INTRODUCTION TO DATA SCIENCE

Lecture based on:

M. Cetinkays-Rundel, „Data Analysis and Statistical Inference", Univ. of Duke
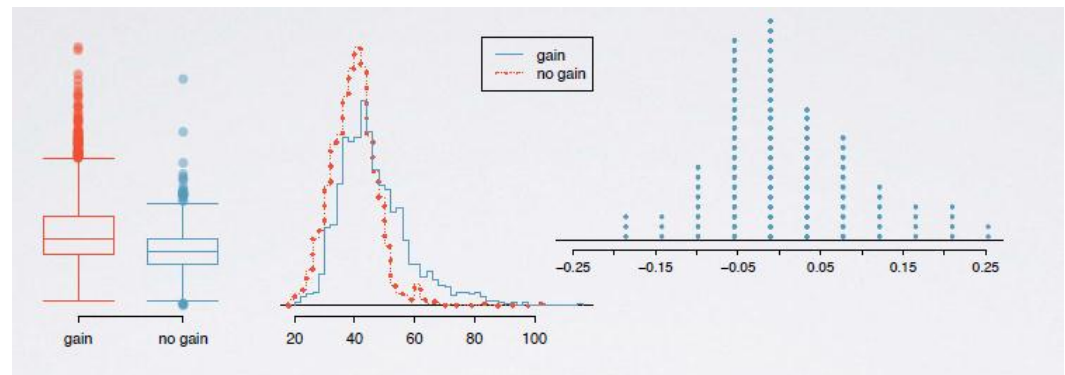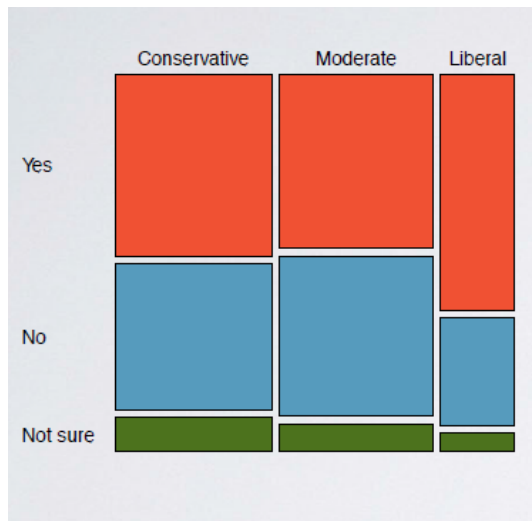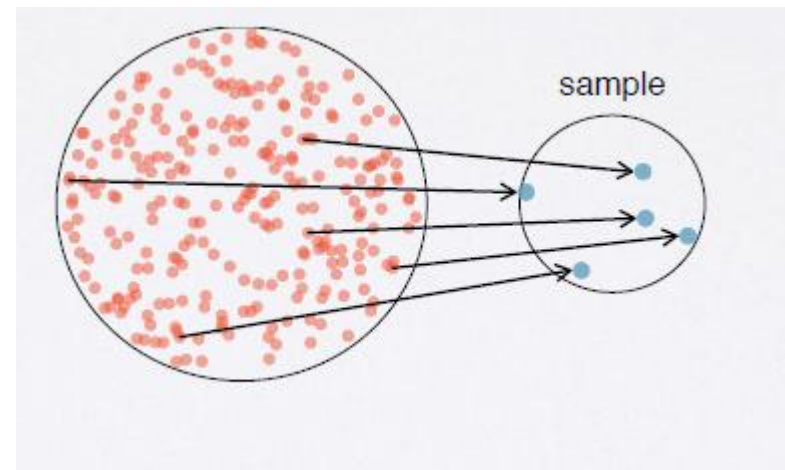
WFAiS UJ, Informatyka Stosowana

I stopień studiów

Exploratory data analysis

How to collect, visualise and interpret the data.

Exploratory data analysis

population	sample	design
scope	exploratory data analysis	inference

4/10/2023

Data: basics

- **Observations, variables, data matrices**
- **Type of variables**
- **Relationship between variables**

Example: data matrix

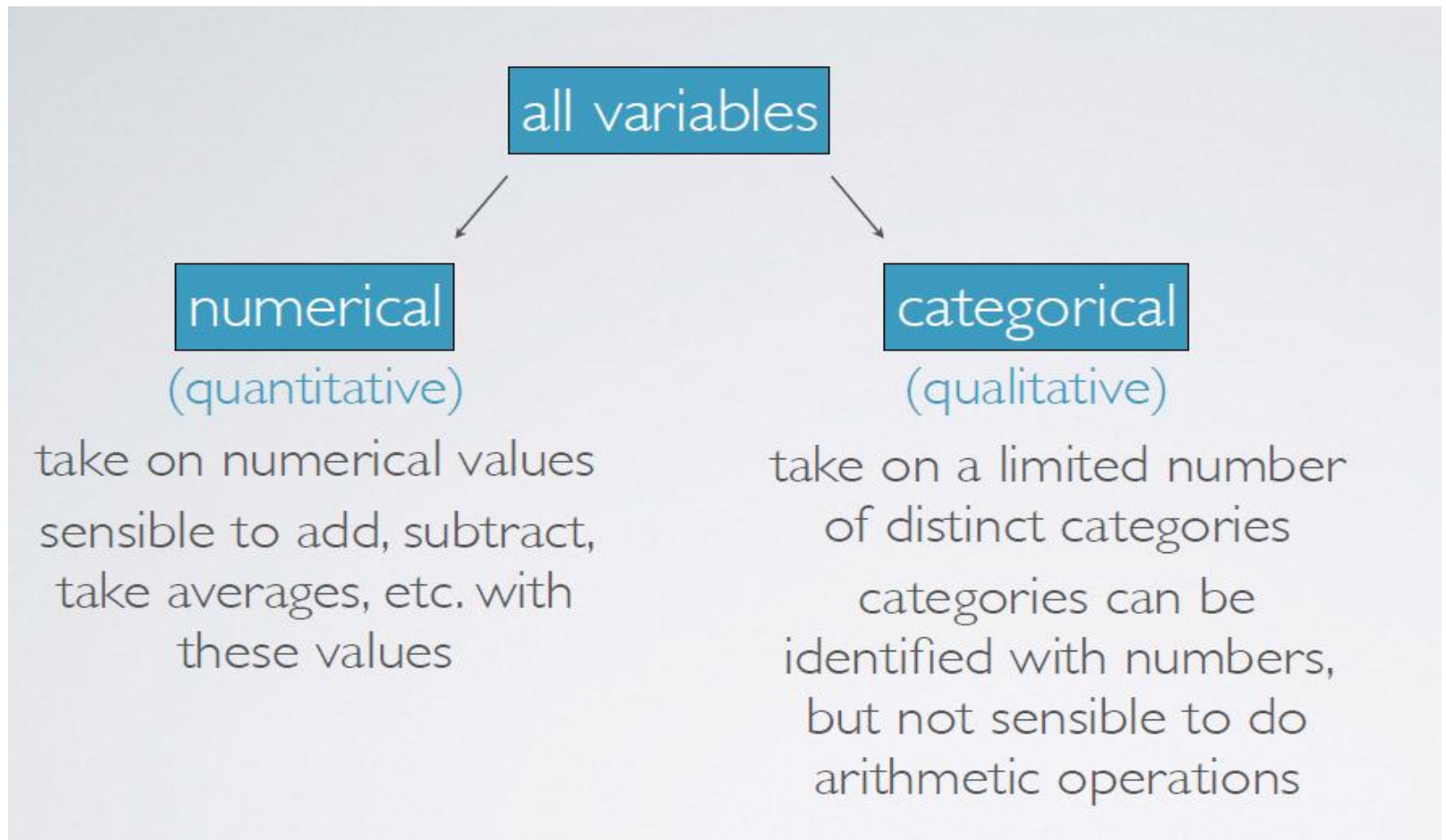**Requests send to Google to remove links from the search engine database.**

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

observation (case)

variable

Type of variables

all variables

numerical
(quantitative)

take on numerical values
sensible to add, subtract,
take averages, etc. with
these values

categorical
(qualitative)

take on a limited number
of distinct categories

categories can be
identified with numbers,
but not sensible to do
arithmetic operations

Numerical variables

all variables

numerical

categorical

continuous

discrete

take on any of an infinite number of values within a given range

take on one of a specific set of numeric values

4/10/2023

Categorical variables

Data matrix

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

country: Name of the country

4/10/2023

Data matrix

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

cr_req: Number of content removal requests made to Google

discrete numerical

Data matrix

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

cr_comply: Percentage of content removal requests Google complied with

continuous numerical

4/10/2023

Data matrix

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

ud_req: Number of user data requests as part of a criminal investigation

4/10/2023

Data matrix

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

`ud_comply`: Percentage of user data requests Google complied with

continuous numerical

4/10/2023

Data matrix

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

categorical `hemisphere`: Hemisphere that the country is located in (southern, northern)

Data matrix

country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...	...	...	...	...	...	...	...
United States	92	63	5950	93	...	northern	very high

hdi: Human Development Index
(very high, high, medium, low)

Relationships between variables

relationships between variables
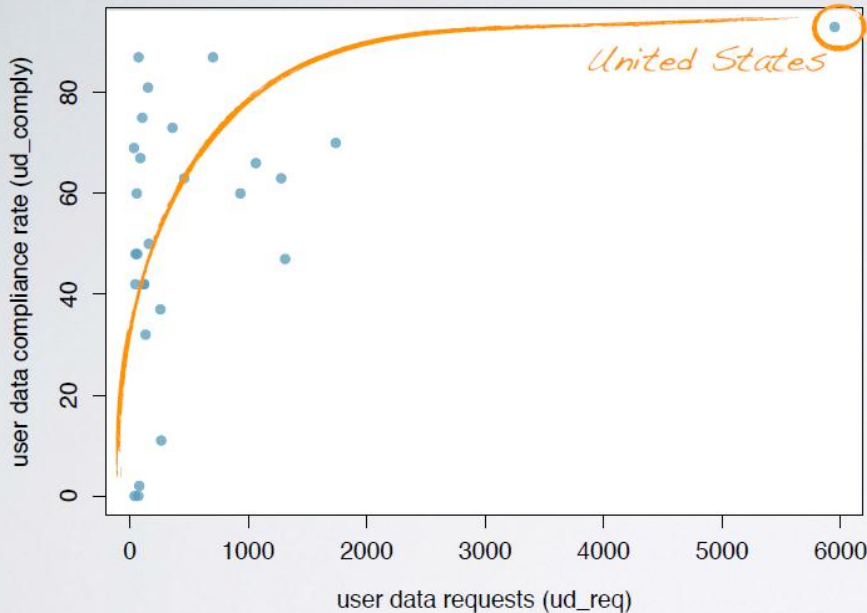
- ▸ Two variables that show some connection with one another are called associated (dependent)
- ▸ Association can be further described as positive or negative
- ▸ If two variables are not associated, they are said to be independent

4/10/2023

Observational studies & experiments

studies

observational

experiment
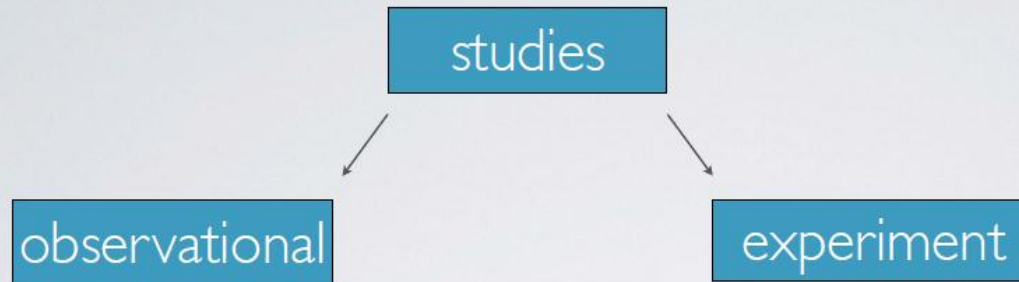
▸ collect data in a way that does not directly interfere with how the data arise ("observe")
▸ only establish an association
▸ retrospective: uses past data
▸ prospective: data are collected throughout the study

▸ randomly assign subjects to treatments
▸ establish causal connections

Observational studies & experiments
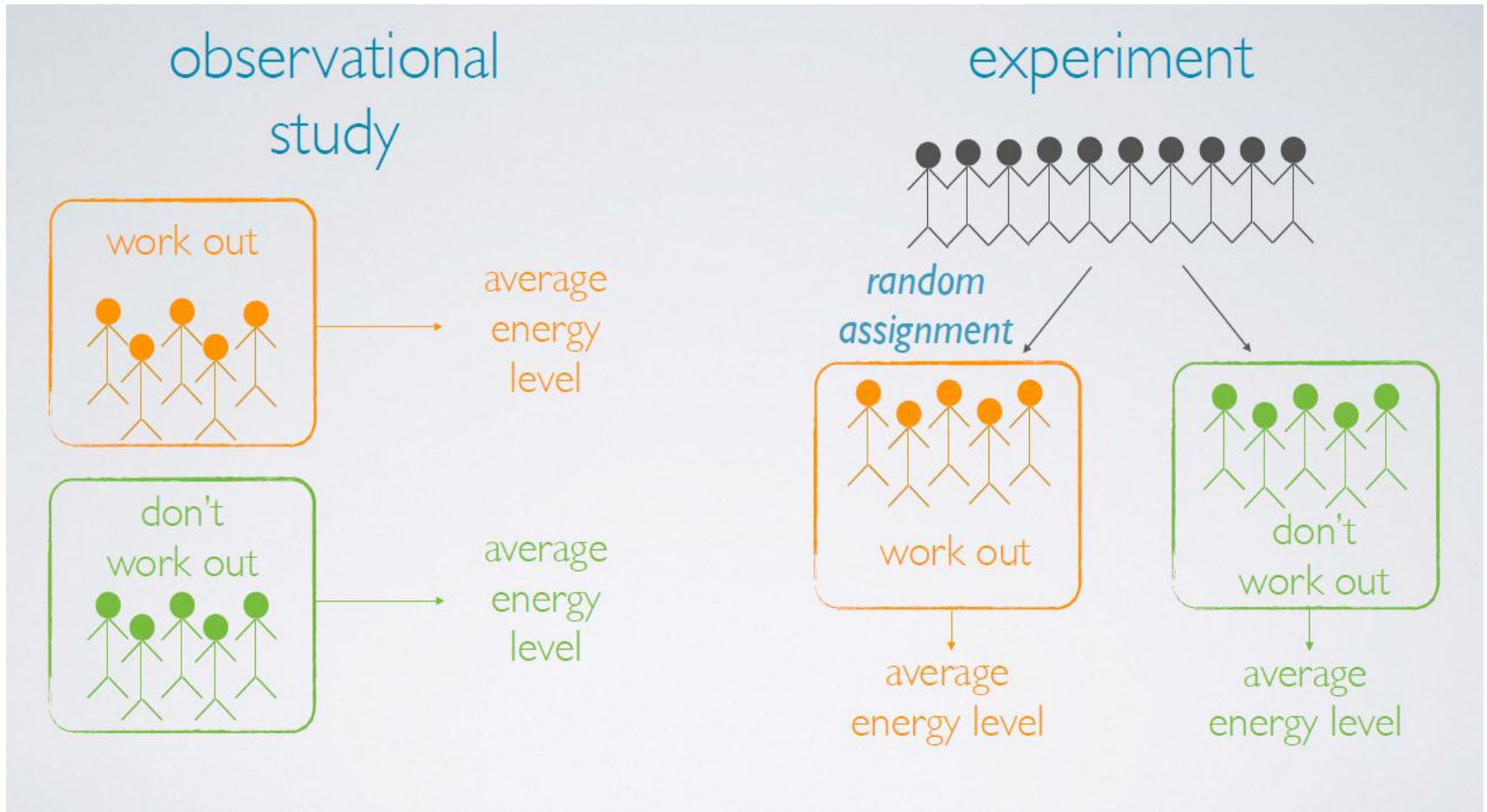
Correlation & Causation

- Case study

Study: Breakfast cereal keeps girls slim

USA TODAY
Sept 8, 2005

[...]

Girls who ate breakfast of any type had a lower average body mass index, a common obesity gauge, than those who said they didn't. The index was even lower for girls who said they ate cereal for breakfast, according to findings of the study conducted by the Maryland Medical Research Institute with funding from the National Institutes of Health (NIH) and cereal-maker General Mills.

[...]

The results were gleaned from a larger NIH survey of 2,379 girls in California, Ohio, and Maryland who were tracked between the ages of 9 and 19.

[...]

As part of the survey, the girls were asked once a year what they had eaten during the previous three days.

[...]

4/10/2023

Possible explanations

1. eating breakfast causes girls to be slimmer

2. being slim causes girls to eat breakfast

3. a third variable is responsible for both
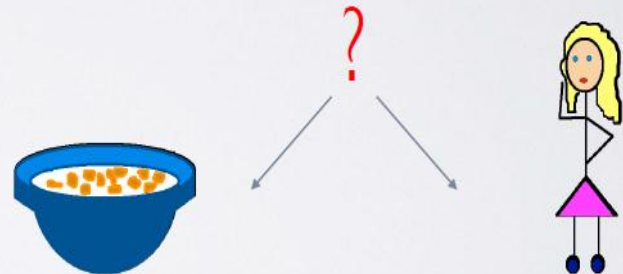
Confounding variables

extraneous variables that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them

confounding variable
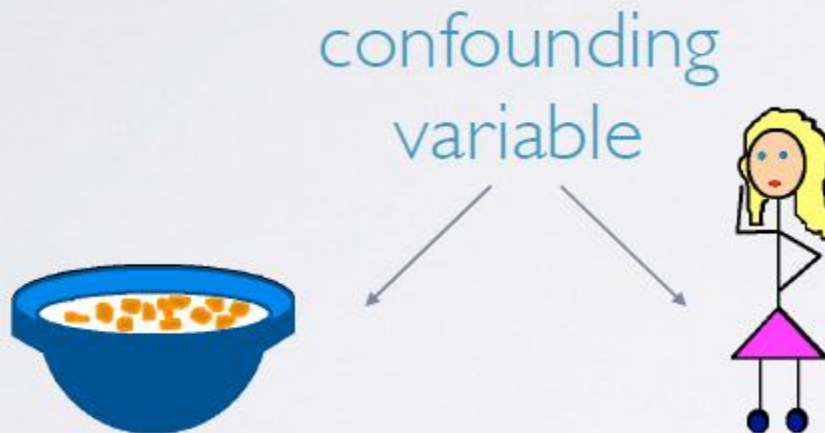
Correlation & Causation

☐ Correlation does not imply causation
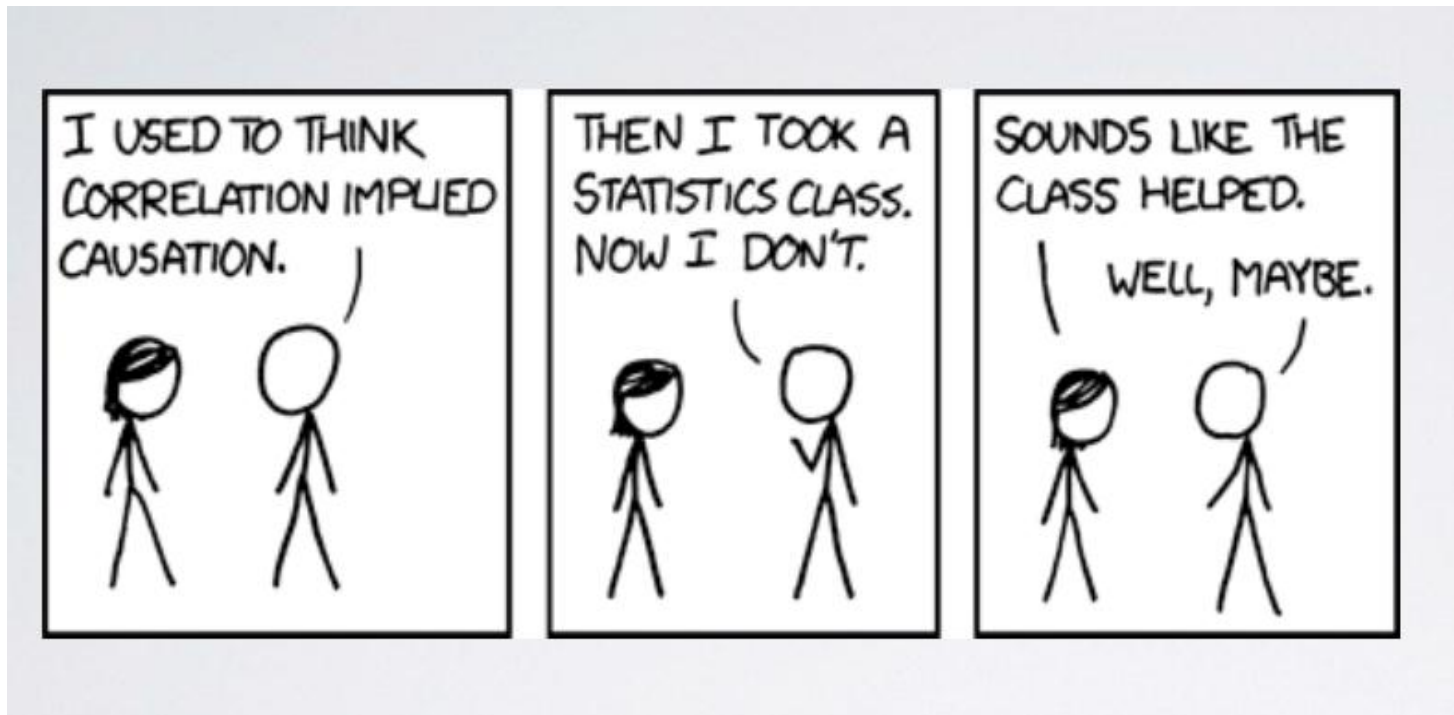
Sampling & sources bias

- Census vs sample

- Source o bias

- Sampling methods

4/10/2023

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 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.

4/10/2023

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

QUICK VOTE

Should the West intervene in Syria?

○ Yes   ○ No

VOTE   or view results

QUICK VOTE

Should the West intervene in Syria?
Yes 34% 534
No 66% 1038
Total Votes: 1572
This is not a scientific poll

A few sources of sampling bias

4/10/2023

Sampling methods

simple random sample (SRS)
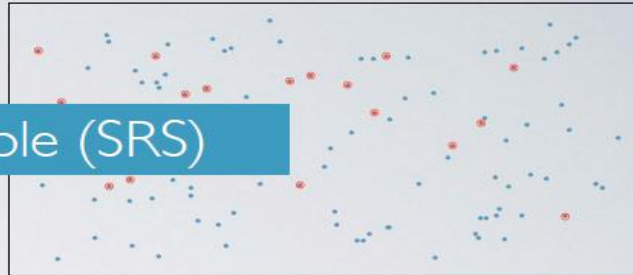
stratified sample
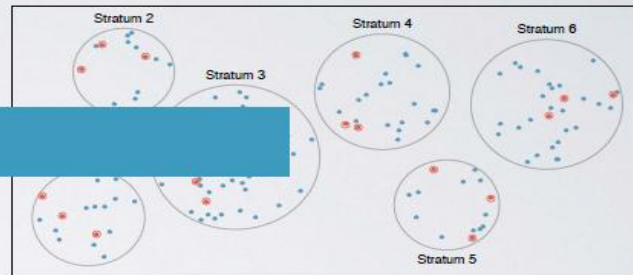
cluster sample
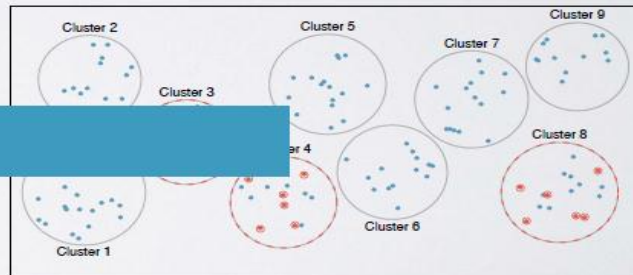
Sampling methods: random sampling
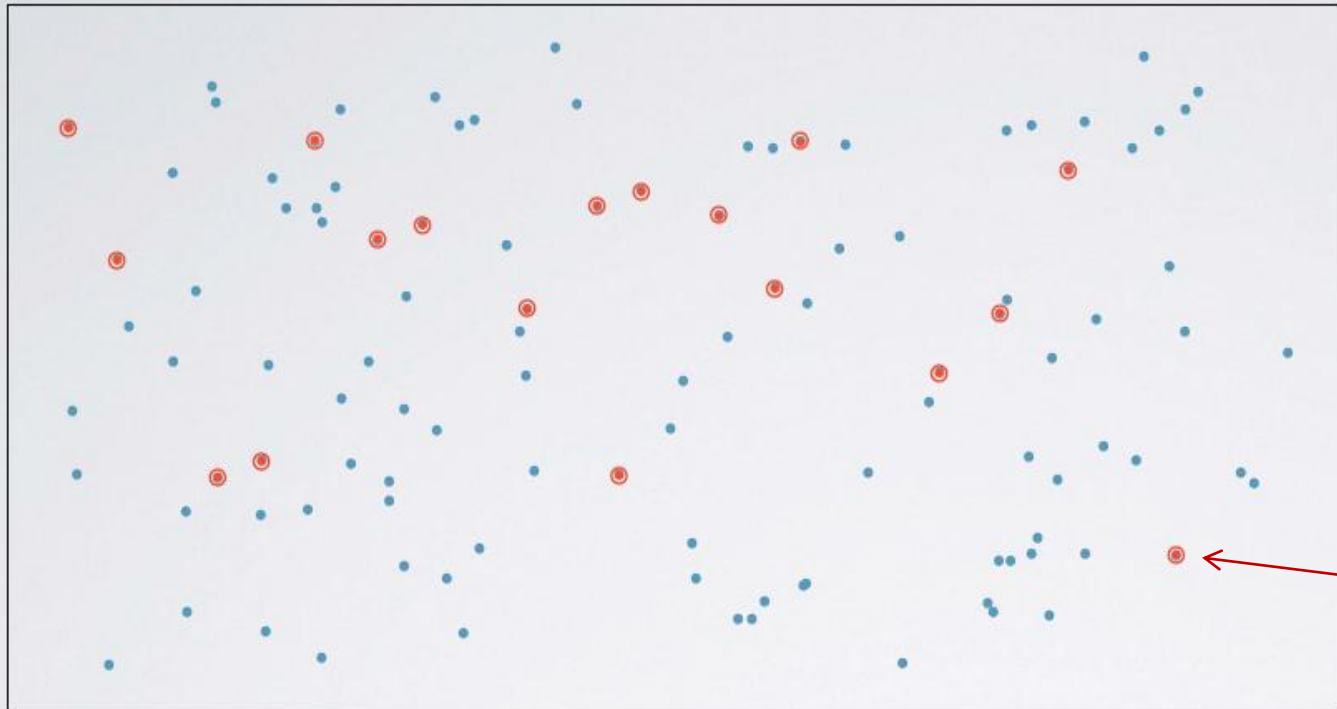
each case is equally likely to be selected

selected

Sampling methods: stratified sample

selected

divide the population into homogenous strata, then randomly sample from within each stratum
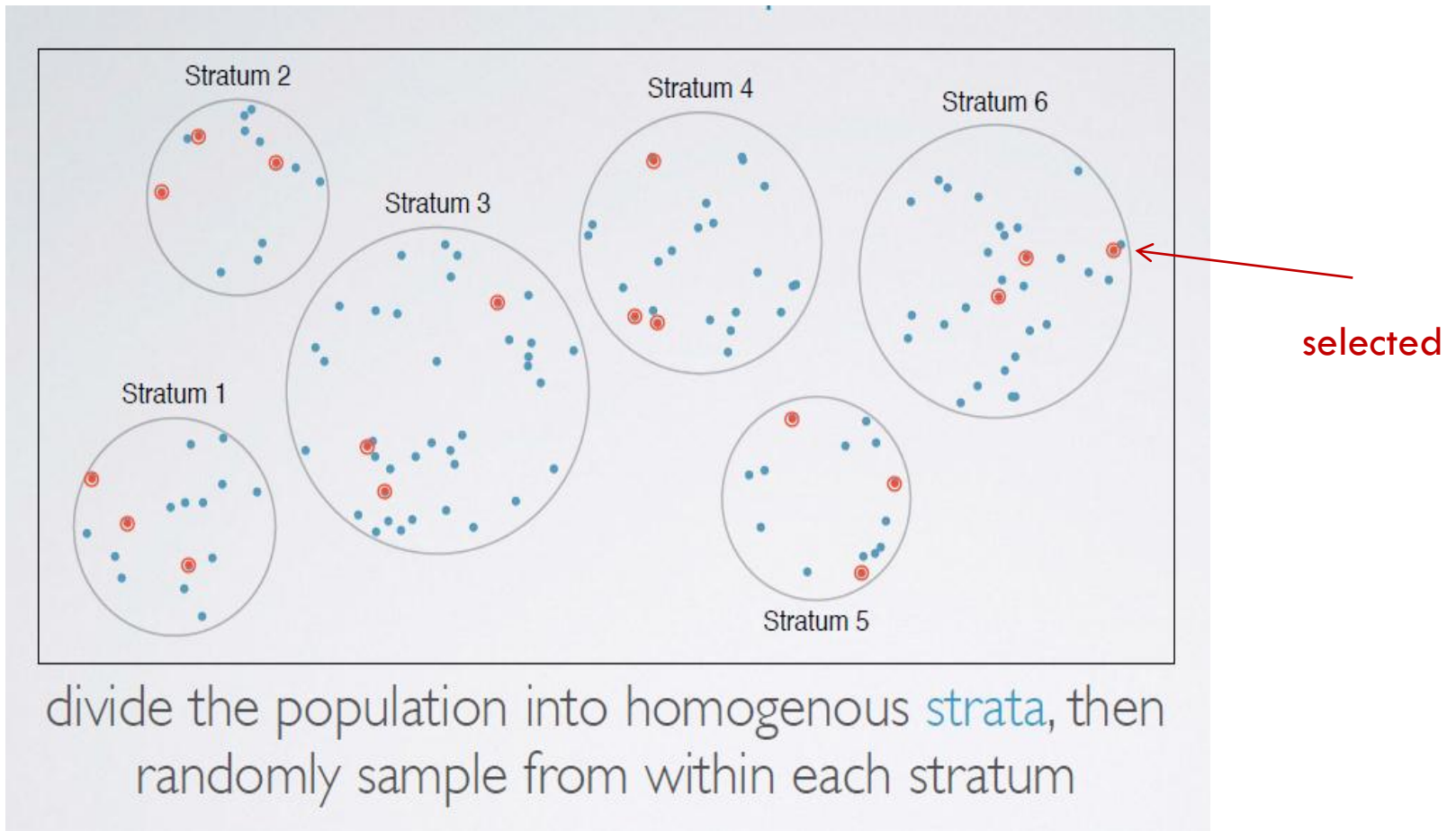
Sampling methods: cluster
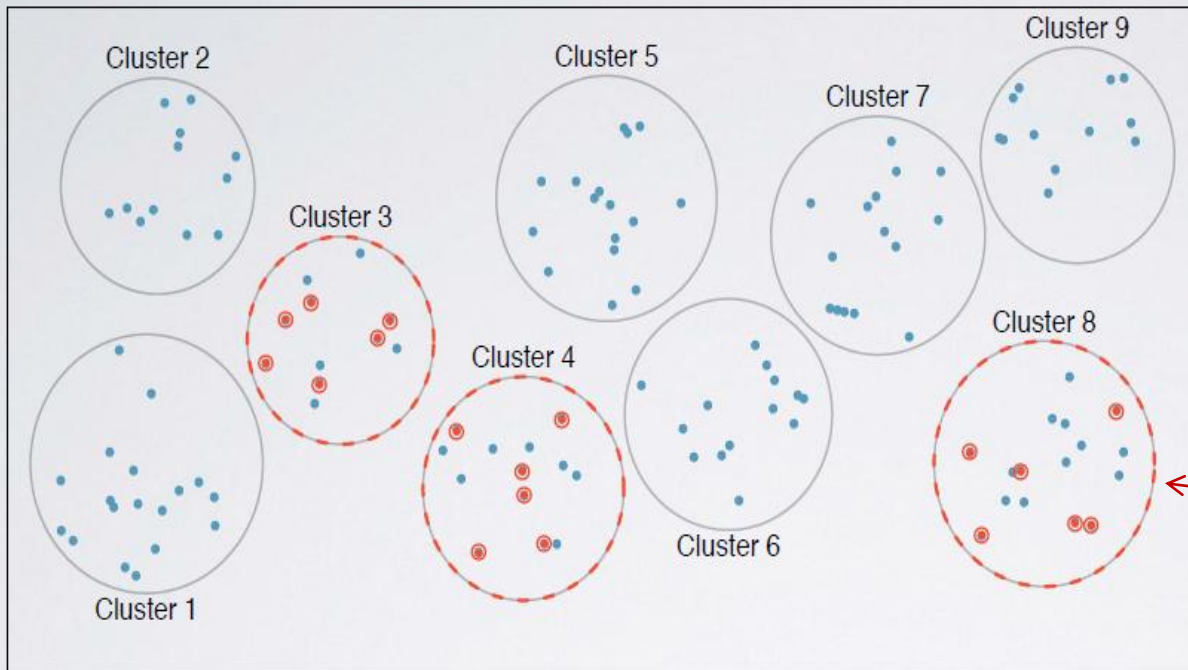
divide the population clusters, randomly sample a few clusters, then randomly sample from within these clusters

selected

4/10/2023

Vizualizing numerical data

- Scatter plots for paired data
- Other visualizations for describing distributions of numerical variables

4/10/2023

Data matrix

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

4/10/2023

Scatterplots

Evaluating their relationship

direction — positive, negative
shape — linear, curved
strength — strong, weak
outliers

4/10/2023
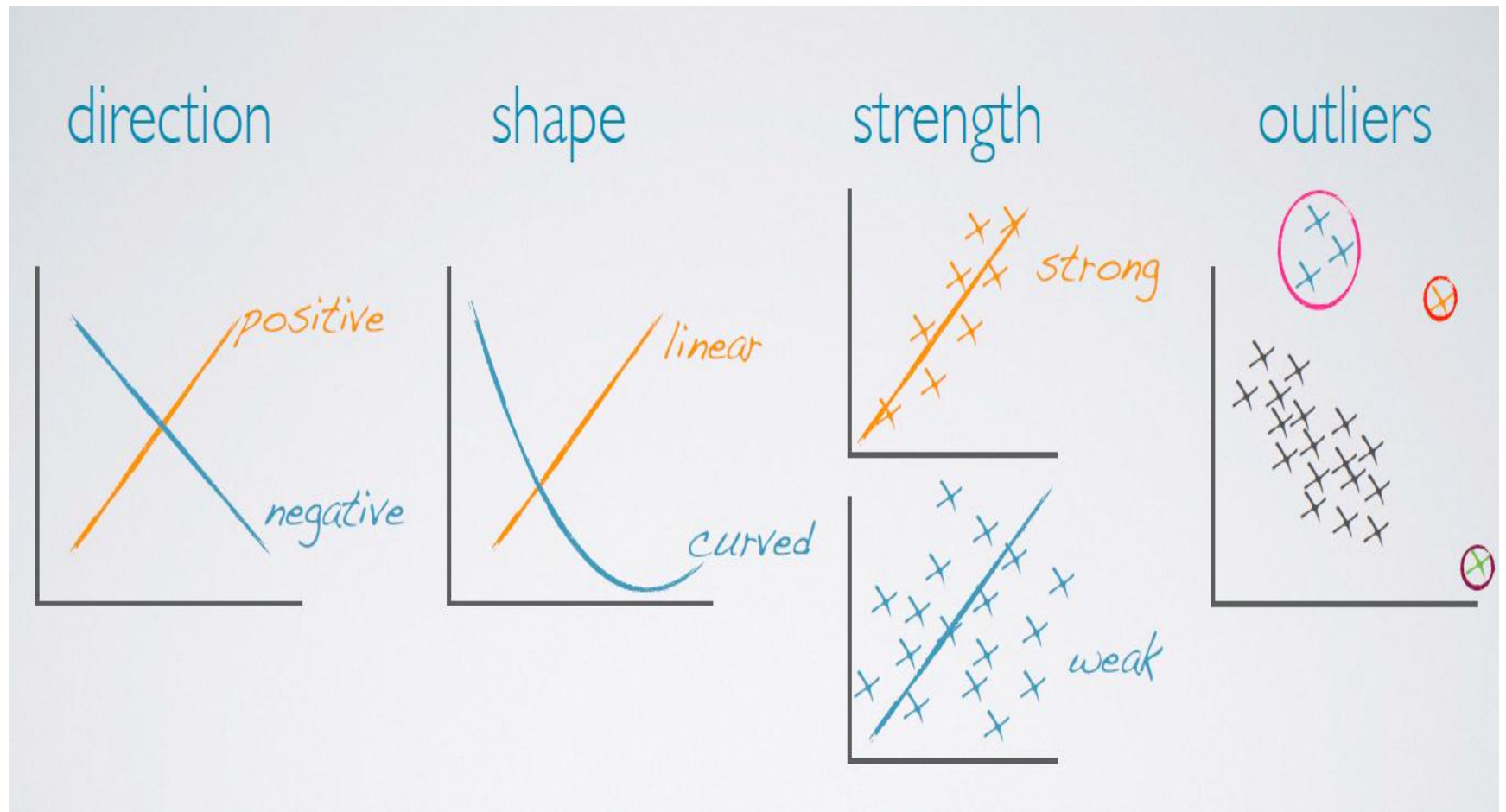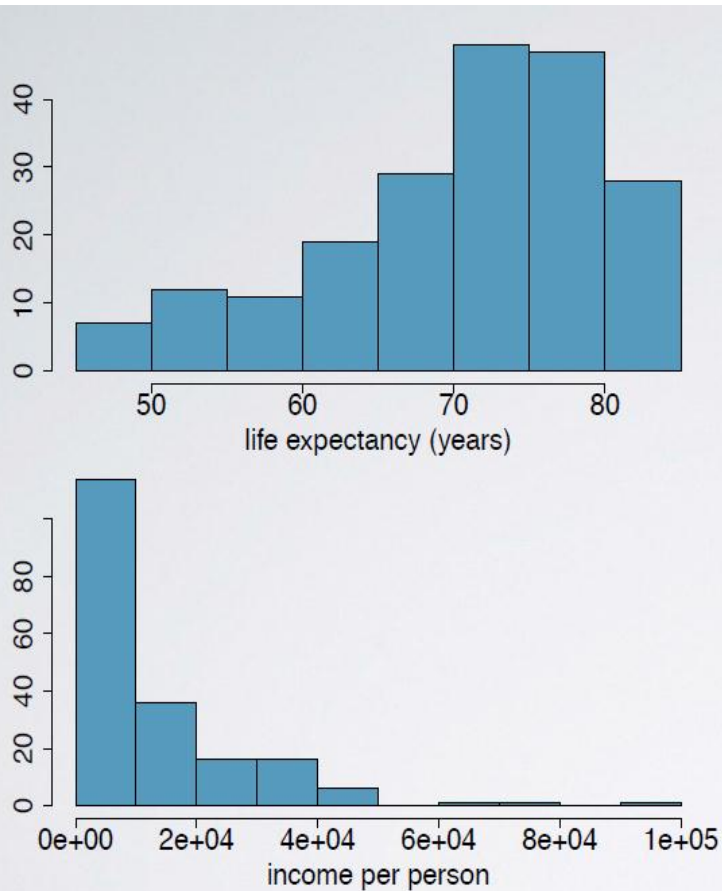
Histogram

histogram

▸ provides a view of the data density

▸ especially useful for describing the shape of the distribution

Histogram

distributions are skewed to the side of the long tail

left skewed

symmetric

right skewed
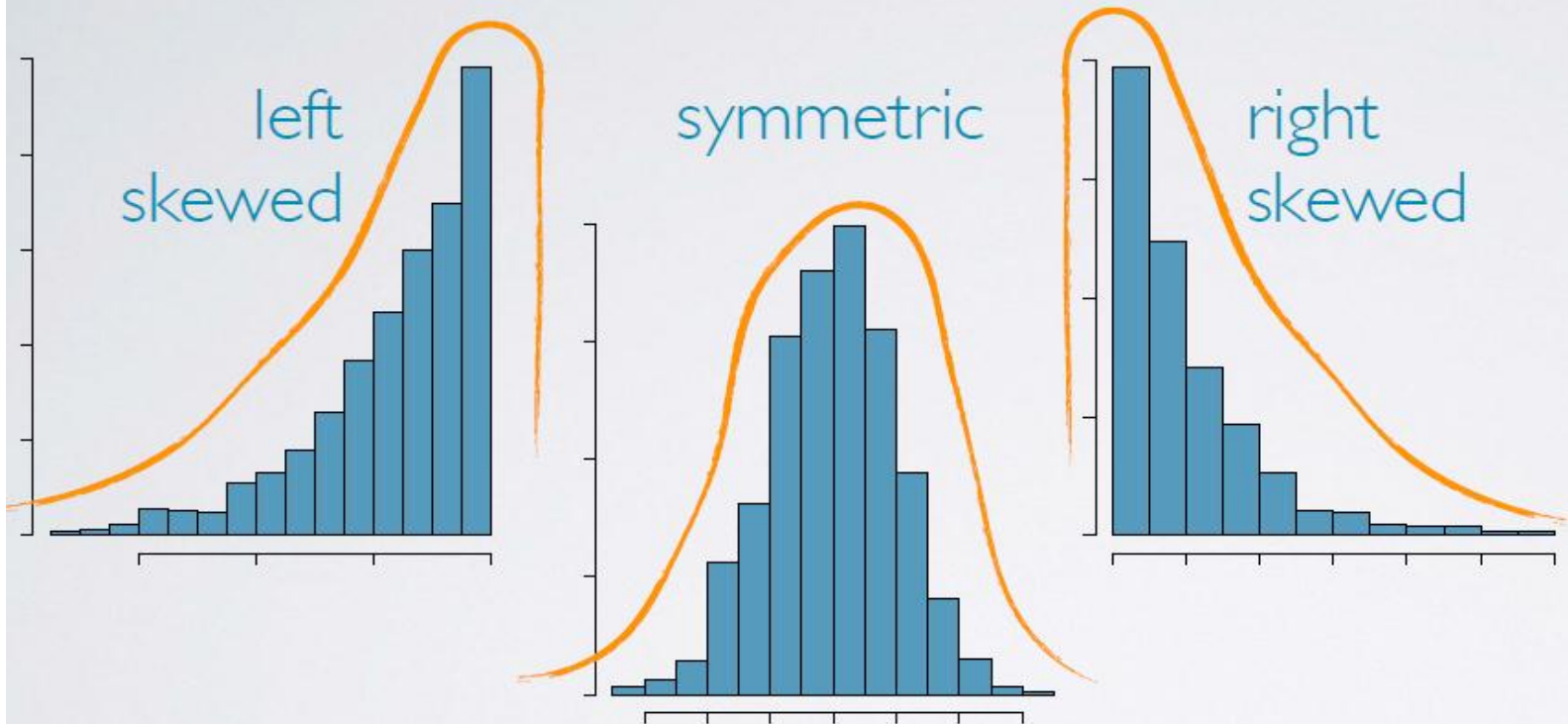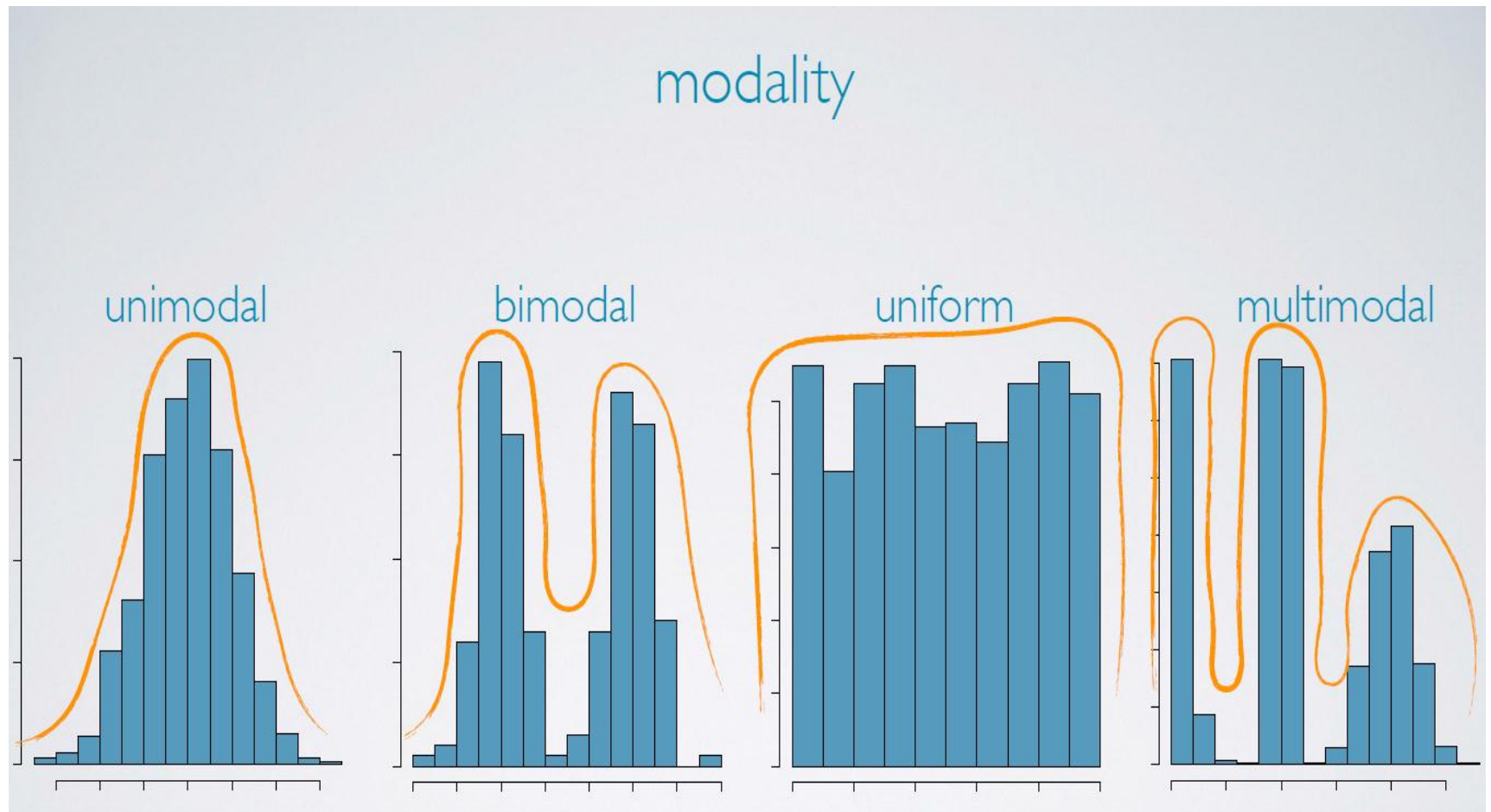
Histogram

Histogram

histogram & bin width

The chosen bin width can alter the story the histogram is telling.
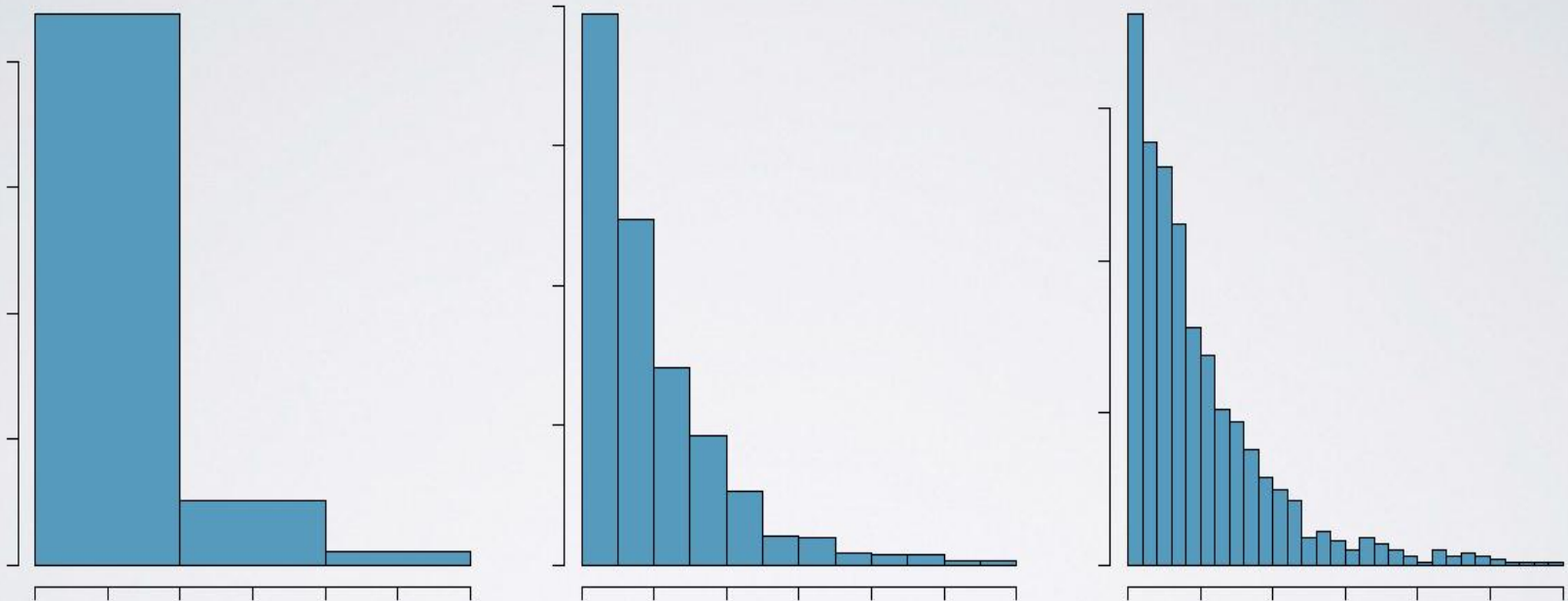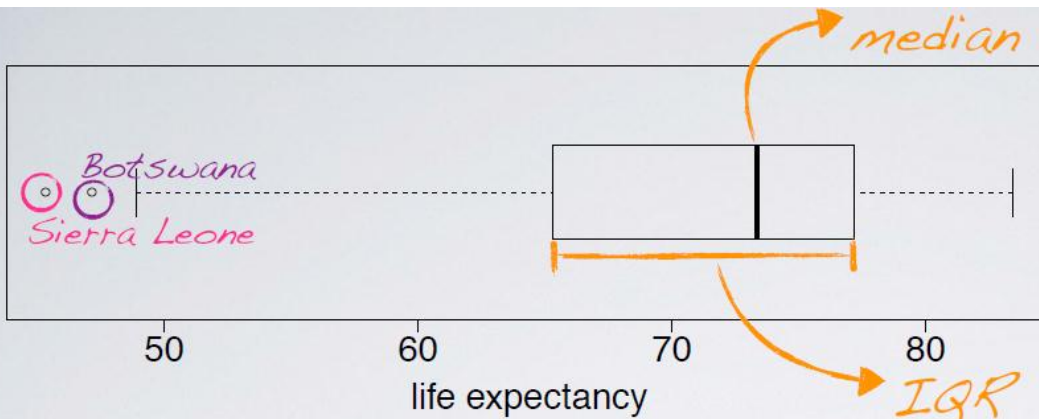
Box plot

box plot

useful for highlighting outliers, median, IQR

Box plot

4/10/2023
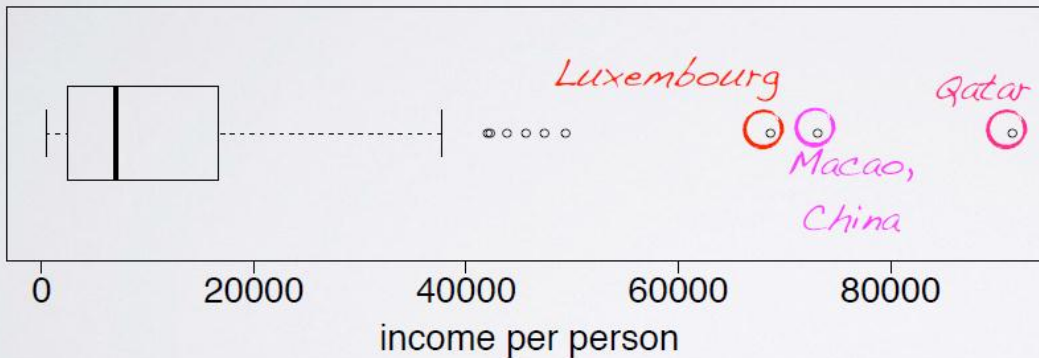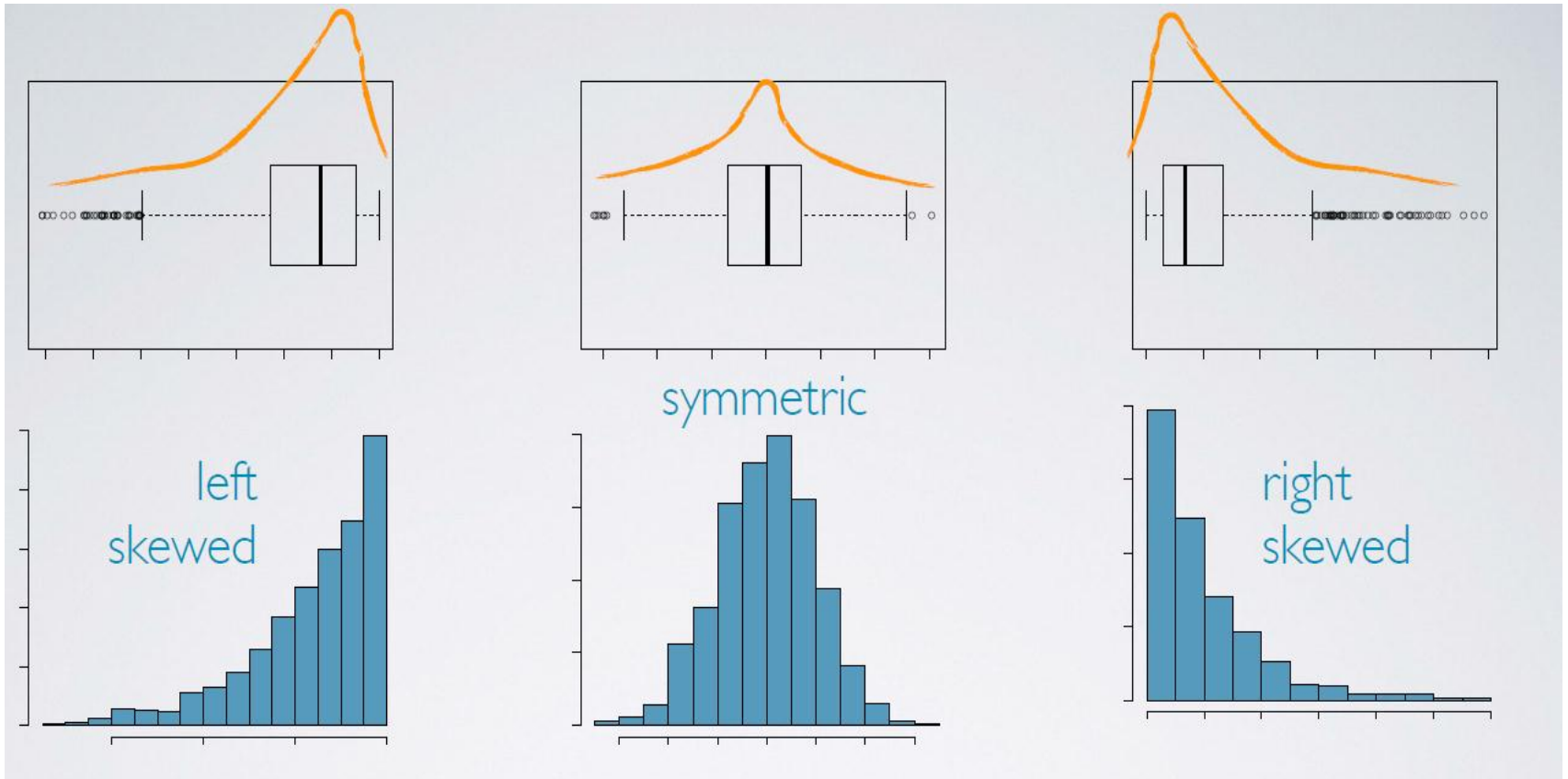
Intensity map

▸ Useful for highlighting the spatial distribution.

45.3   life expectancy   83.4

403   income per person   91490
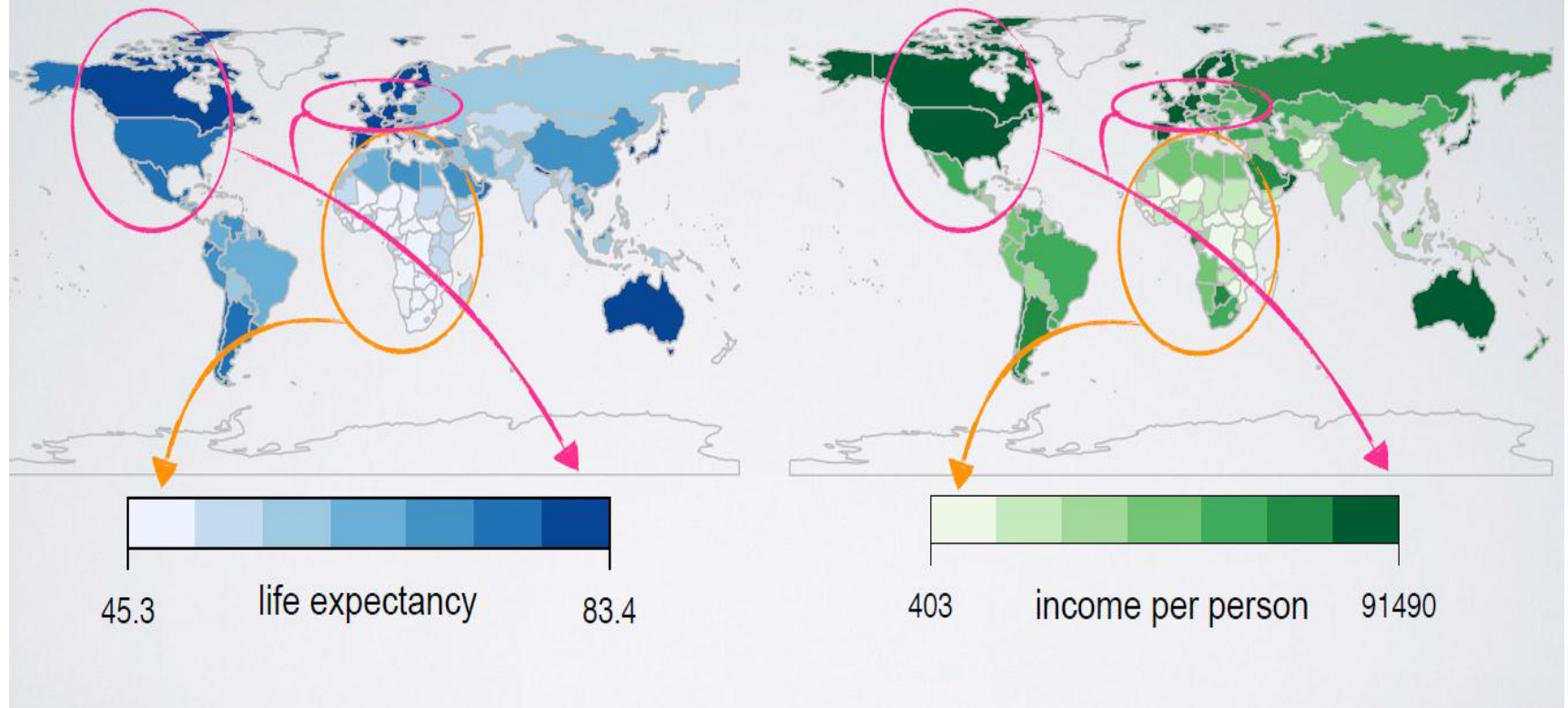
Measures of center
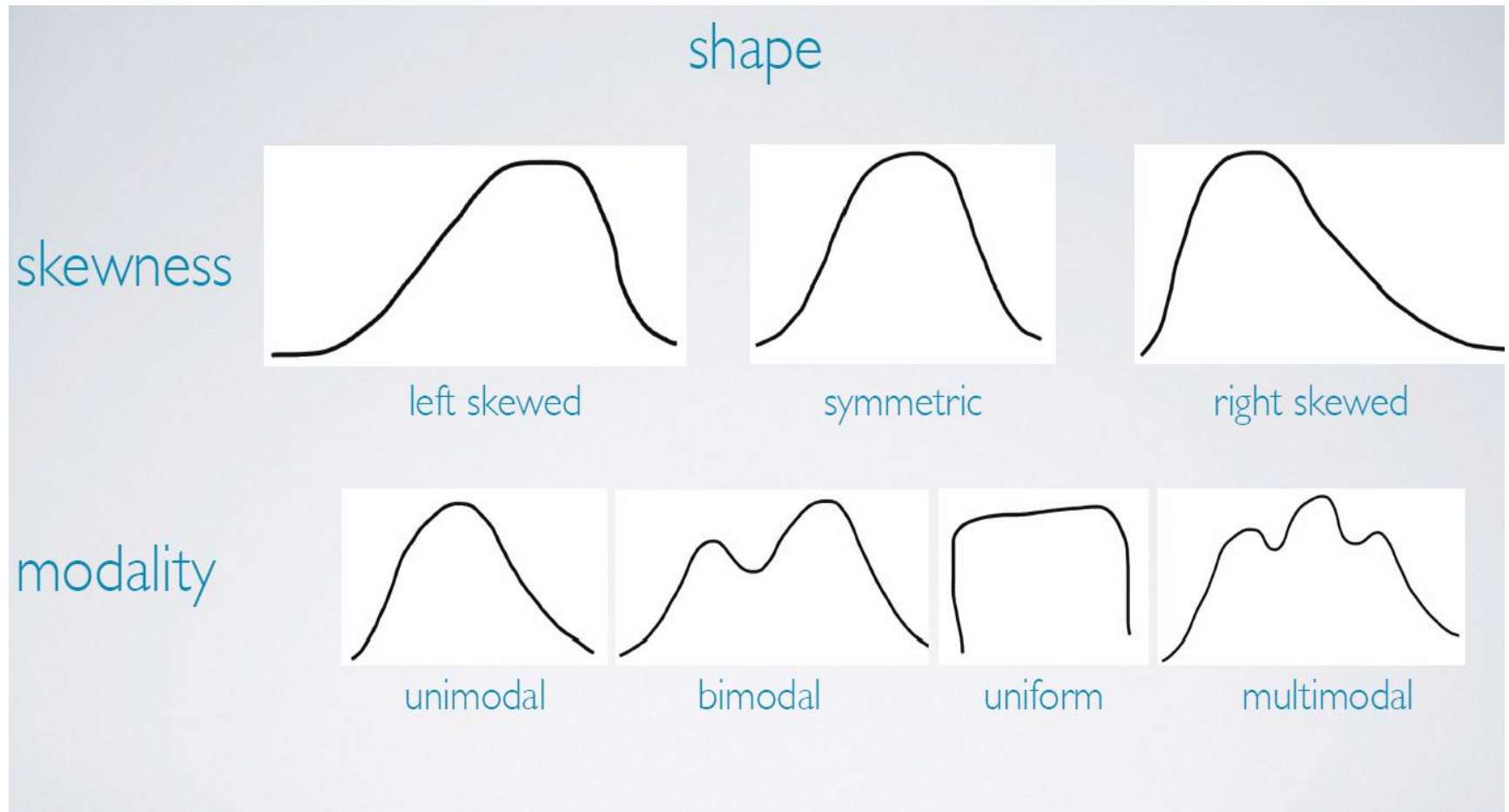
Measures of center

mean	median
arithmetic average	midpoint of the distribution (50th percentile)
$\bar{x}$ sample mean	
μ population mean	

mode	
most frequent observation	sample statistic
	point estimate
	population parameter

Measures of center

example   9 students' exam scores:

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

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

mode: 88

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

4/10/2023

Data matrix

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

Measures of center
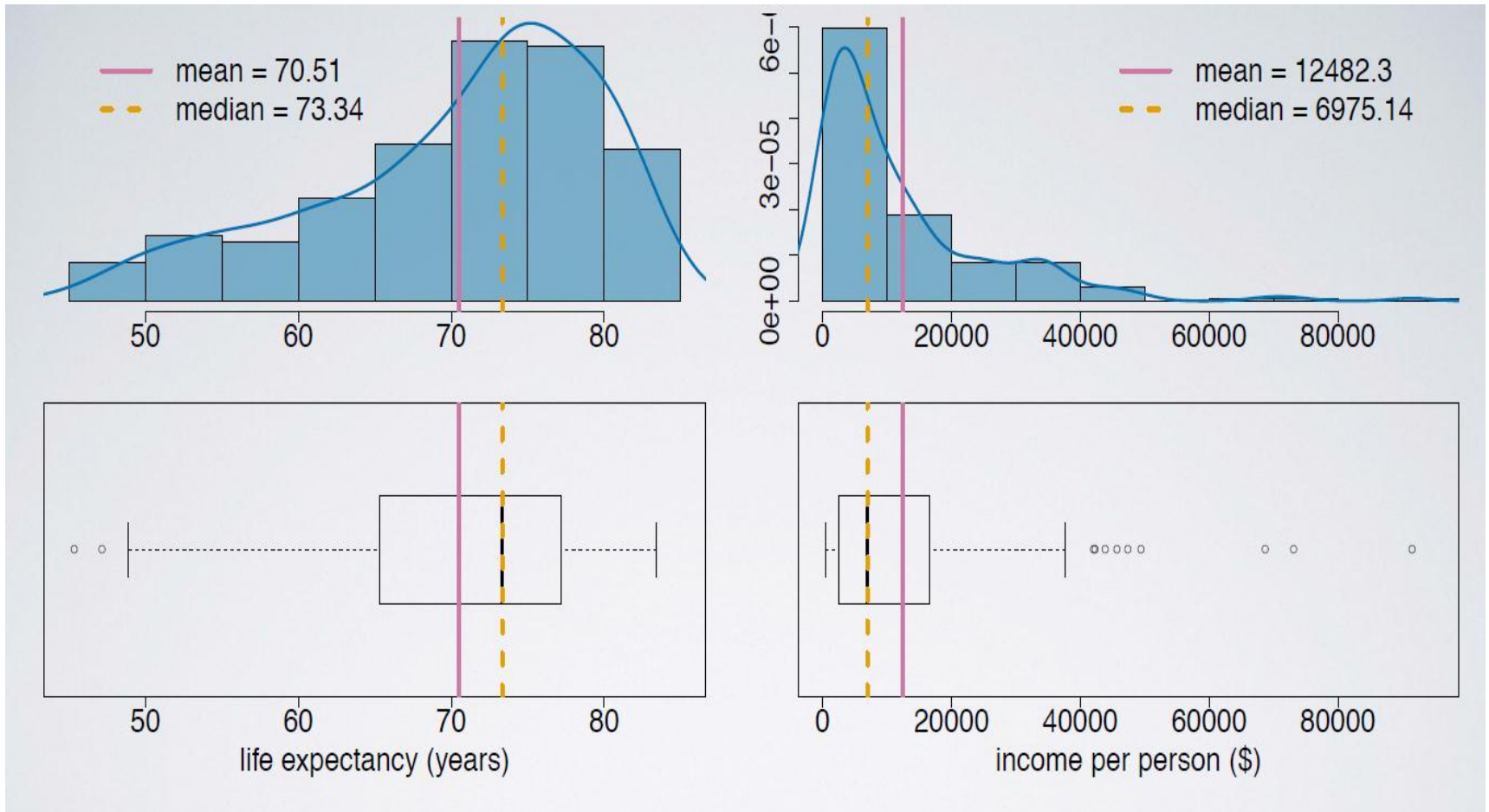
Skewness vs. measures of center
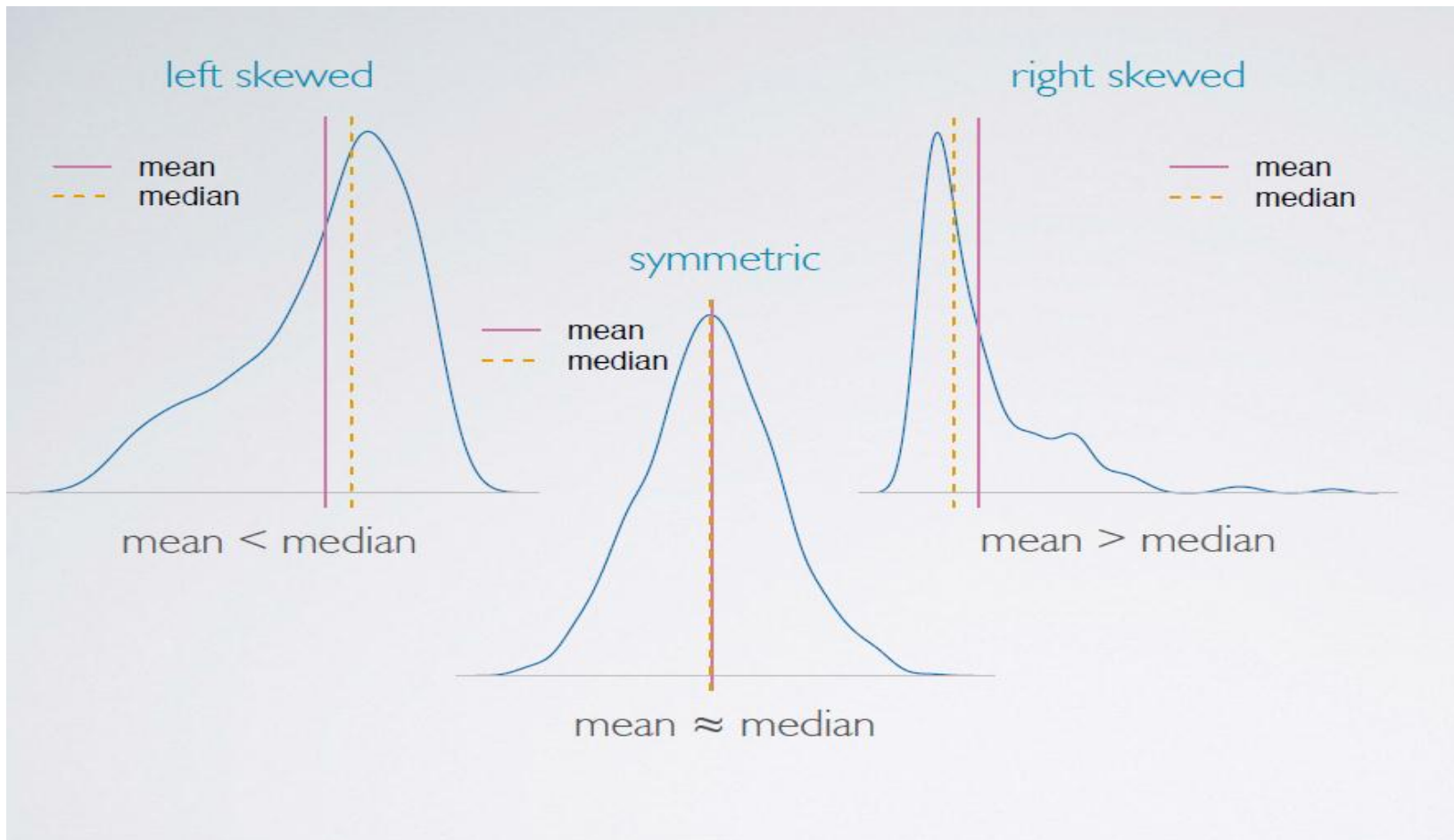
Measures of spread

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

—— less variable
- - - - more variable
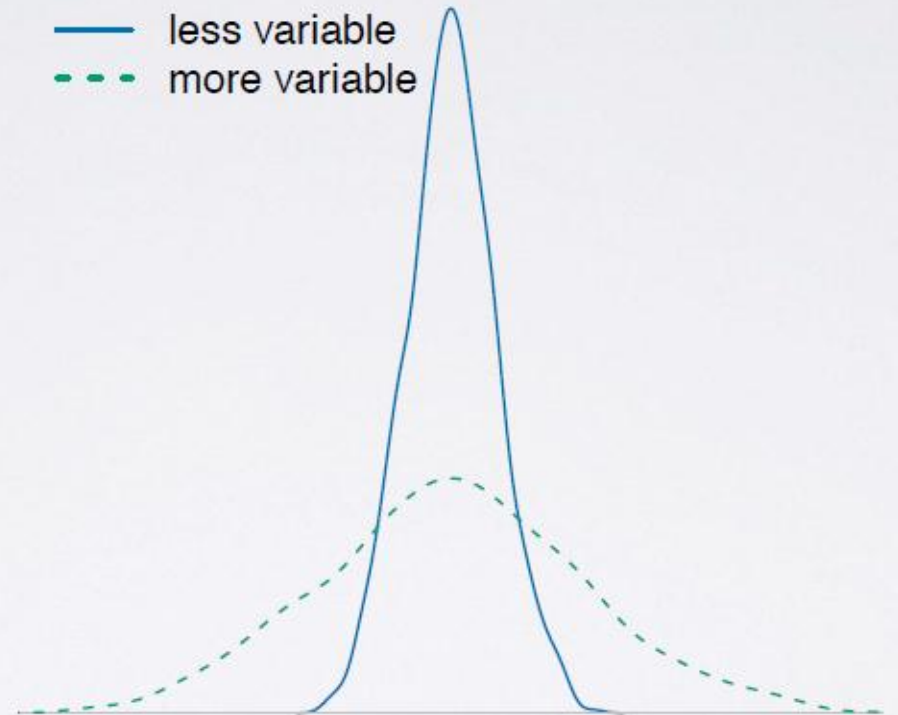
Measures of spread

sample variance s^2

population variance σ^2

> variance

roughly the average squared deviation from the mean

$$s^2 = \frac{\sum_1^{n=1}(x_i - \bar{x})^2}{n-1}$$

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

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

$= 83.06 \; years^2$

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

4/10/2023

Measures of spread

Why do we square the differences?

$$s^2 = \frac{\sum_1^{n=1}(x_i - \bar{x})^2}{n - 1}$$

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

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

$(-2) + 2 = 0$

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

4/10/2023
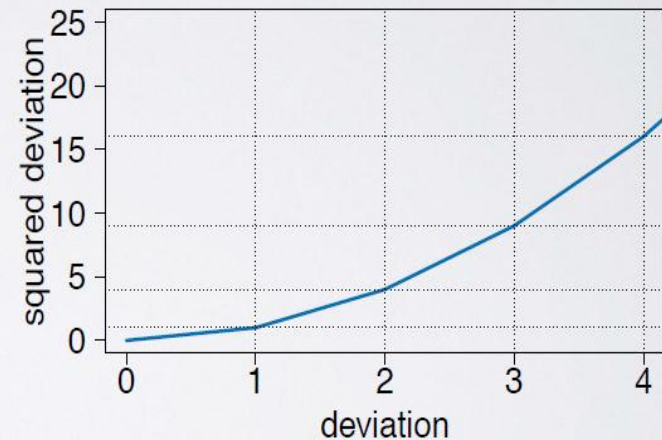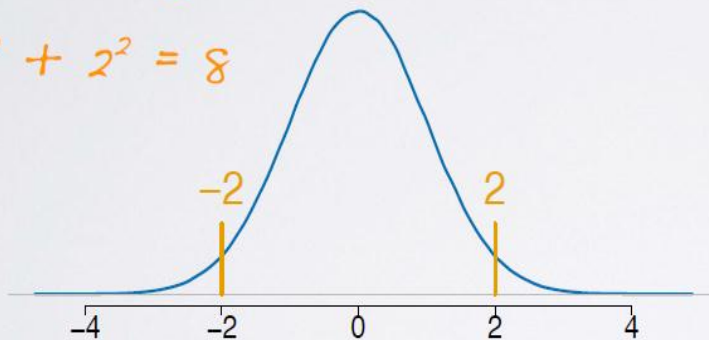
Measures of spread

standard deviation

sample sd
s

population sd
σ

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

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_1^{n=1}(x_i - \bar{x})^2}{n-1}}$$

square root of the variance

4/10/2023

Measures of spread

interquartile range

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

$$IQR = Q3 - Q1$$

$Q1 \approx 65$

$IQR = 77 - 65 = 12$

$Q3 \approx 77$

life expectancy

4/10/2023
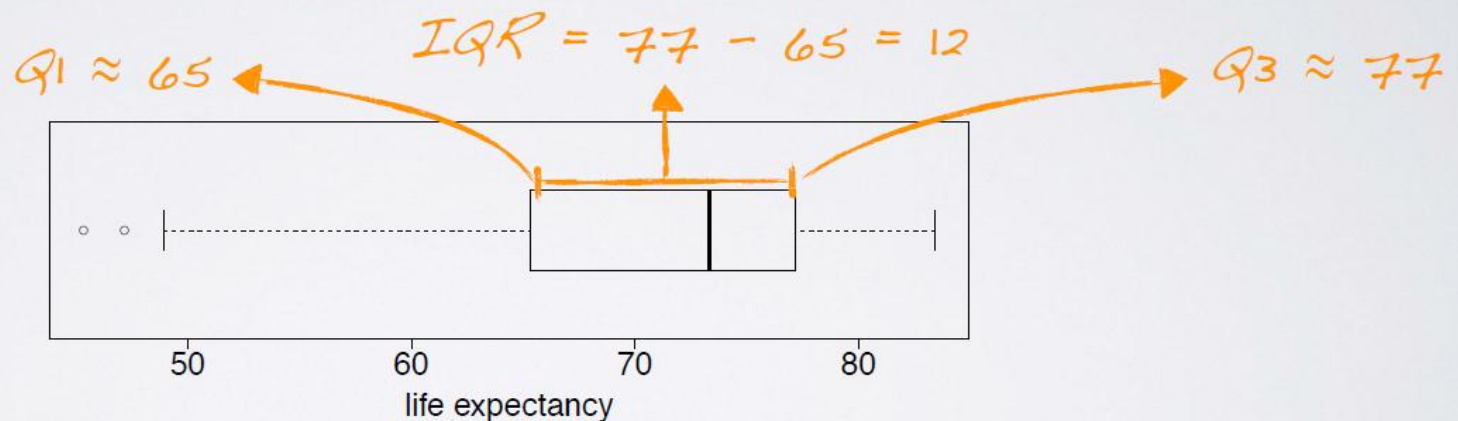
Robust statistics

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

4/10/2023

Robust statistics

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

skewed, with extreme observations

symmetric

Transforming data

▶ 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

goals of transformations

▶ to see the data structure differently

▶ to reduce skew assist in modeling

▶ to straighten a nonlinear relationship in a scatterplot

4/10/2023

Transforming data

(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
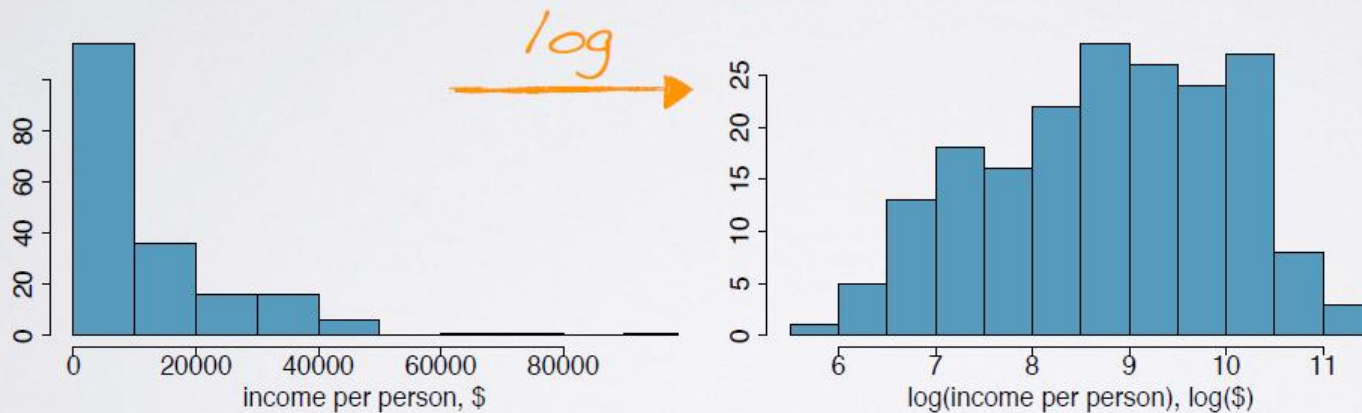
Transforming data

log transformation

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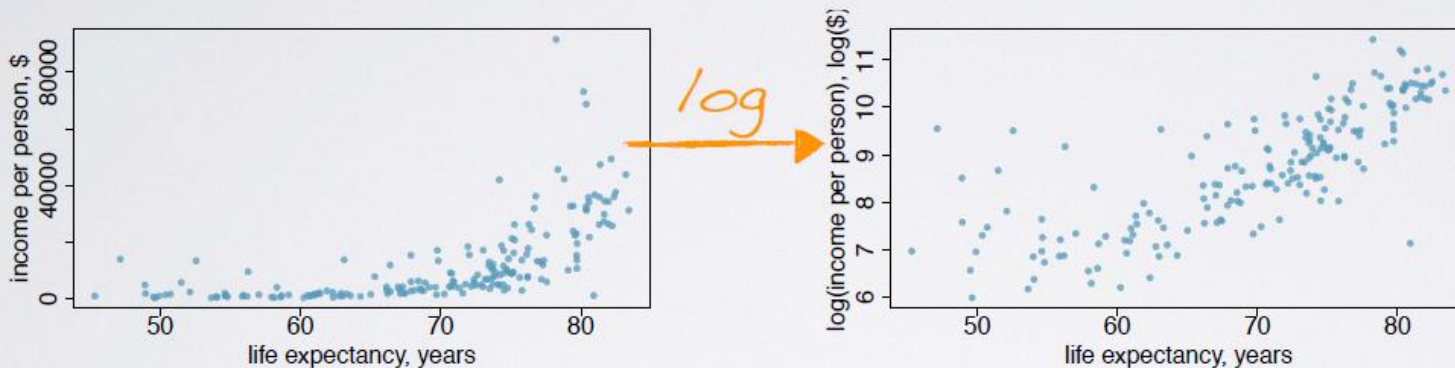
Transforming data
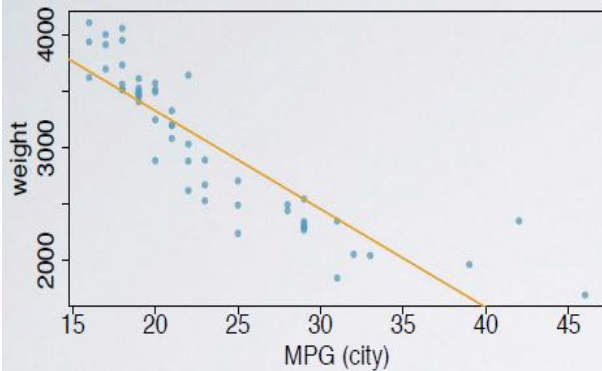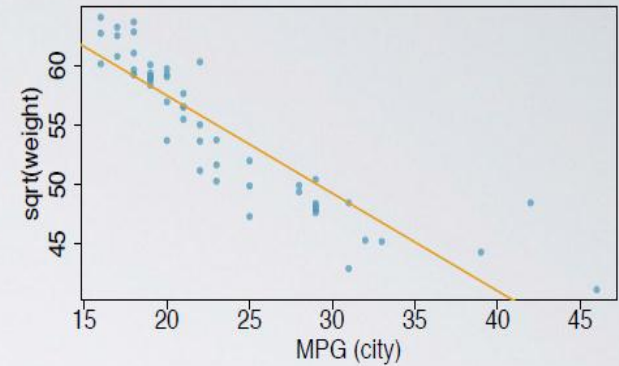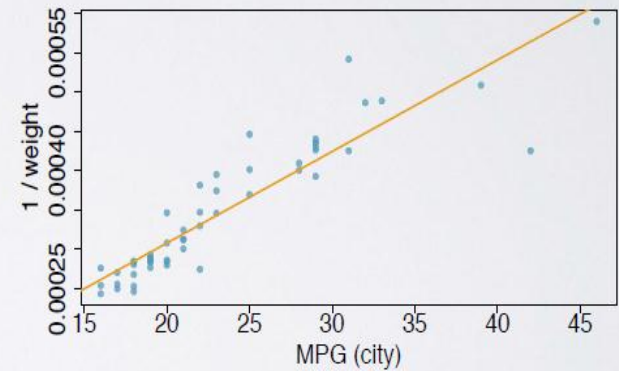
Exploring categorical variables

▶ describe distribution of a single categorical variable

▶ evaluate relationship between two categorical variables

▶ evaluate relationship between a categorical and a numerical variable

Exploring categorical variables

Difficulty saving money	Counts	Frequencies
Very	231	46%
Somewhat	196	39%
Not very	58	12%
Not at all	14	3%
Not sure	1	~0%
Total	500	100%

frequency table & bar plot
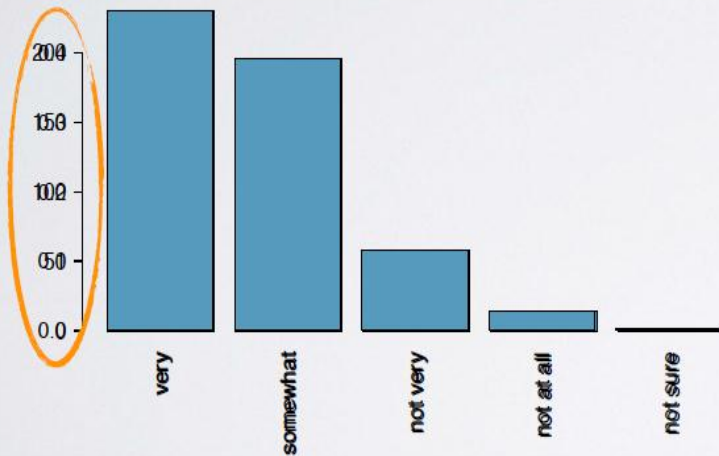
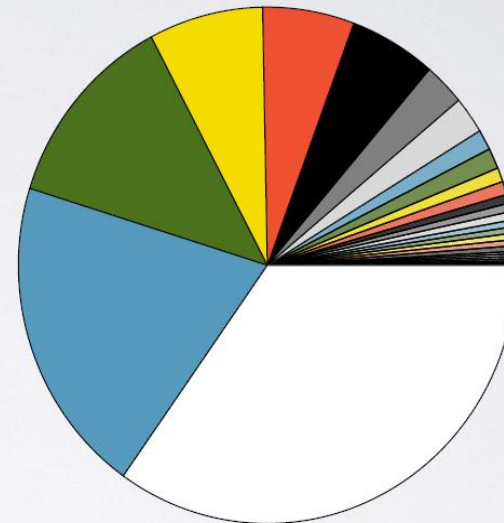# Exploring categorical variables
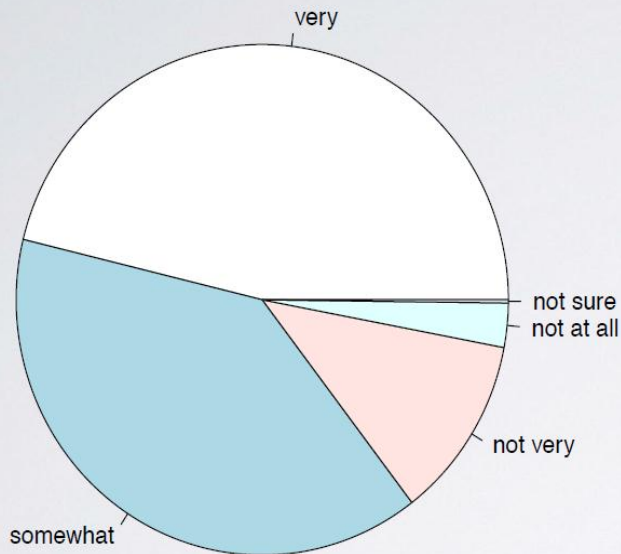
How are bar plots different than histograms?

▸ barplots for categorical variables, histograms for numerical variables

▸ x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable
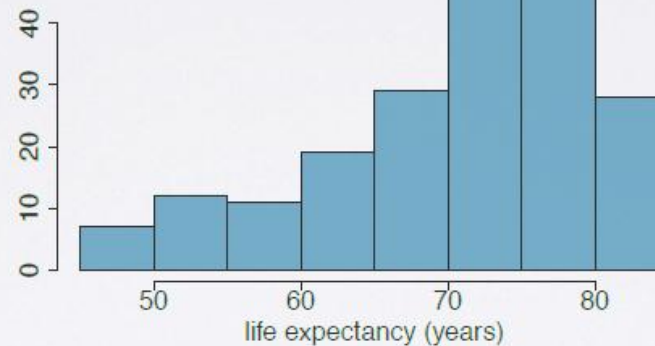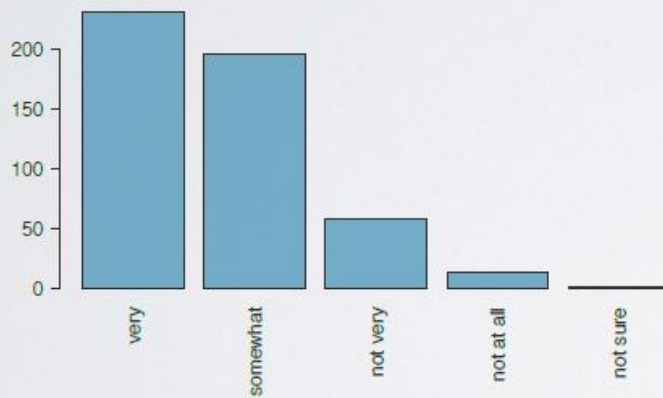
4/10/2023

Exploring categorical variables

Exploring categorical variables

contingency table

		Income				Total
		< $40K	$40-80K	> $80K	Refused	
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

4/10/2023

Exploring categorical variables

- ☐ Relative frequencies

		Income				
		< $40K	$40K - $80K	> $80K	Refused	Total
Difficulty saving	Very	128	63	31	9	231
	Somewhat	54	71	61	10	196
	Not very	17	7	27	7	58
	Not at all	3	6	5	0	14
	Not sure	0	1	0	0	1
	Total	202	148	124	26	500

<$40K: 128 / 202 = 63%   find it very difficult to save

$40K-$80K: 63 / 148 = 43%

>$80K: 31 / 124 = 25%

Refused: 9 / 26 = 35%

feelings about difficulty of saving money and income are associated (dependent)
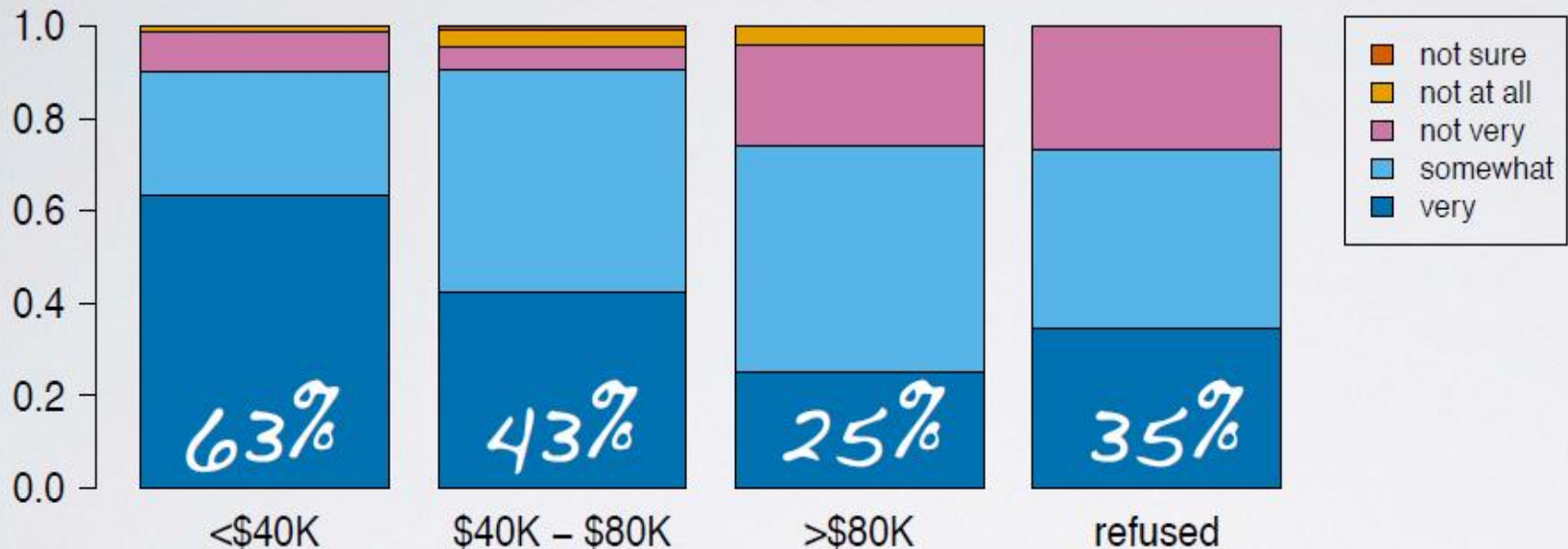
Exploring categorical variables

Exploring categorical variables

Exploring categorical variables

Exploring categorical variables