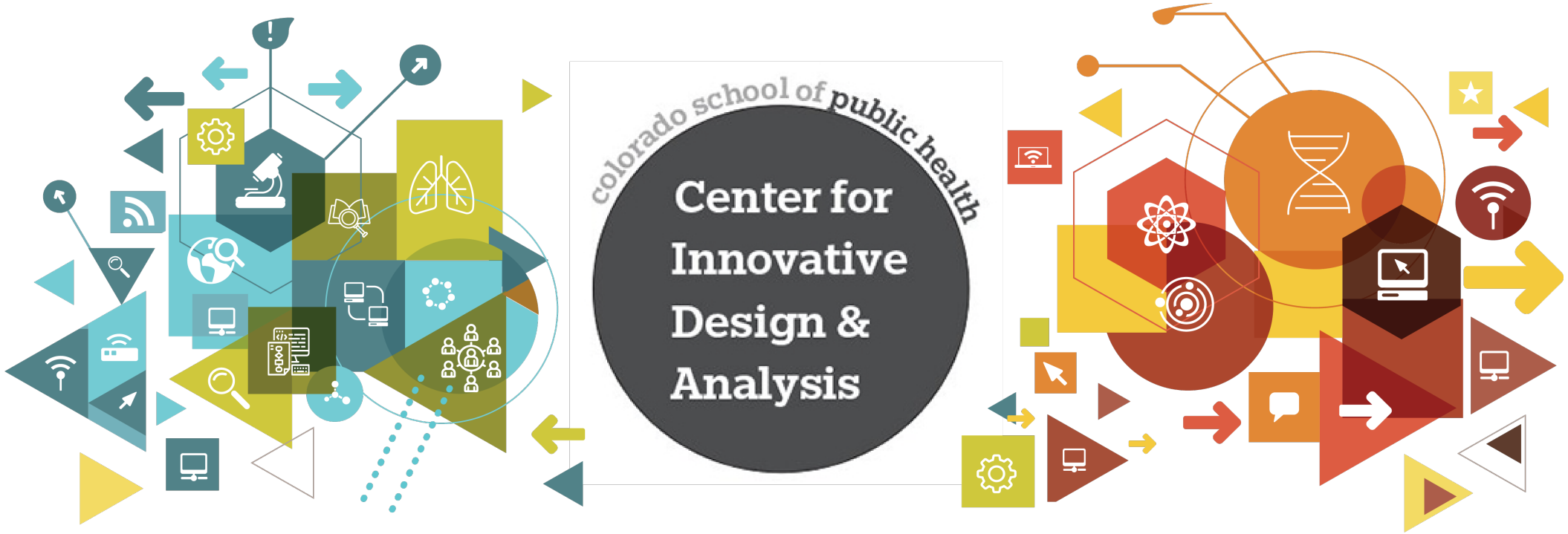
2022 CU Data Week



Basic statistics using R

- Instructor: Conner Jackson
- conner.jackson@cuanschutz.edu

Outline

- Download R
- Download RStudio
- Download our example data for CU Data Week
- Load data into R
  - Describe the data
  - Develop hypotheses
- Summarize data
- T-tests (and non-parametric tests)
- Chi-square tests
- Simple Linear regression
- Questions?

Download R

- Google “R Software”
  - https://www.r-project.org/
- Click download R link
  - https://cran.r-project.org/mirrors.html
- Choose a CRAN
  - I chose University of Kansas, pick a location that is close to you
  - https://rweb.crmda.ku.edu/cran/
- For Mac – don’t forget to download XQuarts

Download RStudio

- Google “download R studio”
  - https://www.rstudio.com/products/rstudio/download/
- Download R studio for your computer
  - After you have downloaded R
- Open R Studio
  - You shouldn’t need to open R

Download our example data

- Data is stored here:
  - https://github.com/CIDA-CSPH/CU-Stats-Talks
- Data comes from SAS Software
  - https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.3/statug/statug_sashelp_sect001.htm
  - Version 9.4 (SAS Institute, Cary NC)
- Click “Lunchtime Basics-Basic Statistics using R”
- Click “Data_PublicSAS”
- Click each .csv file
- Click “Download”
- ‘Right click’ and then Save As

Load data into R

- Open R Studio
- Set our directory to “Downloads”
  - This is where our data was stored after we downloaded it
- Import our data
  - bweight.csv = infant birthweight (continuous variable)
  - birthweight.csv = infant birthweight (binary variable - low birth weight; Y/N)
  - cars.csv = car data from 2004 with various variables (MPG, Weight)
  - class.csv = fictional class with various variables (Age, Height)

Data – Birthweight

- We have both continuous and categorical data to work with
  - Birthweight in grams
  - Birthweight as a categorical binary variable (Y/N)
  - Maternal Smoking status (Y/N)
- Hypotheses to test:
  - Is there a significant relationship between infant birth weight and maternal smoking?
  - Test 1: Maternal smoking status (Y/N) vs Low Birthweight (Y/N)
    - Two categorical (grouping) variables
    - Chi square test
  - Test 2: Maternal smoking status (Y/N) vs Birthweight
    - A continuous variable and a categorical variable
    - T-test vs Wilcoxon Rank Sum/Mann-Whitney test as appropriate

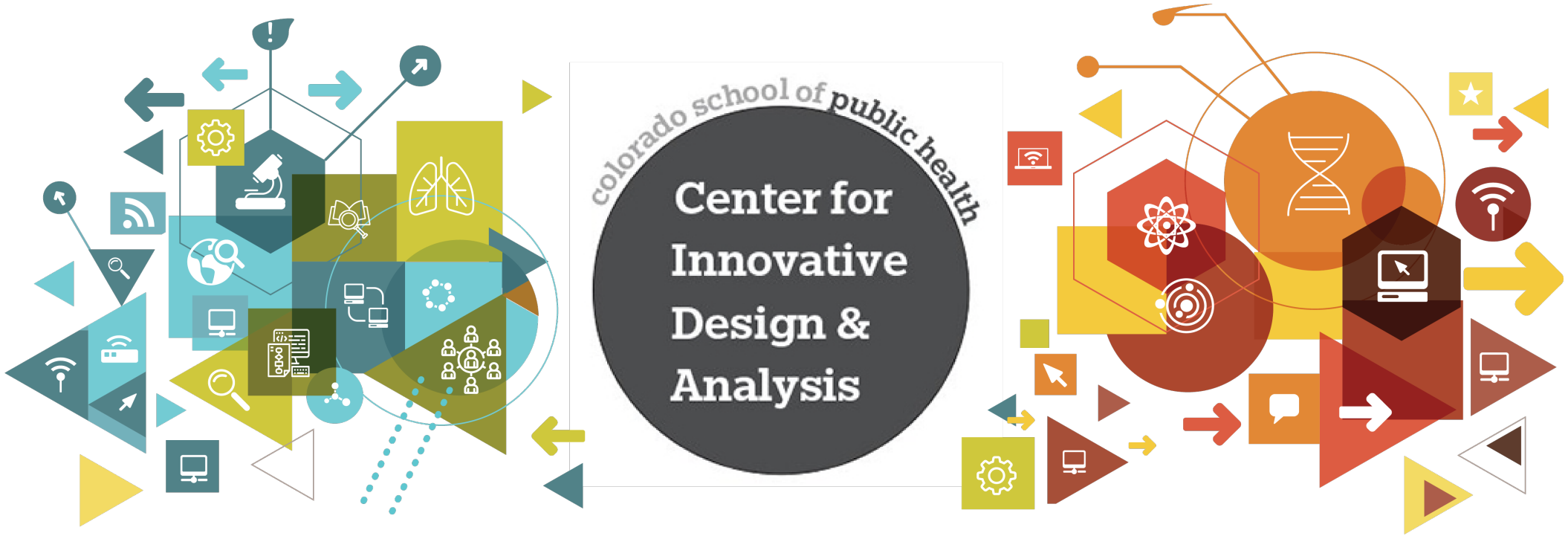
Data – Cars and Class

- We have both continuous and categorical data to work with
  - Focus on continuous data
  - Highway MPG and City MPG
  - Height in inches
  - Weight in pounds
- Hypotheses to test:
  - Cars
    - Test 3: Is there a significant difference between Highway and City MPG?
      - 2 paired continuous variables, difference
      - Paired t-test or Wilcoxon Signed Rank as appropriate
  - Class
    - Test 4: Is there a significant relationship between weight and height of school children?
      - 2 continuous variables, trend
      - Simple linear regression

R is difficult to learn – these can help

- Swirl
  - An interactive tool that assists you as you code
  - https://swirlstats.com/students.html
- R for Data Science
  - Free online text that uses tidyverse
  - https://r4ds.had.co.nz/
- R is like learning a language
  - You need to practice regularly
  - Challenge yourself to use R for something you would normally do elsewhere and compare your results

Questions?


colorado school of public health
Center for Innovative Design & Analysis