Statistics: is
a systematic method for drawing conclusions from data with variability
the scientific study of how to:
  collect data (design)
  summarize data (description)
  draw conclusions and make predictions from incomplete data
    (estimation, inference, prediction)

A: My model for scientific investigations: Question $\longrightarrow$ (Design) Data $\longrightarrow$ (Analysis) Conclusion

Five components: design, description, estimation, inference, prediction
  Design: How to collect data (not a focus in this class)
  Description: describing patterns in the sample
  Estimation: estimating population quantities from sample statistics
  Inference (p 8): "conclusion that patterns in the data are present in some broader context"
  Statistical Inference:
    "above justified by a probability model linking the data to a broader context"
  Prediction: predicting characteristics, e.g., values, of new observations

B: Usual terminology (p 19):
Population / parameter; Sample / statistic

Probability model connects sample to the population
  Depends on how data are collected

Data are presumed to come from some random mechanism.
  Design-based: estimation and inference justified only by study design,
    i.e. only on how data collected.
  Model-based: presume a model for the random mechanism.

Design-based: two possible random mechanisms:
Choice of units: random, haphazard, or arbitrary
Assignment of treatments to units: randomly or not

Four types of studies: (book display 1.5)

	Treatment assignment	
Choice of units	Random	Not
Random	experiment (rare)	survey
Not	experiment	observation

C: Random assignment of treatments really matters:
retraction/re-publication of Med. diet study

Surveys seem simple: just ask questions. Actually very difficult to do well

Literary Digest 1936 “survey”

Data description: numerical or graphical, no inference

D: What are we interested in?: location, spread, shape

Notation: Σ: sum, X_i: i'th observation

Graphical portrayals of the data:
dot plots, histograms, box plots

Numerical measures:
E: Location: mean: $\overline{X} = \Sigma X_i/n$, median: 1/2 above, 1/2 below
How choose?
traditional: do what your group has always done
better: does it matter? symmetric distributions: no. use mean: more precise
if does: what's the question? total $\Rightarrow$ mean, typical individual $\Rightarrow$ median.

F: Spread:
sample variance: $s^2 = \Sigma(X_i - \overline{X})^2/(n-1)$
standard deviation: $s = \sqrt{\text{Variance}}$
coefficient of variation, cv: $s/\overline{X}$, measures relative variation.

G: Inference: two types: Confidence interval, hypothesis test
Can be based on how the data were collected (design-based inference)
Or on a probability model for the data (model-based inference)
Spoiler: multiple opinions and lots of controversy.
Lots of tricky epistemological and practical issues. I give the traditional approach

H: Hypothesis test:
Assume something (null hypothesis) and ask whether observations are unusual
p-value: probability of observing data or something more extreme given null hypothesis
Does not prove null hypothesis or show that alternative more probable than null.

Scale of evidence: Display 2.12
p > 0.10: no evidence against null, (not: "null is true", e.g., not: "X has no effect")
0.05 < p < 0.10: weak evidence for the alternative
0.01 < p < 0.05: (moderate) evidence for the alternative
0.001 < p < 0.01: strong evidence for the alternative
p < 0.001: (convincing) very strong evidence for the alternative

> "All we know about the world teaches us that the effects of A and B are always different - in some decimal place - for any A and B. Thus asking "are the effects different?" is foolish" (John Tukey)

I: Randomization test: design-based inference
randomly reassign labels, compute S = statistic of interest (e.g., difference in means)
p = P[observed S as or more extreme than random S], e.g.,
Compare means of two populations, choose $S = \mid \overline{Y}_A - \overline{Y}_B \mid$
p $= \mathrm{P}\Big[\text{observed } \mid \overline{Y}_A - \overline{Y}_B \mid \geq \text{ random } \mid \overline{Y}_A - \overline{Y}_B\Big]$