University *of* New Haven

Descibing Data

BANL 6100: Business Analytics

Mehmet Balcilar

University of New Haven

Fall 2023

sampling & sources of bias

- census vs. sample
- sources of bias
- sampling methods

Young, Underemployed, and Optimistic

Coming of Age, Slowly, in a Tough Economy

Young adults hit hard by the recession. A plurality of the public (41%) believes young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.

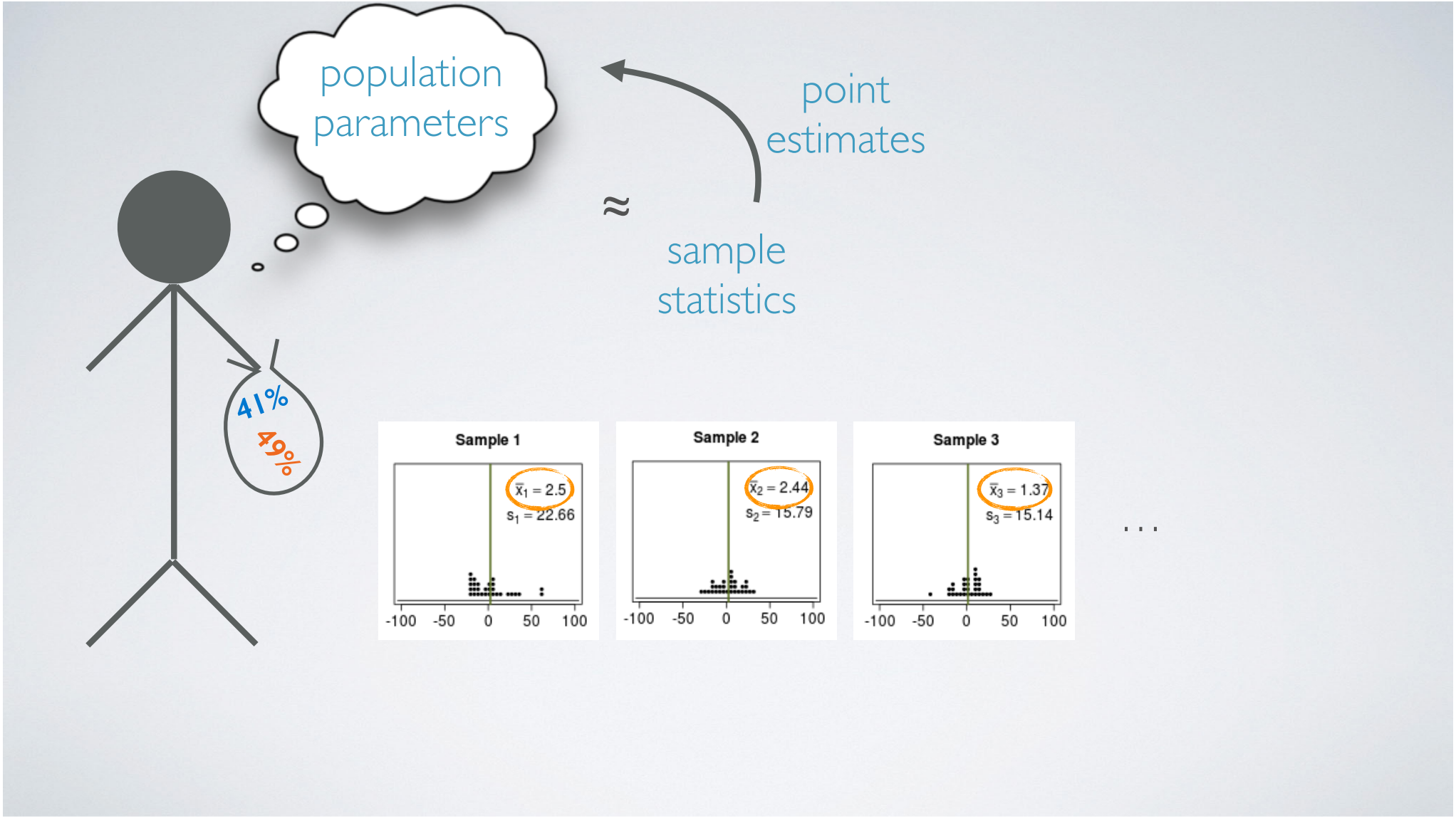
Tough economic times altering young adults' daily lives, long-term plans. While negative trends in the labor market have been felt most acutely by the youngest workers, many adults in their late 20s and early 30s have also felt the impact of the weak economy. Among all 18- to 34-year-olds, fully half (49%) say they have taken a job they didn't want just to pay the bills, with 24% saying they have taken an unpaid job to gain work experience.

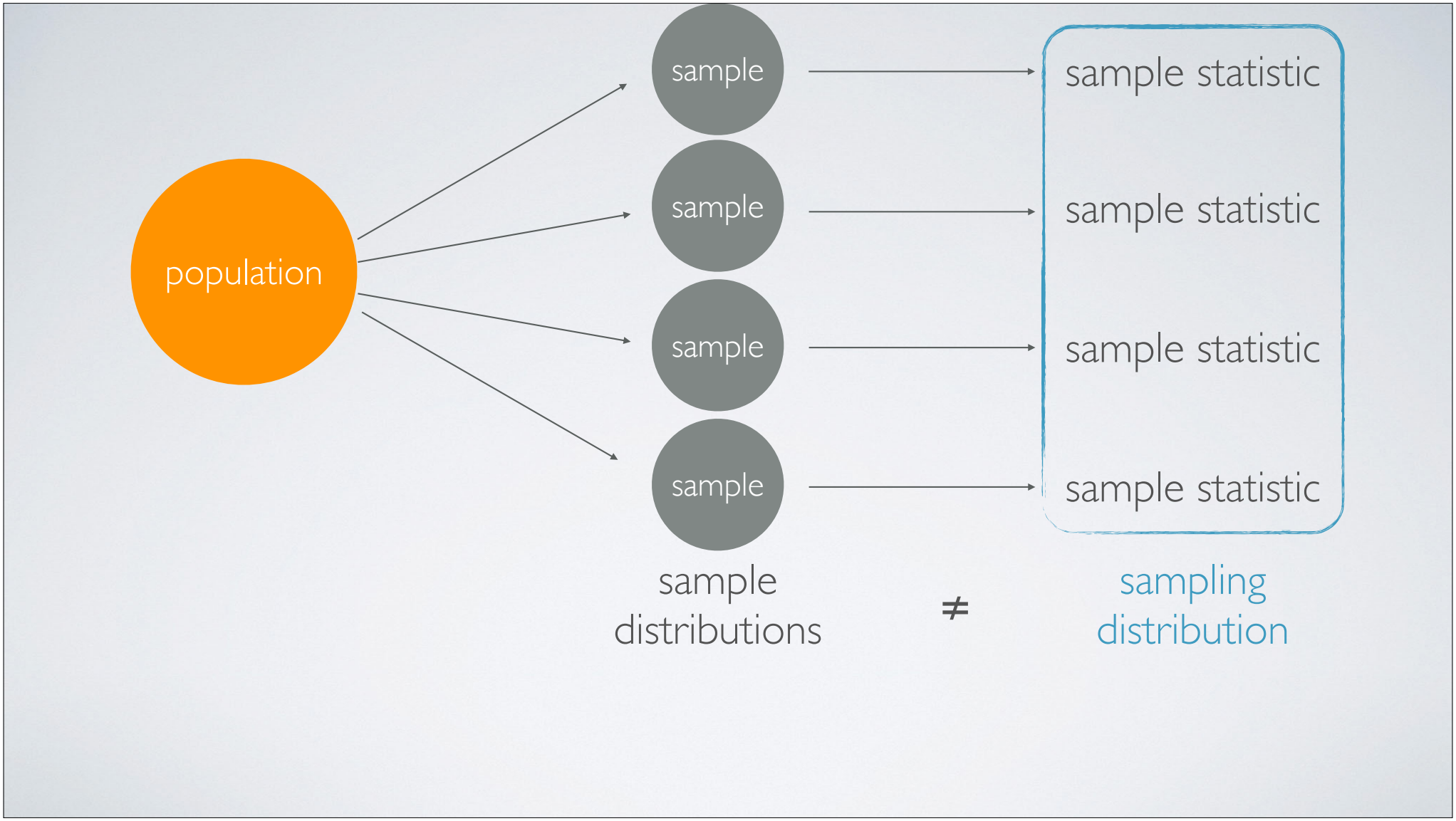
Link to poll: http://www.pewsocialtrends.org/2012/02/09/young-underemployed-and-optimistic/

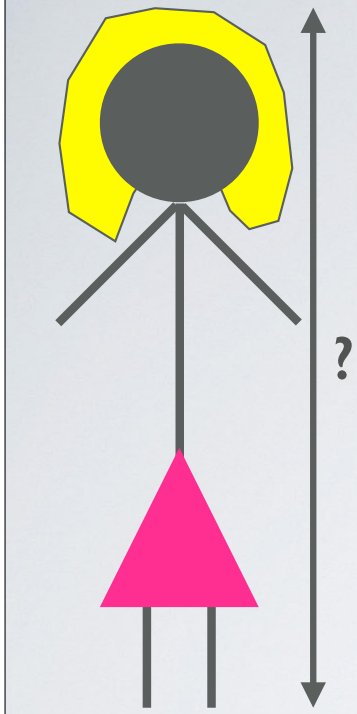
The general public survey is based on telephone interviews conducted Dec. 6-19, 2011, with a nationally representative sample of 2,048 adults ages 18 and older living in the continental United States […] Margin of sampling error is plus or minus 2.9 percentage points for results based on the total sample and 4.4 percentage points for adults ages 18-34 at the 95% confidence level.

- 41% ± 2.9%: We are 95% confident that 38.1% to 43.9% of the public believe young adults, rather than middle-aged or older adults, are having the toughest time in today's economy.

- 49% ± 4.4%: We are 95% confident that 44.6% to 53.4% of 18-34 years olds have taken a job they didn't want just to pay the bills.


population parameters
point estimates
≈
sample statistics
41%
49%
Sample 1
$\bar{x}_1 = 2.5$
$s_1 = 22.66$
Sample 2
$\bar{x}_2 = 2.44$
$s_2 = 15.79$
Sample 3
$\bar{x}_3 = 1.37$
$s_3 = 15.14$
-100 -50 0 50 100
...


population
sample
sample
sample
sample
sample statistic
sample statistic
sample statistic
sample statistic
sample distributions
≠
sampling distribution



US
women
N = pop size
μ

AL: $x_{AL,1}, x_{AL,2}, \cdots, x_{AL,1000}$

NC: $x_{NC,1}, x_{NC,2}, \cdots, x_{NC,1000}$

WY: $x_{WY,1}, x_{WY,2}, \cdots, x_{WY,1000}$

$\bar{x}_{AL}$ ... $\bar{x}_{NC}$ ... $\bar{x}_{WY}$

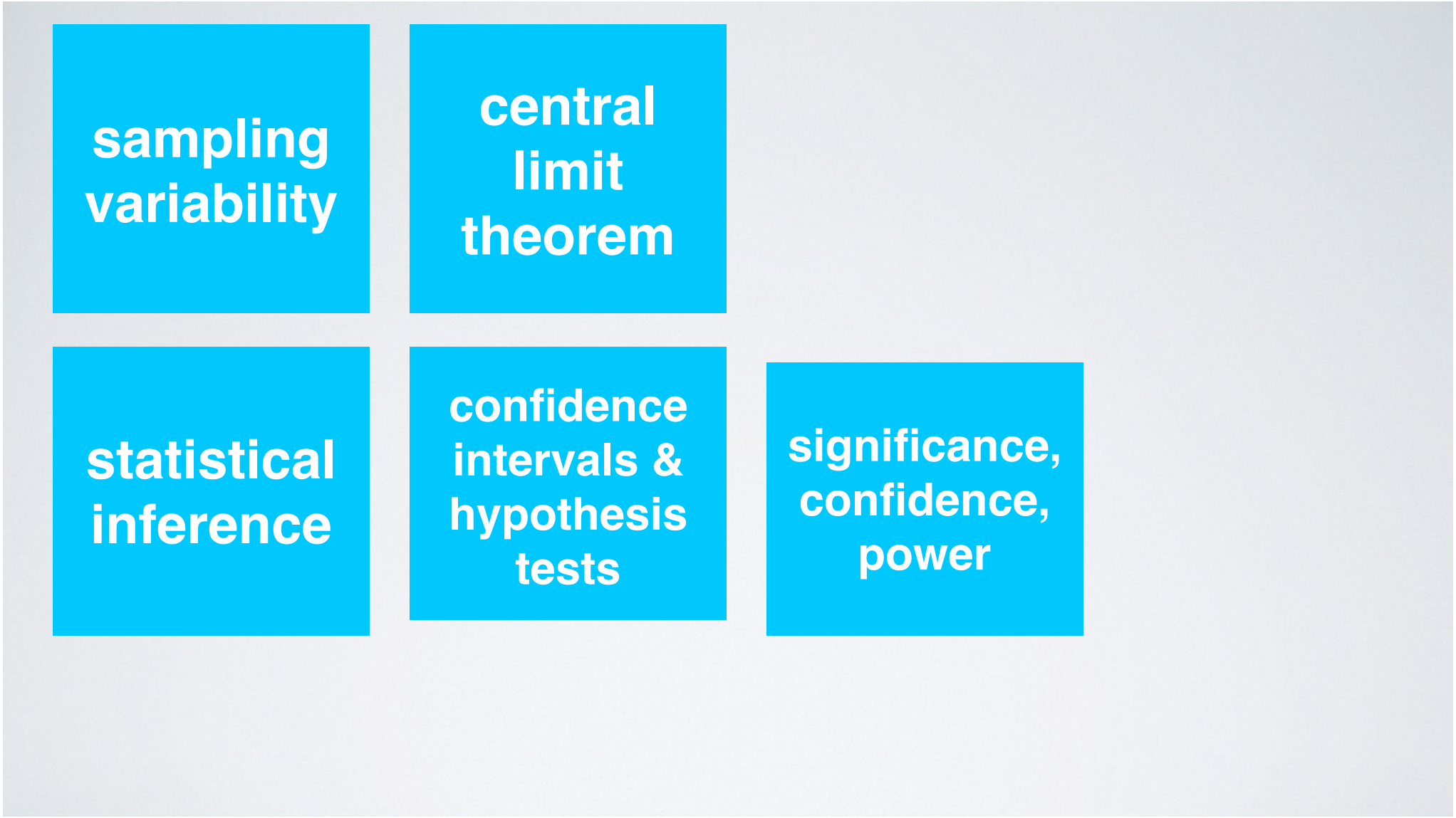
sampling distribution

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}$$

$mean(\bar{x}) \approx \mu$

$n \uparrow$ ↓ standard error $SD(\bar{x}) < \sigma$


sampling variability
central limit theorem
statistical inference
confidence intervals & hypothesis tests
significance, confidence, power

census

Wouldn't it be better to just include everyone and "sample" the entire population, i.e. conduct a census?

- Some individuals are hard to locate or measure, and these people may be different from the rest of the population.
- Populations rarely stand still.

Illegal Immigrants Reluctant To Fill Out Census Form

by PETER O'DOWD

March 31, 2010 4:00 AM

npr from KJZZ

There is an effort underway to make sure Hispanics are accurately counted in the 2010 Census. Phoenix has some of the country's "hardest-to-count" districts. Some Latinos, especially illegal residents, fear that participating in the count will expose them to immigration raids or government harassment.

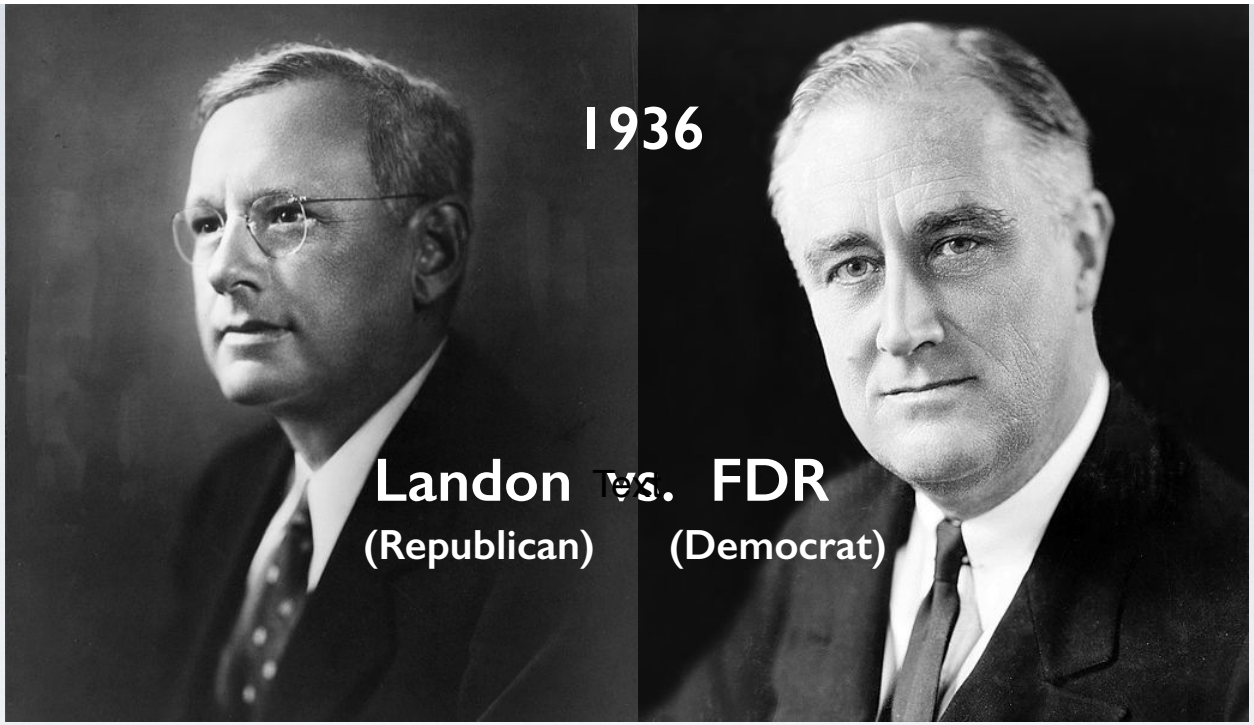
Listen to the NPR story at http://www.npr.org/templates/story/story.php?storyId=125380052

Image credit: Wonderlane CC BY 2.0 http://www.flickr.com/photos/wonderlane/6231888661

a few sources of sampling bias

- Convenience sample: Individuals who are easily accessible are more likely to be included in the sample
- Non-response: If only a (non-random) fraction of the randomly sampled people respond to a survey such that the sample is no longer representative of the population
- Voluntary response: Occurs when the sample consists of people who volunteer to respond because they have strong opinions on the issue

Poll source: edition.cnn.com, August 29, 2013



The Literary Digest

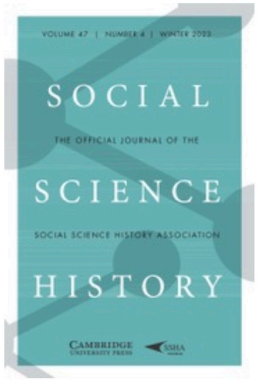
Election results

Lose with 43% of the votes

Win with 62% of the votes

Based on the poll, The Literary Digest predicted that Landon would win the 1936 presidential election with 57.1% of the popular vote and an electoral college margin of 370 to 161.

Image sources: http://en.wikipedia.org/wiki/File:LandonPortr.jpg, http://en.wikipedia.org/wiki/File:FDR_in_1933.jpg, and http://en.wikipedia.org/wiki/File:LiteraryDigest-19210219.jpg



Social Science History

Article contents

"President" Landon and the 1936 *Literary Digest* Poll

Were Automobile and Telephone Owners to Blame?

Published online by Cambridge University Press: **04 January 2016**

Dominic Lusinchi

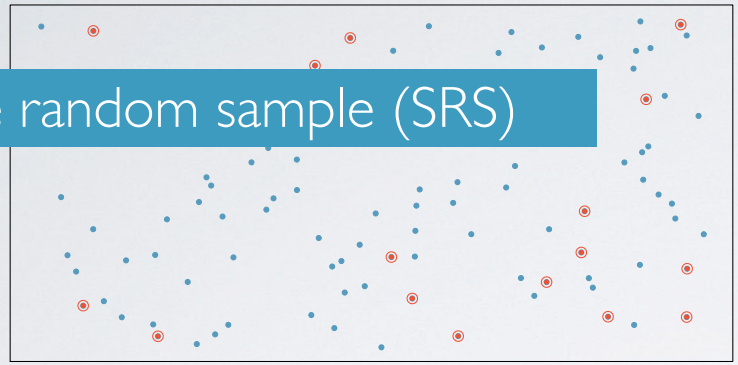
Abstract

The disastrous prediction of an Alf Landon victory in the 1936 presidential election by the Literary Digest poll is a landmark event in the history of American survey research in general and polling in particular. It marks both the demise of the straw poll, of which the Digest was the most conspicuous and well-regarded example, and the rise to prominence of the self-proclaimed "scientific" poll. Why did the Digest poll fail so miserably? One view has come to prevail over the years: because the Digest selected its sample primarily from telephone books and car registration lists and since these contained, at the time, mostly well-to-do folks who would vote Republican, it is no wonder the magazine mistakenly predicted a Republican win. This "conventional explanation" has found its way into countless publications (scholarly and in the press) and college courses. It has been used to illustrate the disastrous effects of a poorly designed poll. But is it correct? Empirical evidence, in the form of a 1937 Gallup poll, shows that this "conventional explanation" is wrong, because voters with telephones and cars backed Franklin D. Roosevelt and because it was those who failed to participate in the poll (overwhelmingly supporters of Roosevelt) who were mainly responsible for the faulty prediction.

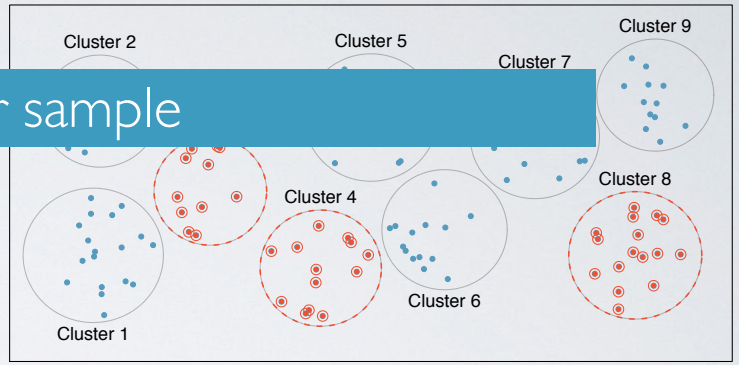
Image credit: Wonderlane CC BY 2.0 http://www.flickr.com/photos/wonderlane/6231888661

sampling methods

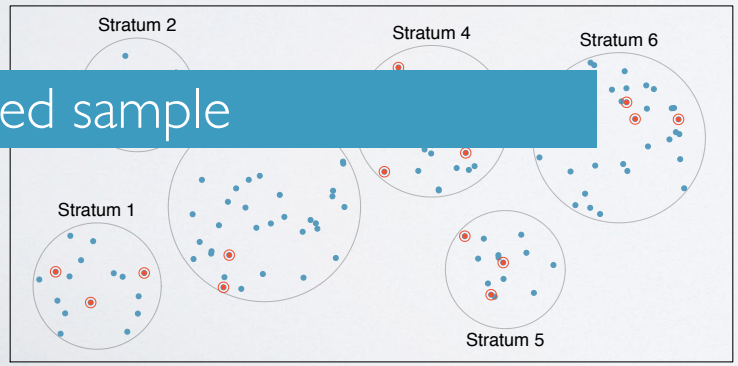
simple random sample (SRS)



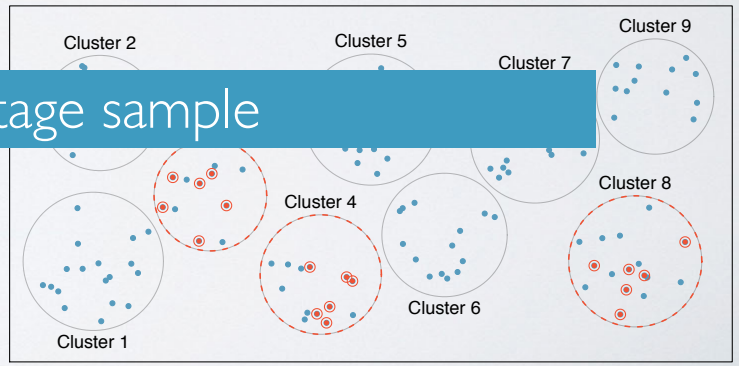
cluster sample



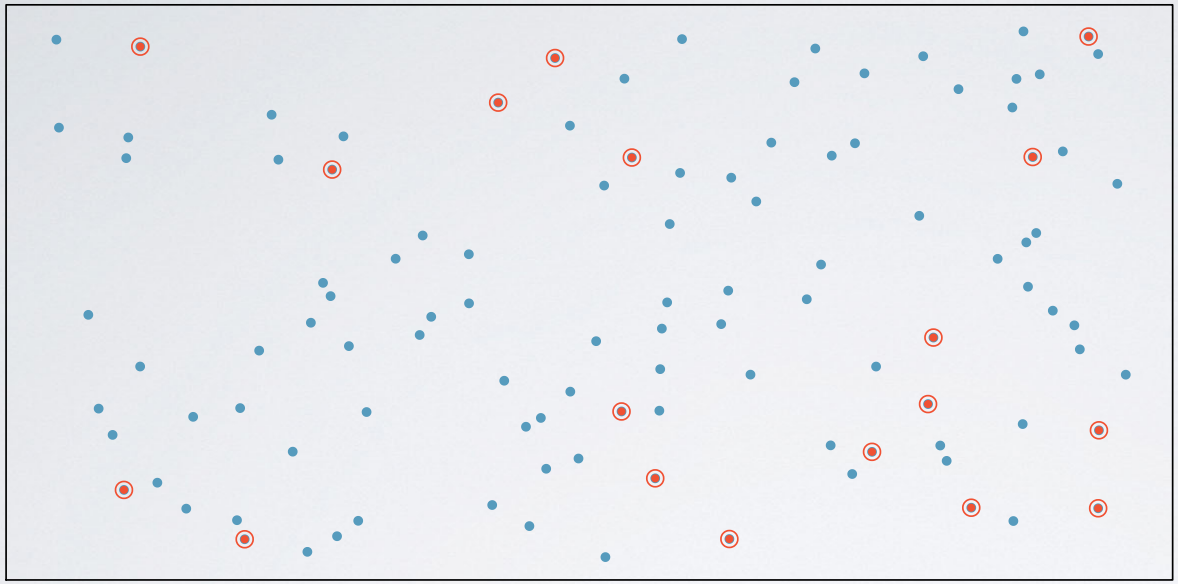
stratified sample



multistage sample

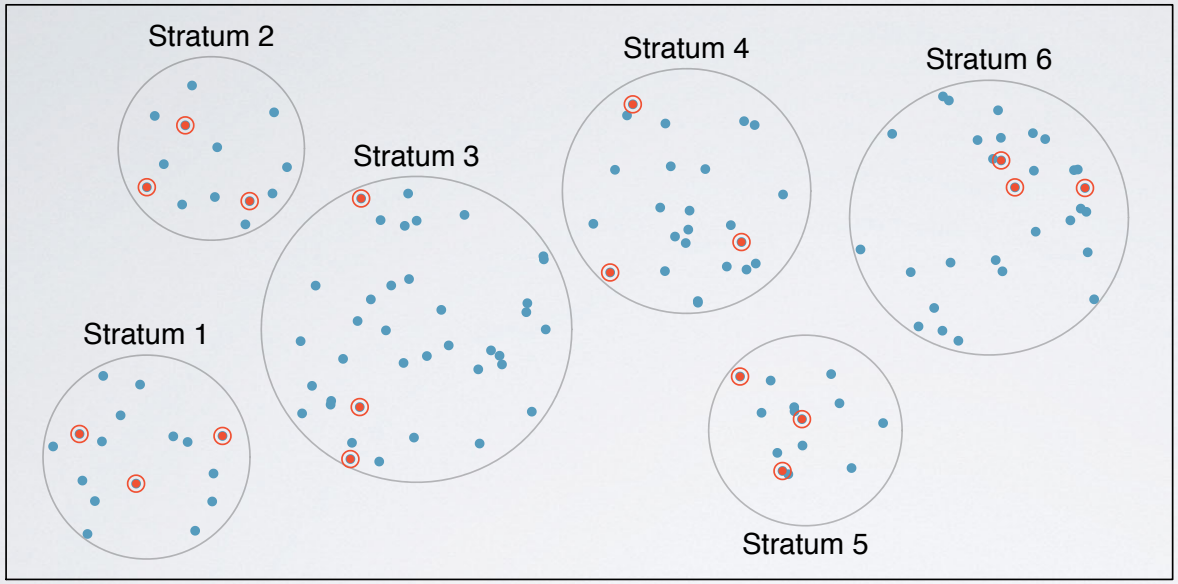


simple random sample (SRS)



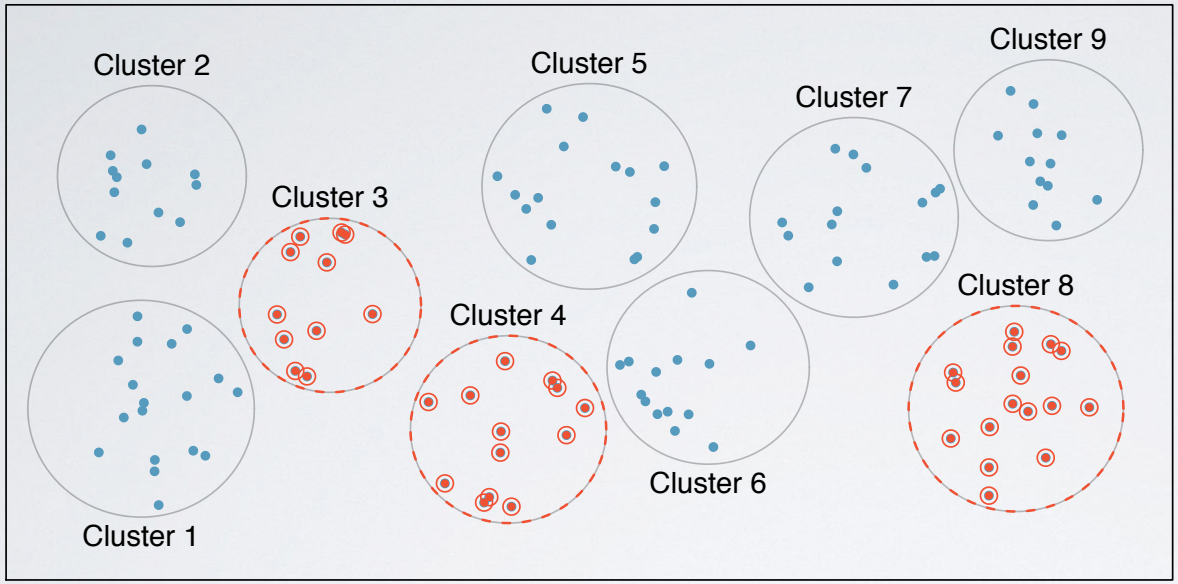
each case is equally likely to be selected

stratified sample



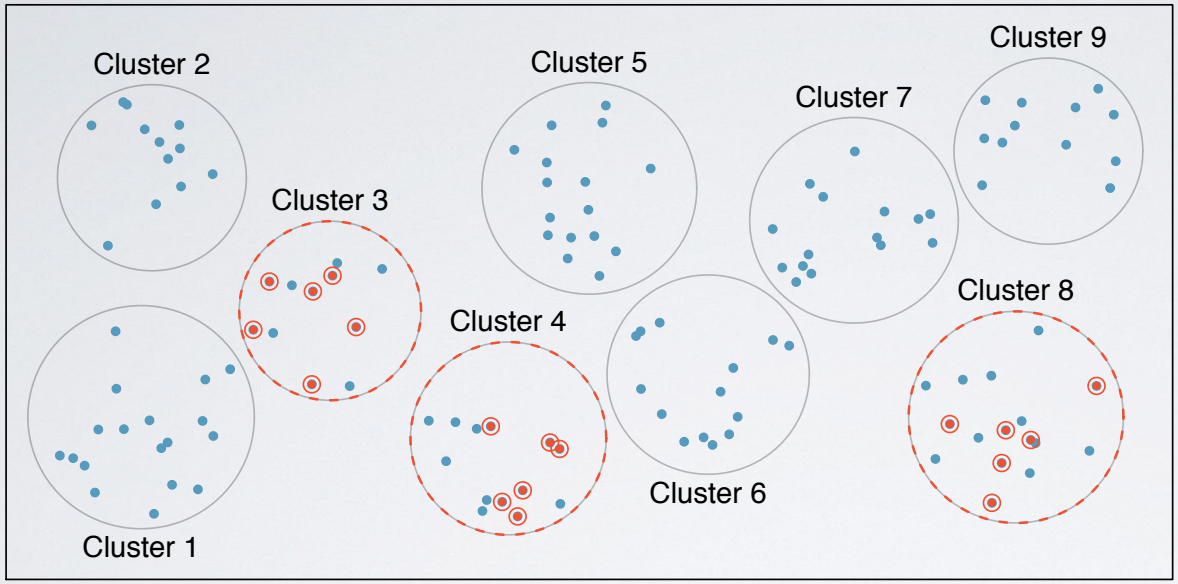
divide the population into homogenous strata,
then randomly sample from within each stratum

cluster sample



divide the population clusters,
randomly sample a few clusters,
then sample all observations within these clusters

multistage sample

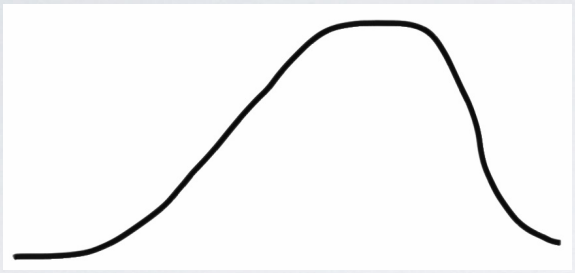


divide the population clusters,
randomly sample a few clusters,
then randomly sample within these clusters

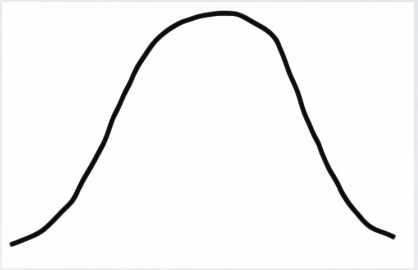
measures of center

shape

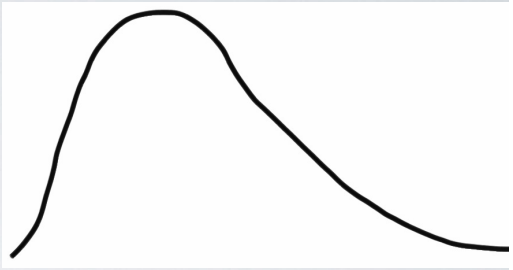
skewness



left skewed

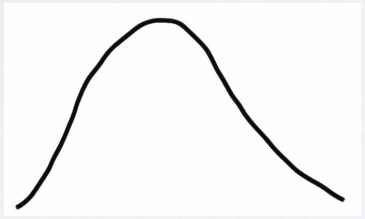


symmetric

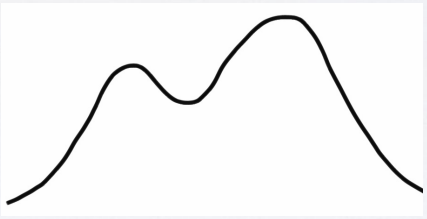


right skewed

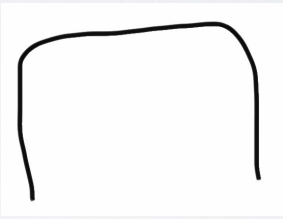
modality



unimodal



bimodal



uniform

multimodal

measures of center

mean

arithmetic average

$\bar{x}$ sample mean

μ population mean

median

midpoint of the distribution (50th percentile)

mode

most frequent observation

sample statistic

point estimate

population parameter

Mean

- The *sample mean*, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \cdots, x_n$ represent the n observed values.
- The *population mean* is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.
- The sample mean is a *sample statistic*, and serves as a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population), it is usually a pretty good estimate.

Median

- The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

- If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

- Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50^{th} *percentile*.

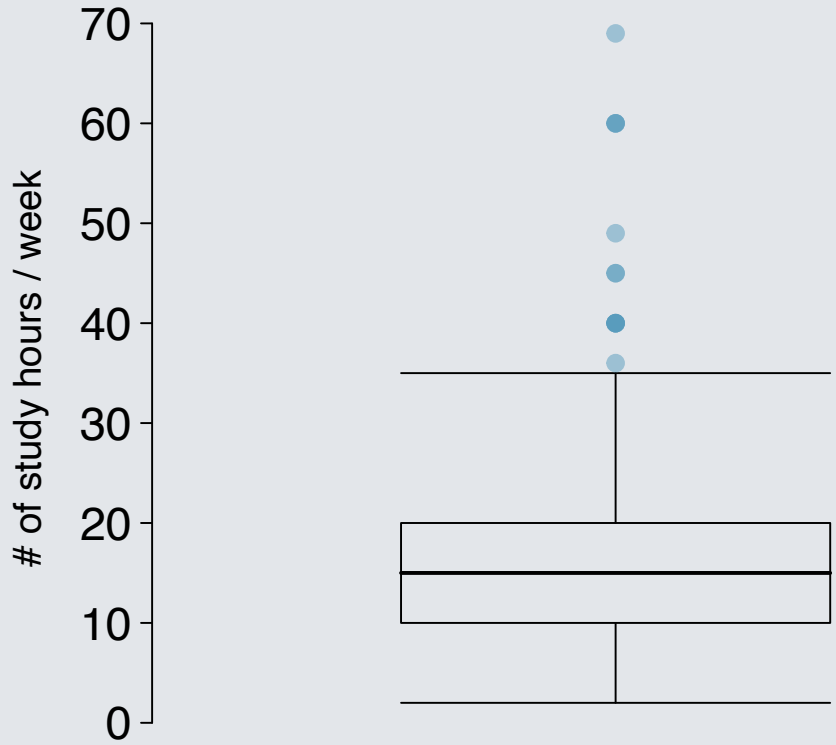
Q1, Q3, and IQR

- The 25^{th} percentile is also called the first quartile, *Q1*.
- The 50^{th} percentile is also called the median.
- The 75^{th} percentile is also called the third quartile, *Q3*.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the *interquartile range*, or the *IQR*.

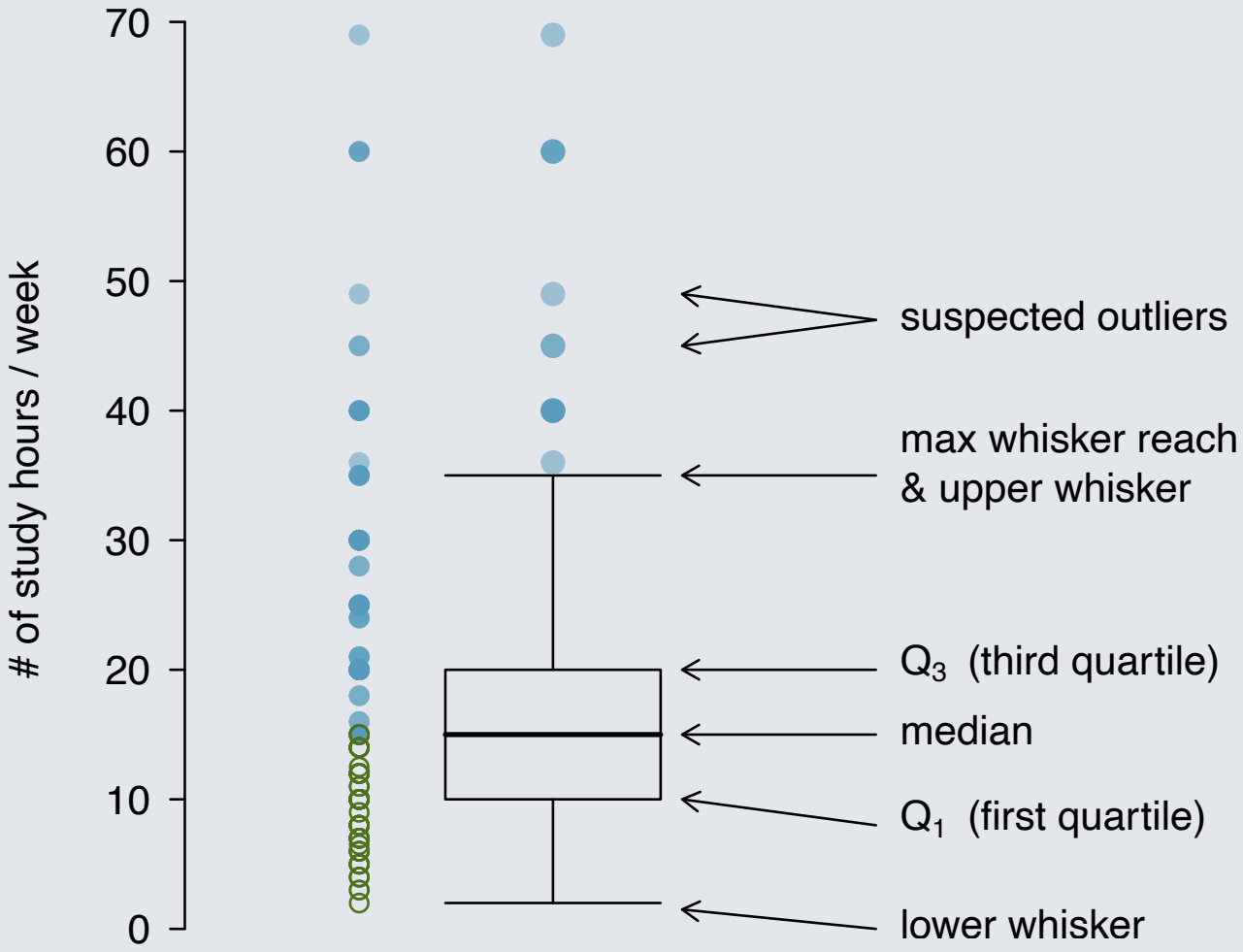
$$IQR = Q3 - Q1$$

Box plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a box plot



Whiskers and outliers

- *Whiskers*
  of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

example 9 students' exam scores:

75, 69, 88, 93, 95, 54, 87, 88, 27

mean: $\frac{75+69+88+93+95+54+87+88+27}{9} = 75.11$

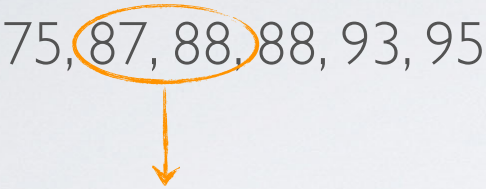
mode: 88

median: 27, 54, 69, 75, 87, 88, 88, 93, 95

example

10 students' exam scores:

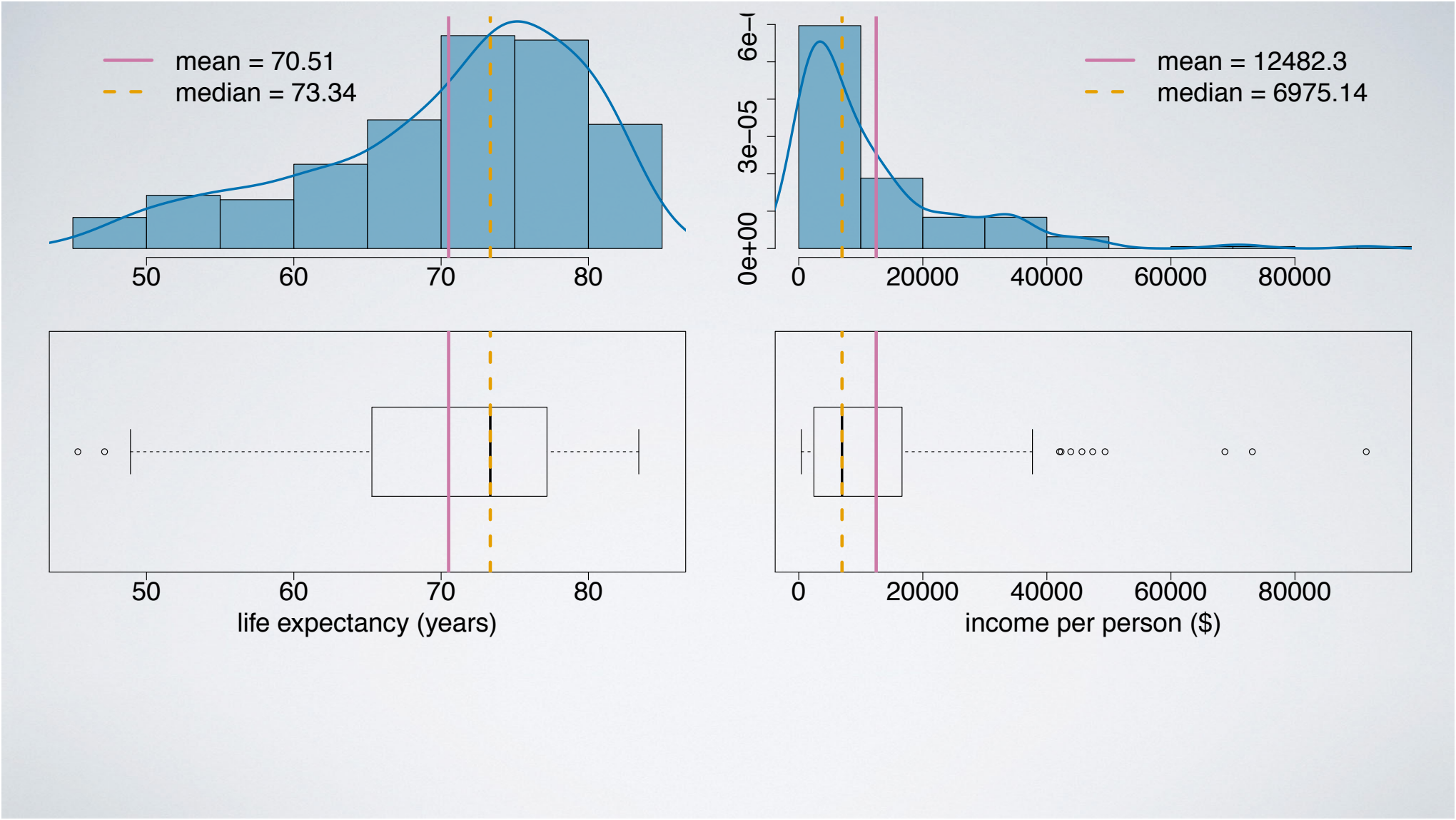
27, 54, 69, 75, 87, 88, 88, 93, 95, 100



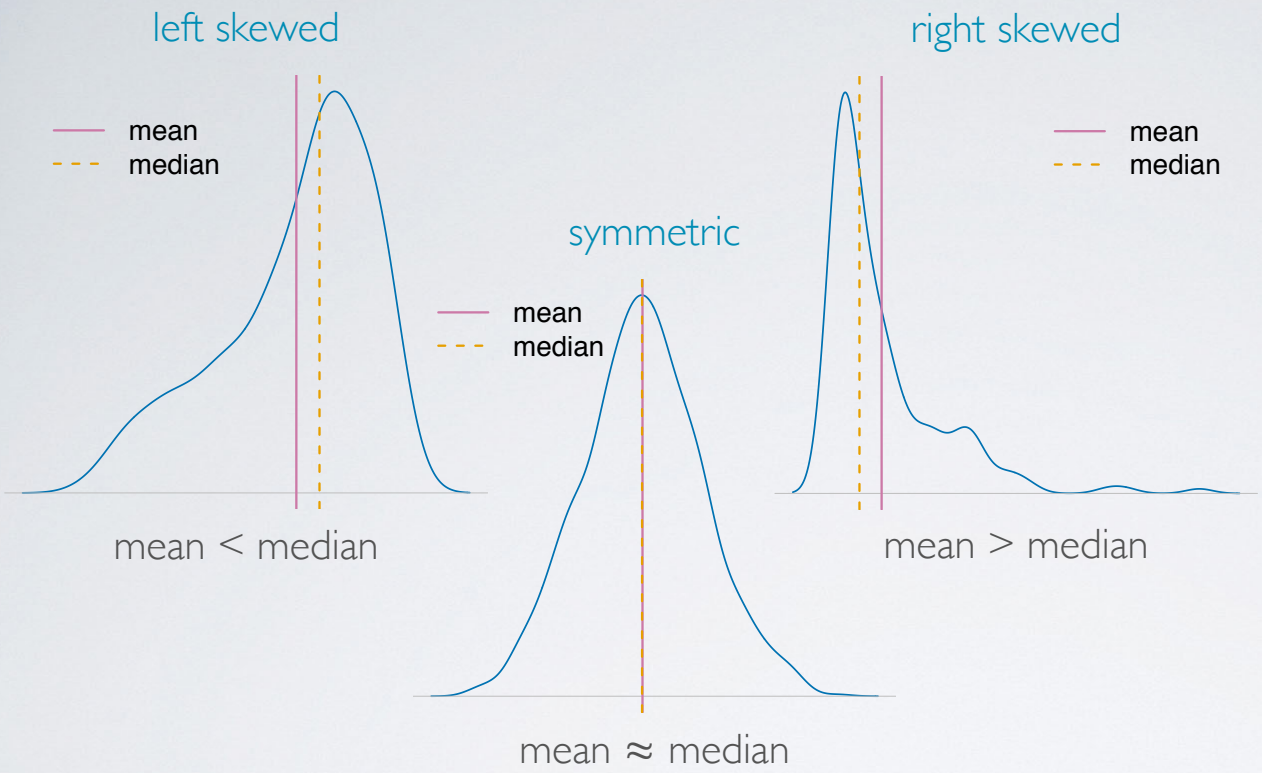
$$\frac{87 + 88}{2} = 87.5$$

data	income per person ($, 2012)	life expectancy (years, 2012)
Afghanistan	1359.7	60.254
Albania	6969.3	77.185
Algeria	6419.1	70.874
…	…	…
Zimbabwe	545.3	58.142

Source: gapminder.com

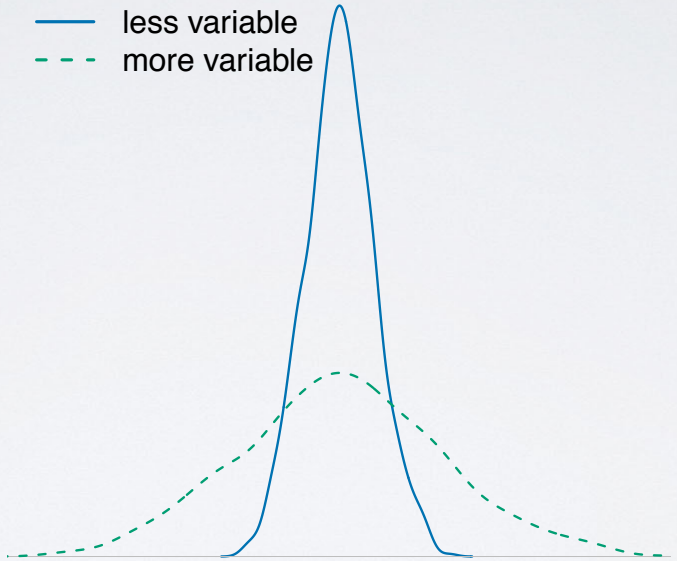

mean = 70.51
median = 73.34
50
60
70
80
mean = 12482.3
median = 6975.14
0e+00
3e−05
6e−
0
20000
40000
60000
80000
life expectancy (years)
income per person ($)

skewness vs. measures of center



measures of spread

- range: *(max - min)*
- variance
- standard deviation
- inter-quartile range



variance

sample variance s^2

population variance σ^2

roughly the average squared deviation from the mean

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

example Given that the average life expectancy is 70.5, and there are 201 countries in the dataset:

$$s^2 = \frac{(60.3 - 70.5)^2 + (77.2 - 70.5)^2 + \ldots + (58.1 - 70.5)^2}{201 - 1}$$

$= 83.06 \text{ years}^2$

	country	life exp
1	Afghanistan	60.3
2	Albania	77.2
3	Algeria	70.9
	…	…
201	Zimbabwe	58.1

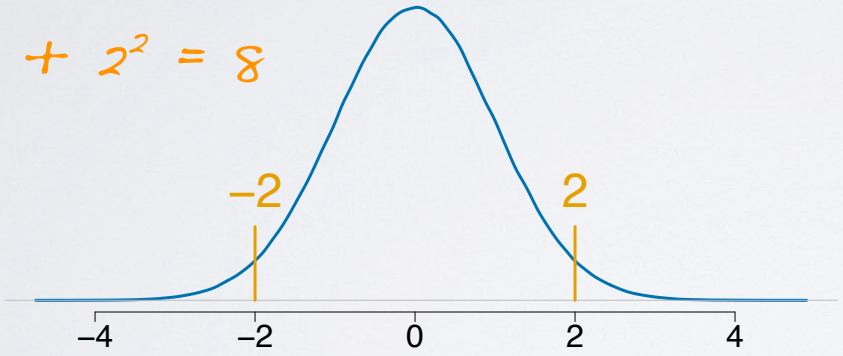
Why do we square the differences?

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

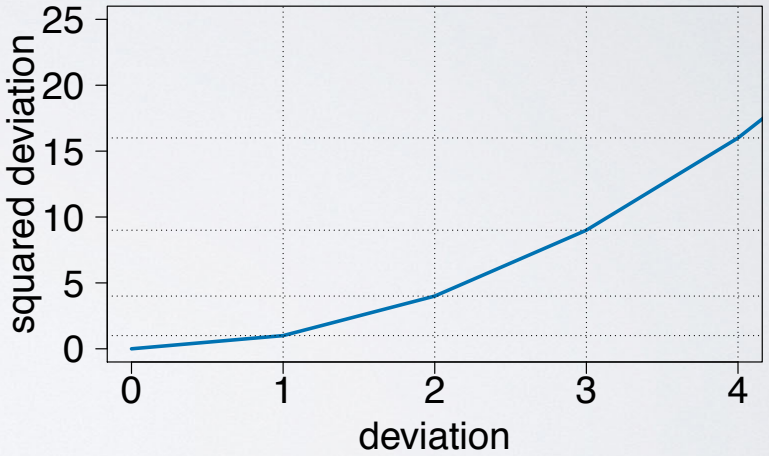
- get rid of negatives so that negatives and positives don't cancel each other when added together

$(-2) + 2 = 0$

$(-2)^2 + 2^2 = 8$



- increase larger deviations more than smaller ones so that they are weighed more heavily



standard deviation

sample sd
s

population sd
σ

roughly the average deviation around the mean, and has the same units as the data.

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

square root of the variance

variability vs. diversity

Which of the following sets of cars has a more **diverse** composition of colors?

☐ SET 1

☐ SET 2

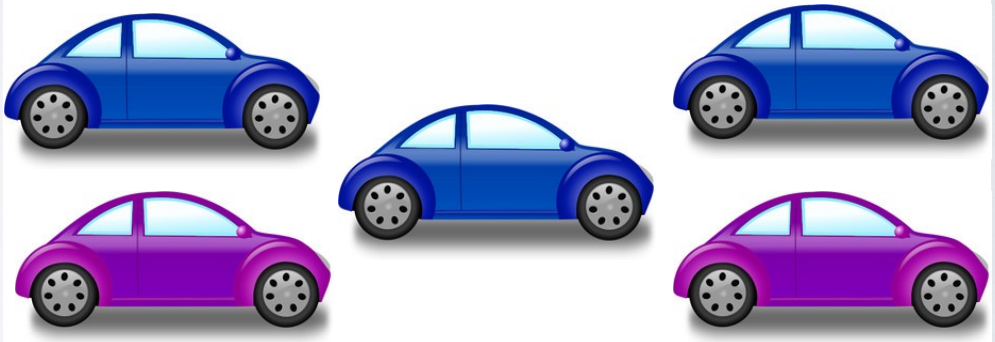


Image source: http://direct.clipart.co.za/

variability vs. diversity

Which of the following sets of cars has a more **diverse** composition of colors?

☑ SET 1

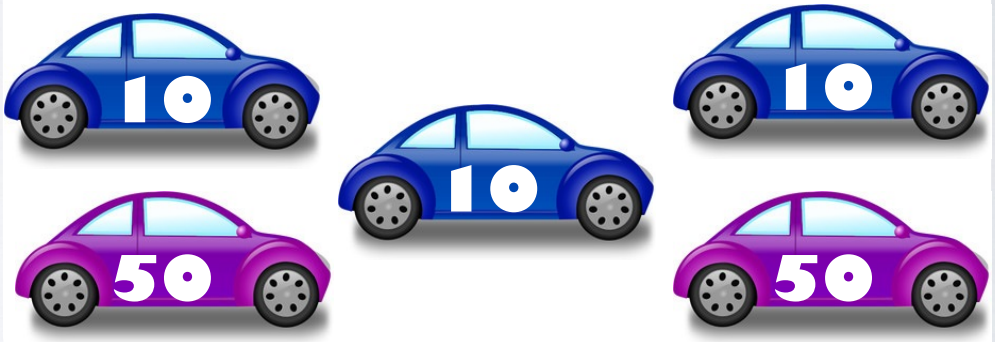
☐ SET 2

variability vs. diversity

Which of the following sets of cars has more **variable** mileage?

☐ SET 1

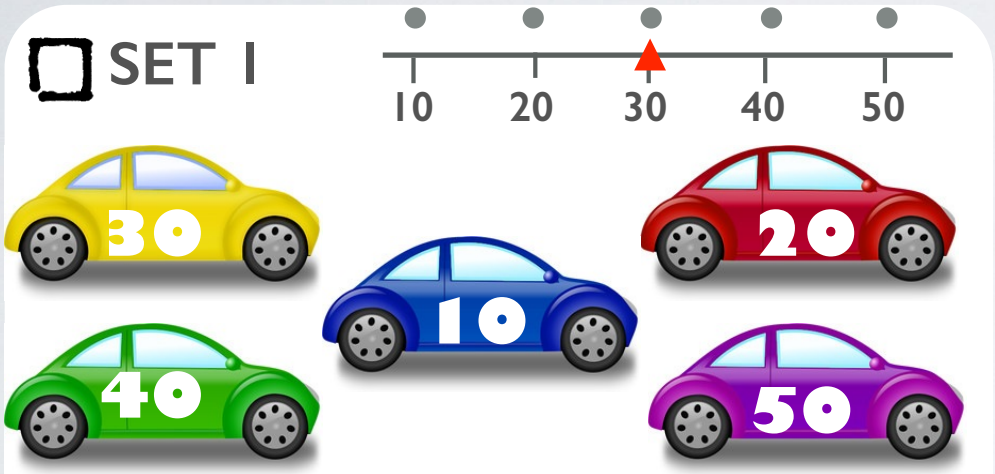
☐ SET 2



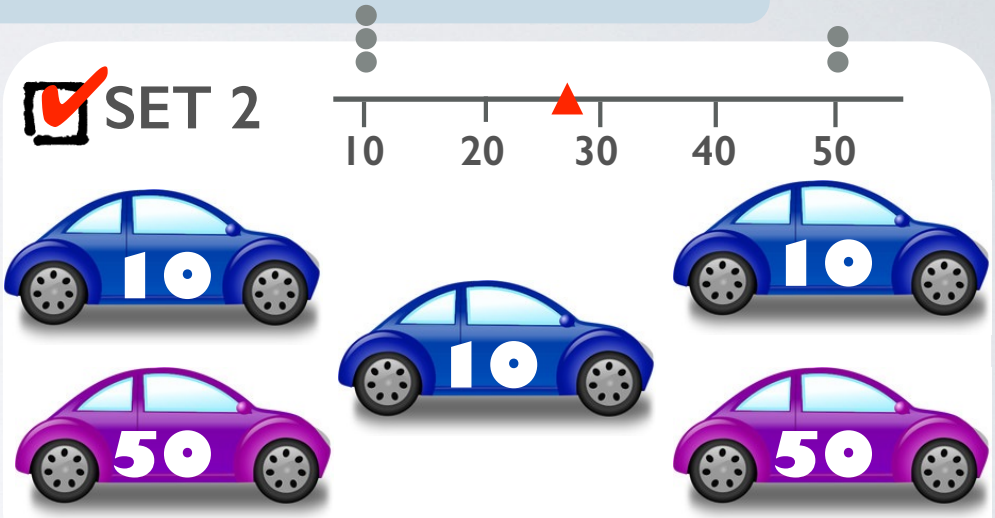
variability vs. diversity

Which of the following sets of cars has more **variable** mileage?

☐ SET 1



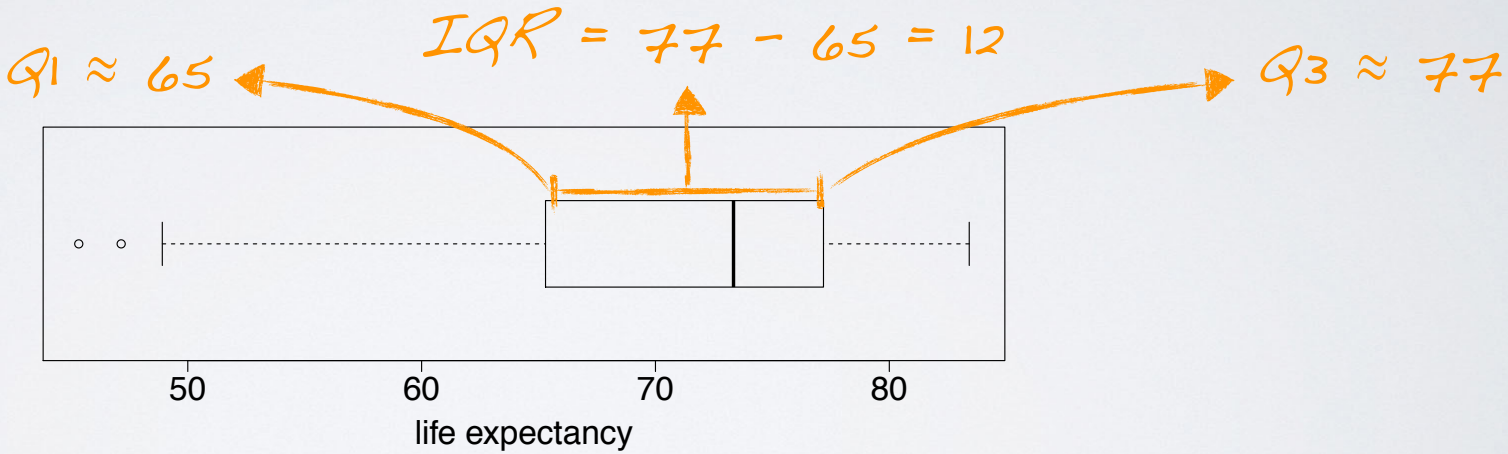
☑ SET 2



interquartile range

range of the middle 50% of the data, distance between the first quartile (25th percentile) and third quartile (75th percentile)

$$IQR = Q3 - Q1$$



Coefficient of Variation

- The *coefficient of variation* statistic measures of the spread of a set of measurements of a sample.
- It allows us to directly compare variation in samples measured with different units, or with very different means.
- It is also known as relative standard deviation (RSD), and defined as the standard deviation divided by the mean:

$$c_v = \frac{s}{\bar{x}}$$

- c_v should only be computed for data measured on a *ratio scale (having a real zero)*, as these are the measurements that can only take non-negative values.

robust statistics

- define robust statistics
- robust measures of center & spread

robust statistics

we define robust statistics as measures on which extreme observations have little effect

example

data	mean	median
1, 2, 3, 4, 5, 6	3.5	3.5
1, 2, 3, 4, 5, 1000	169	3.5

	robust	non-robust
center	median	mean
spread	IQR	SD, range

transforming data

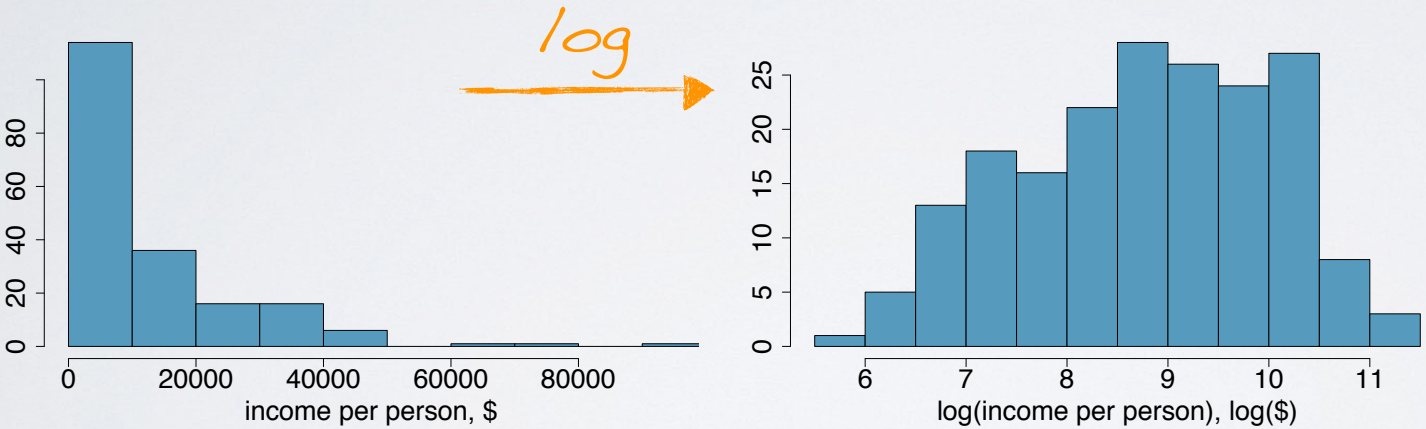
- define transformations
- review when it might be useful/ necessary to transform data

transformations

- a transformation is a rescaling of the data using a function
- when data are very strongly skewed, we sometimes transform them so they are easier to model

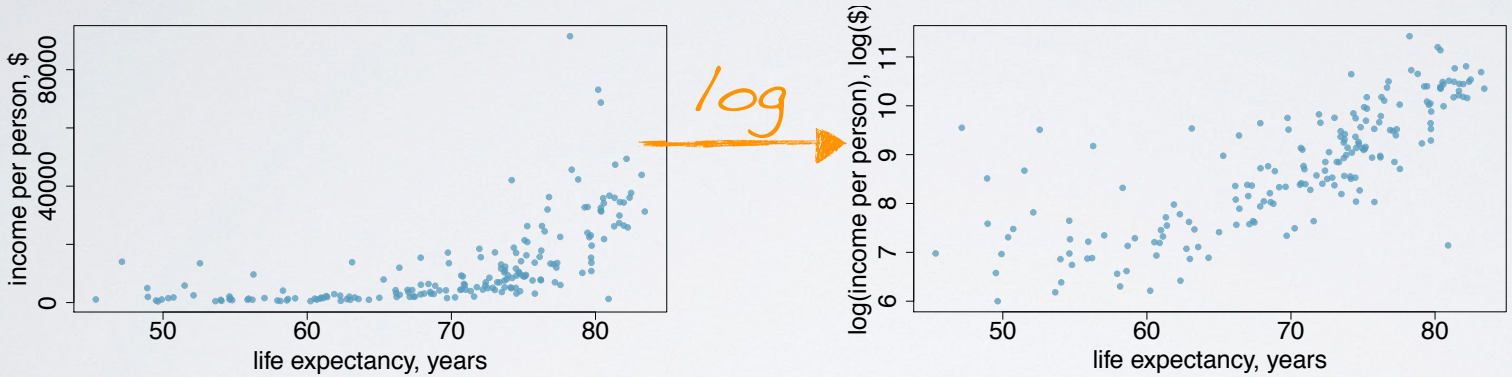
(natural) log transformation

often applied when much of the data cluster near zero (relative to the larger values in the data set) and all observations are positive

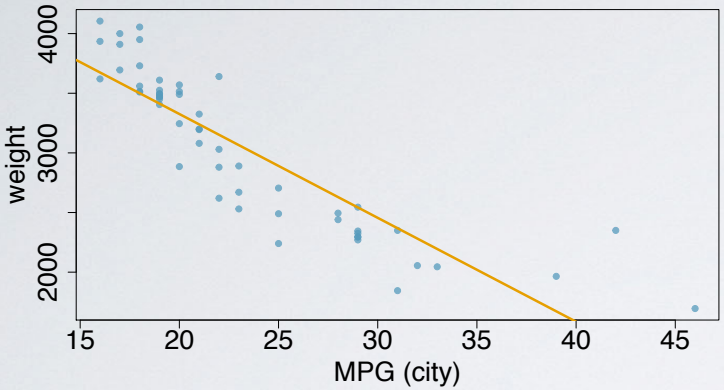


log transformation

to make the relationship between the variables more linear, and hence easier to model with simple methods

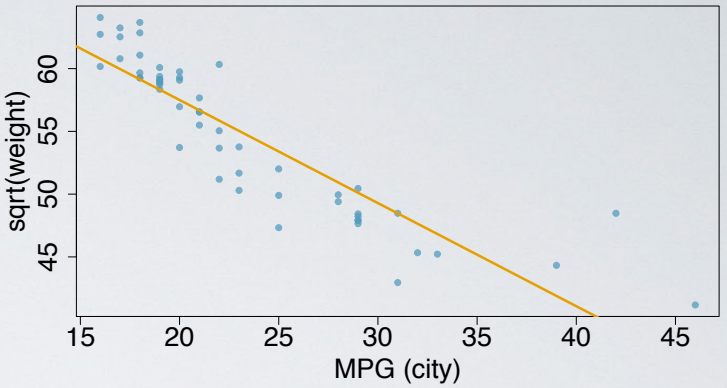


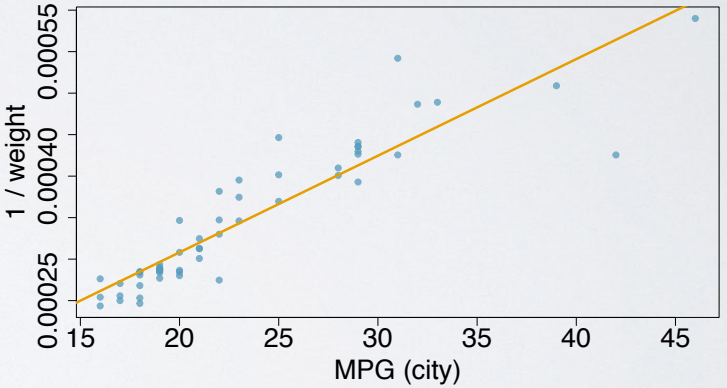
other transformations



square root

inverse





goals of transformations

- to see the data structure differently
- to reduce skew assist in modeling
- to straighten a nonlinear relationship in a scatterplot