

INTRODUCTION TO DATA SCIENCE

Lecture based on:

M. Cetinkays-Rundel, „Data Analysis and Statistical Inference”, Univ. of Duke

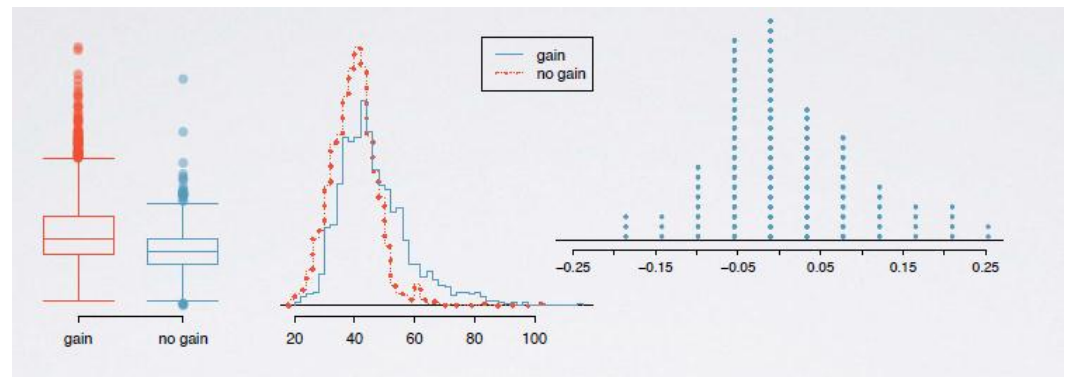
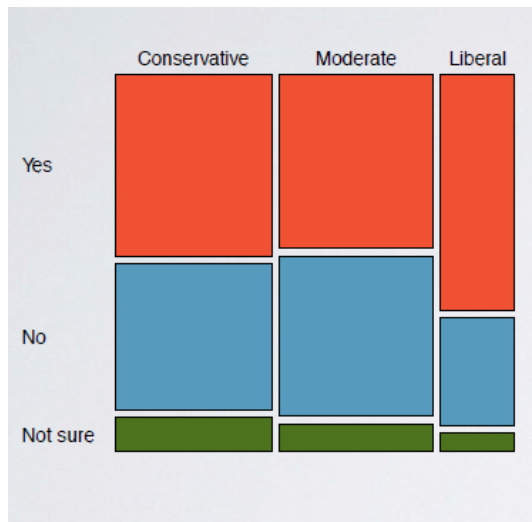
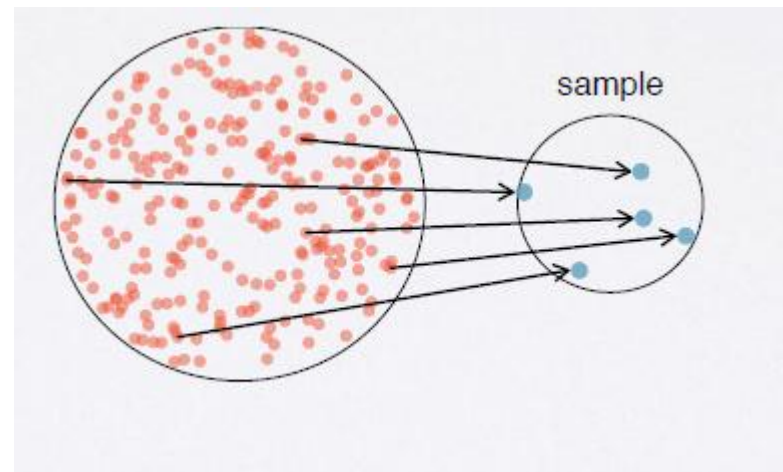
2/10/2018

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Exploratory data analysis

2

How to collect, visualise and interpret the data.



early
smoking
research

anecdotal
evidence

*“My uncle smokes
three packs a day
and he’s in perfectly
good health”*

*“smoking is a complex
human behavior, by its
nature difficult to study,
confounded by human
variability”*

Source: Brandt, The Cigarette Century (2009)

populations
and
samples

research
question

population

sample

generalize
to

Are consumers
of certain
alcohol brands
more likely to
end up in the
emergency
room with
injuries?

Everyone

ER patients
at the Johns
Hopkins
Hospital
in Baltimore
in the US

Residents
of
Baltimore

Source: Jernigan, David H.,
et al. "Alcohol brand use
and injury in the
emergency department: A
pilot study." Substance
use & misuse (2013).

Image credit: Jon Sullivan

<http://www.public-domain-image.com/food-and-drink-public-domain-images-pictures/wine-public-domain-images-pictures/champagne-flutes-glasses-bubbles.jpg.html>

Exploratory data analysis

5

population

sample

design

scope

exploratory
data
analysis

inference

Data: basics

6

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

Example: data matrix

7

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

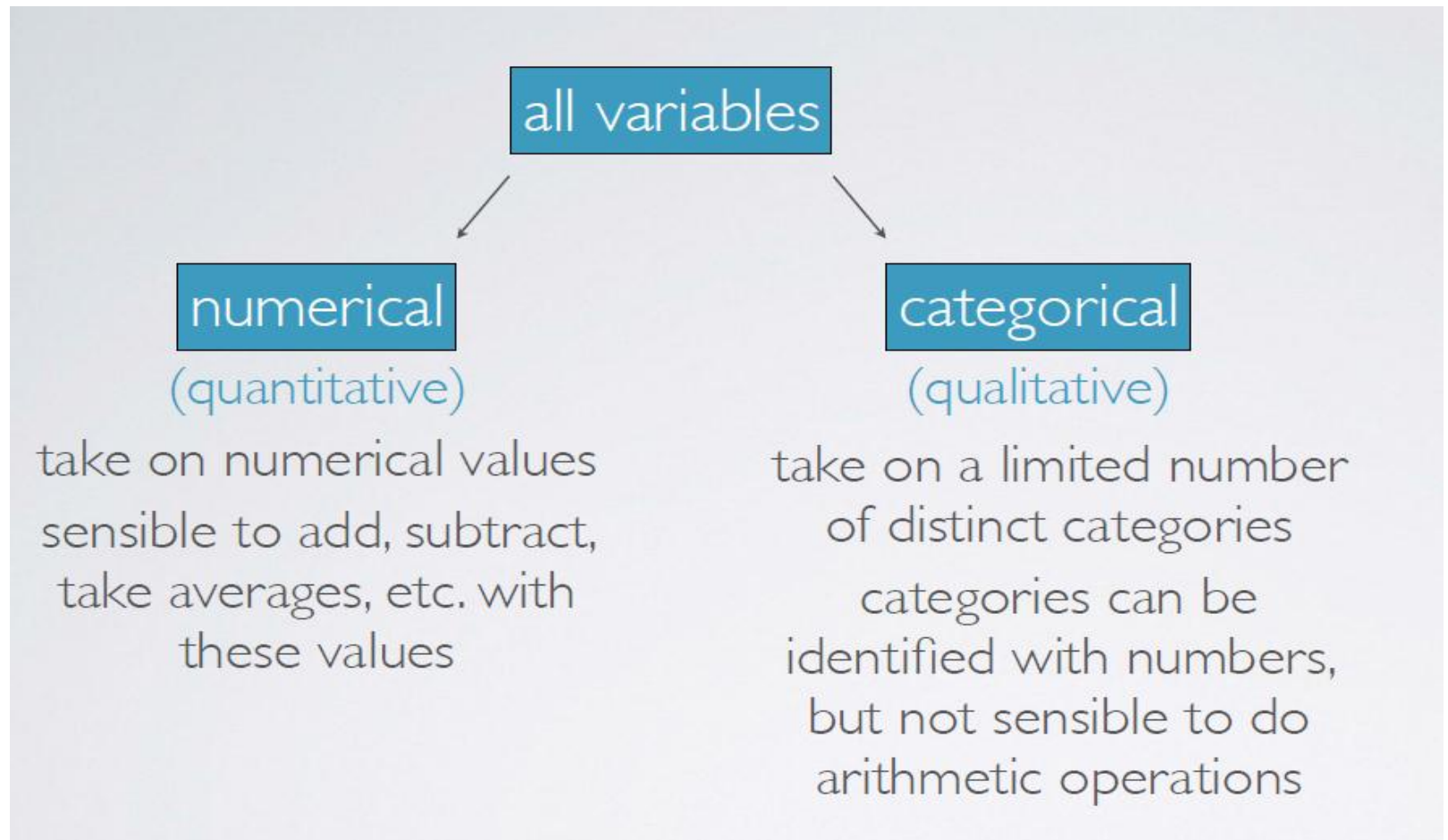
country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...
United States	92	63	5950	93	...	northern	very high

observation
(case)

variable

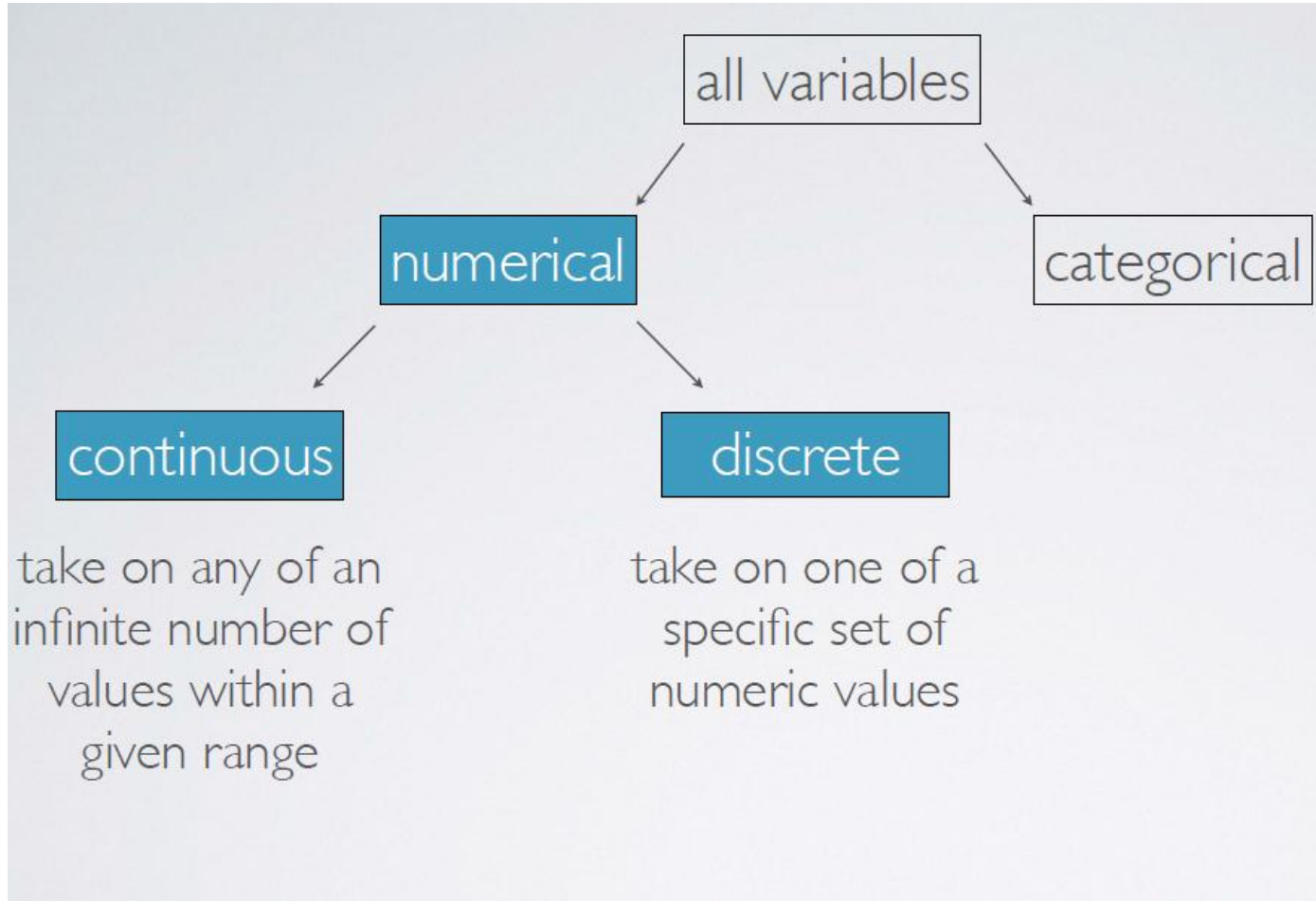
Type of variables

8



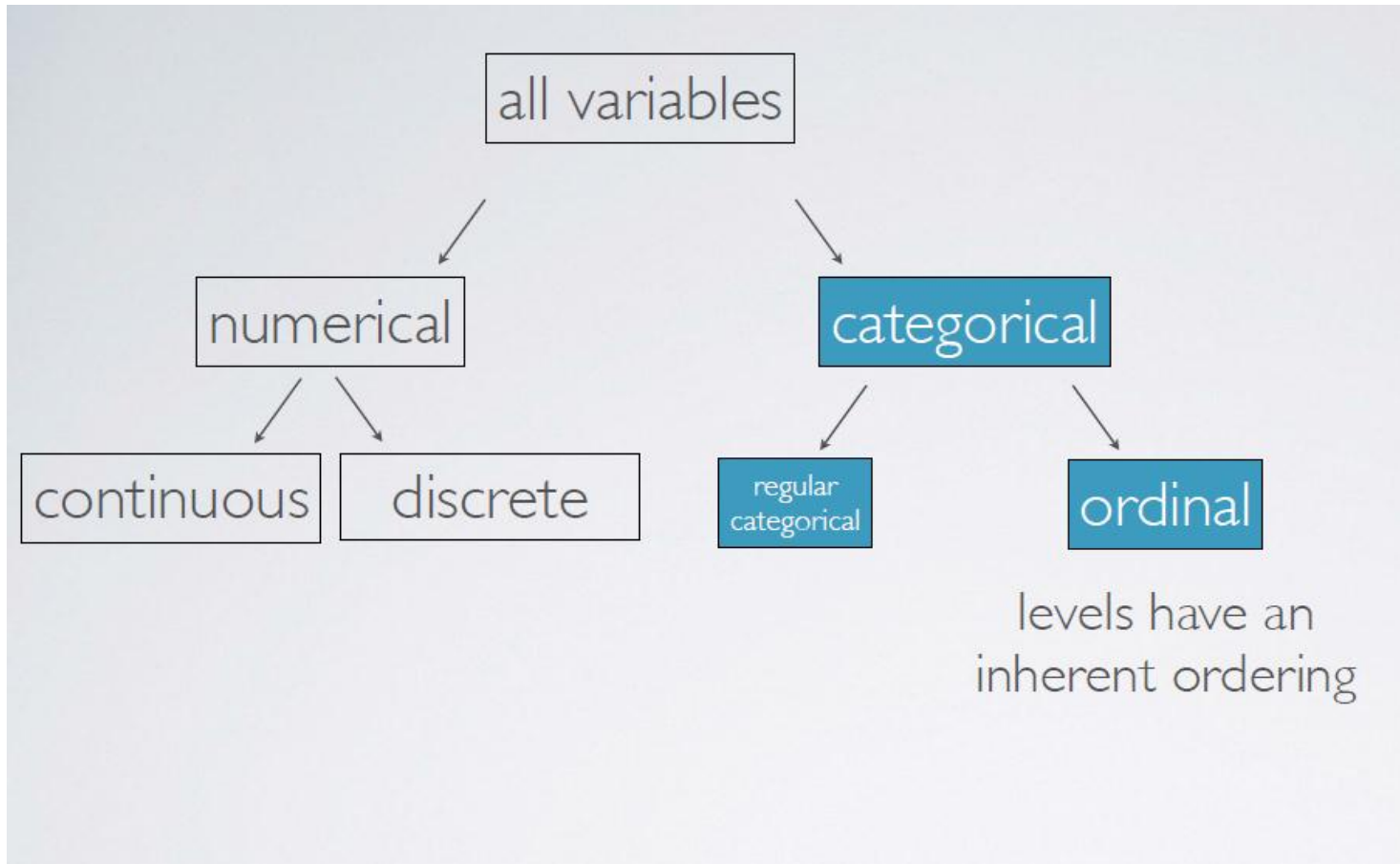
Numerical variables

9



Categorical variables

10



Data matrix

11

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

Data matrix

12

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

13

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

Data matrix

14

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

Data matrix

15

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

Data matrix

16

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

17

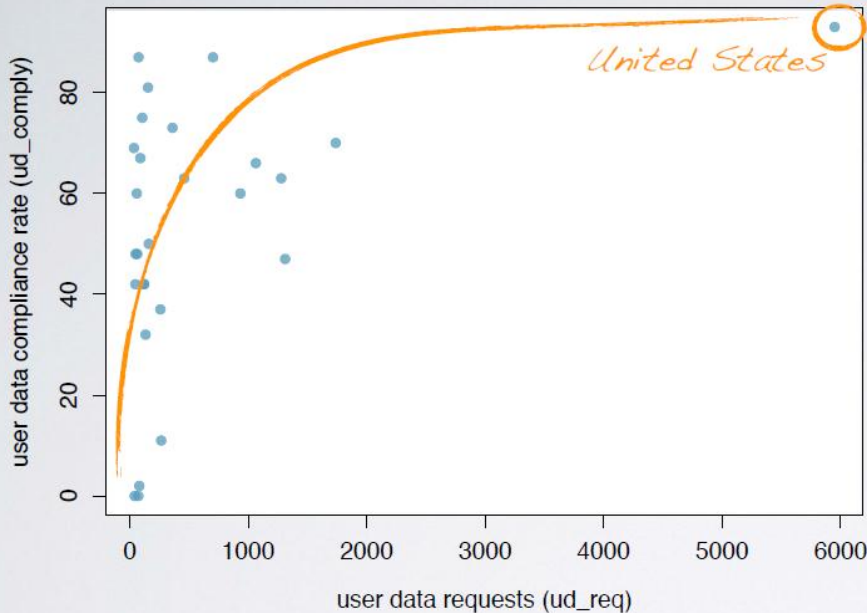
country	cr_req	cr_comply	ud_req	ud_comply	...	hemisphere	hdi
Argentina	21	100	134	32	...	southern	very high
Australia	10	40	361	73	...	southern	very high
Belgium	<10	100	90	67	...	northern	very high
Brazil	224	67	703	82	...	southern	high
...
United States	92	63	5950	93	...	northern	very high

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

Relationships between variables

18

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**

Observational studies & experiments

19

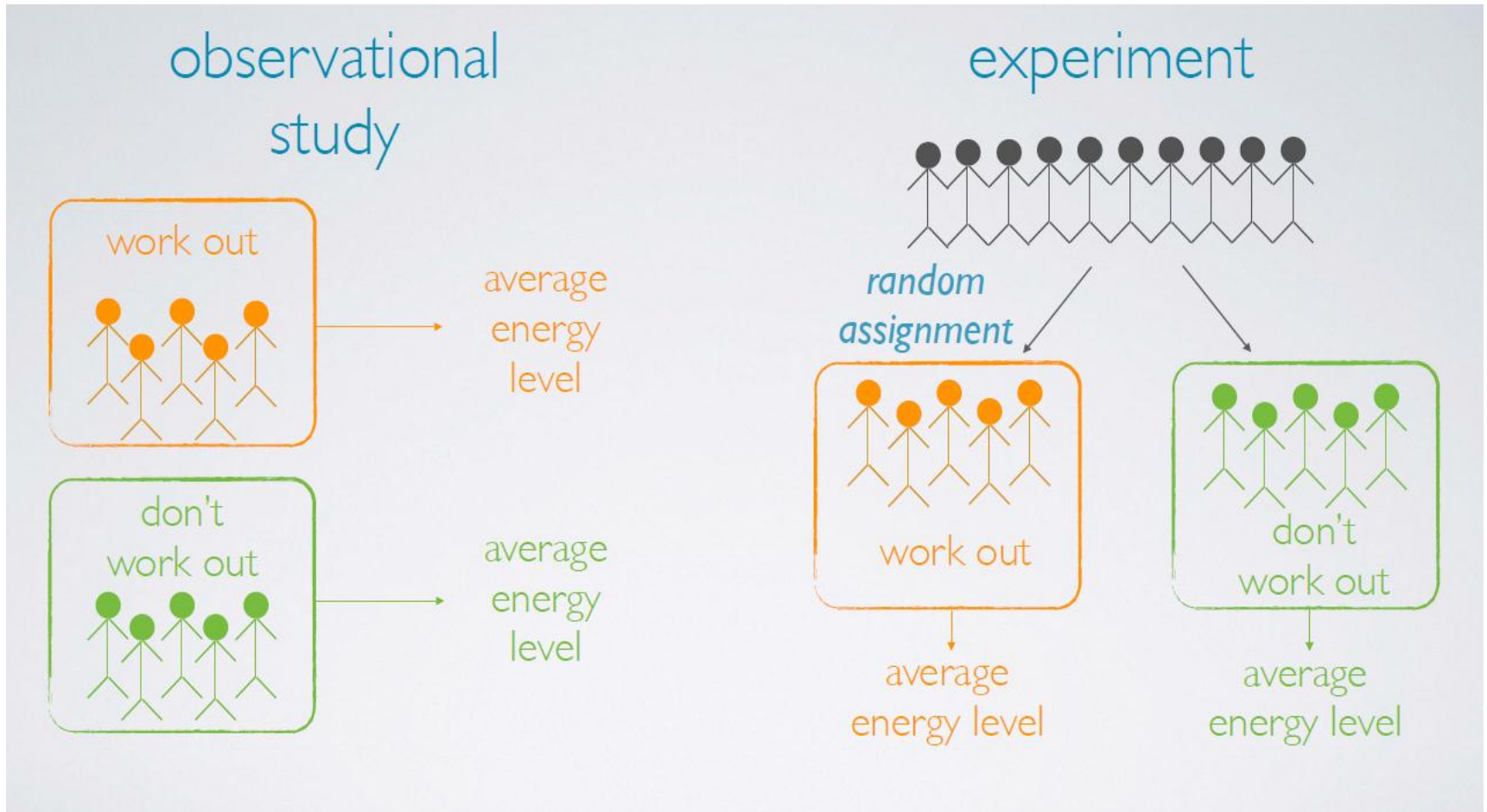


- ▶ collect data in a way that does not directly interfere with how the data arise (“observe”)
- ▶ only establish an association
- ▶ **retrospective**: uses past data
- ▶ **prospective**: data are collected throughout the study

- ▶ randomly assign subjects to treatments
- ▶ establish causal connections

Observational studies & experiments

20



Correlation & Causation

21

□ 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.

[...]

Possible explanations

22

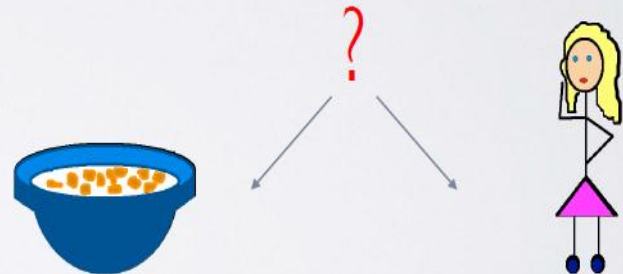
1. eating breakfast causes girls to be slimmer



2. being slim causes girls to eat breakfast



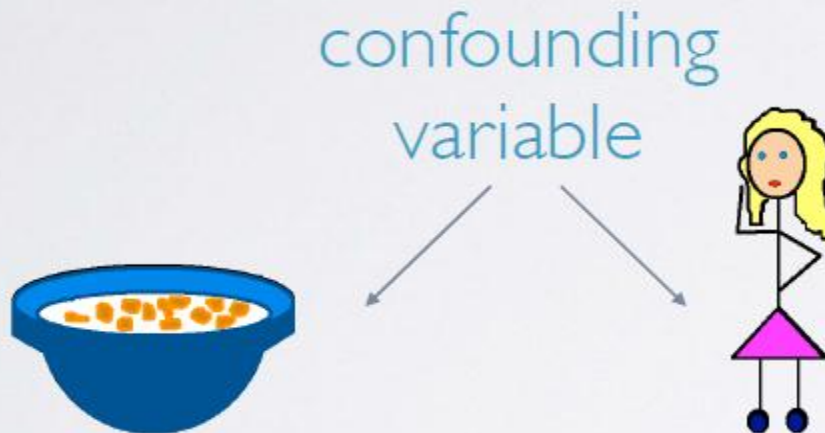
3. a third variable is responsible for both



Confounding variables

23

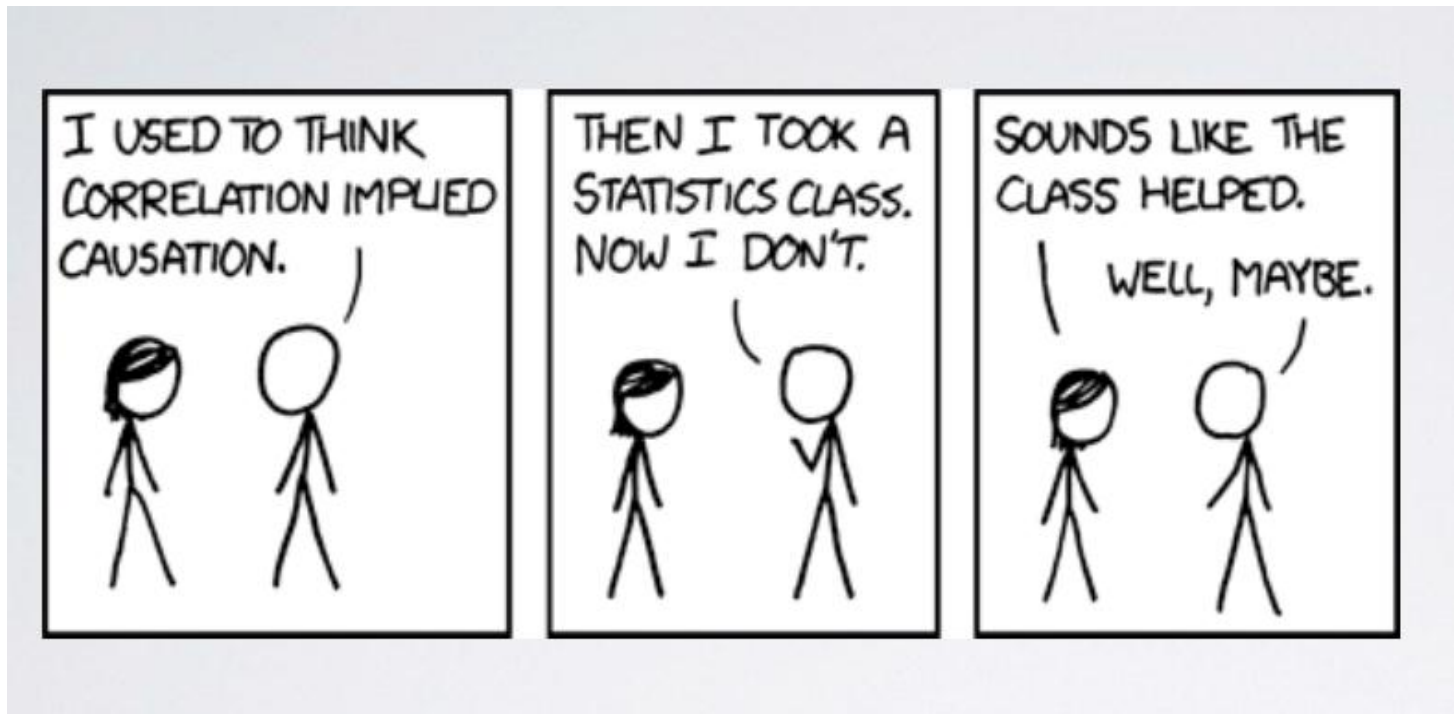
extraneous variables that affect both the explanatory and the response variable, and that make it seem like there is a relationship between them



Correlation & Causation

24

- Correlation does not imply causation



Sampling & sources bias

25

- Census vs sample
- Source o bias
- Sampling methods

Census

26

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



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.

A few sources of sampling bias

27

- ▶ **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

28

1936

Landon vs. FDR
(Republican) (Democrat)

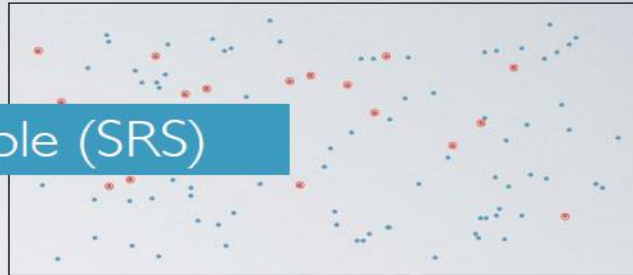
The Literary Digest
Election results

Lose with 43% of the votes
Win with 62% of the votes

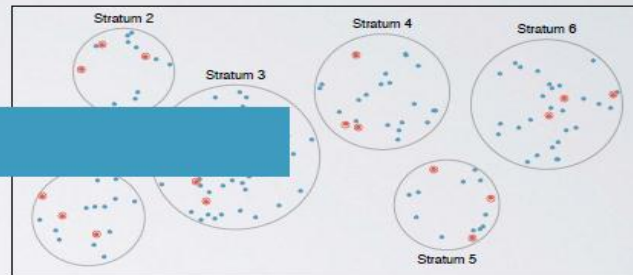
Sampling methods

29

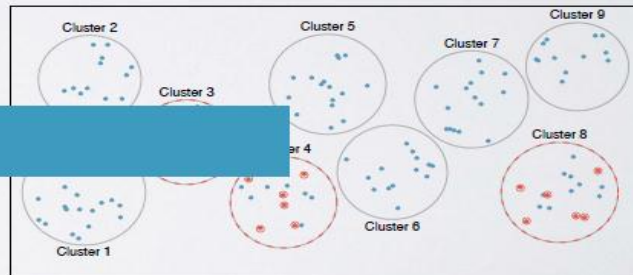
simple random sample (SRS)



stratified sample

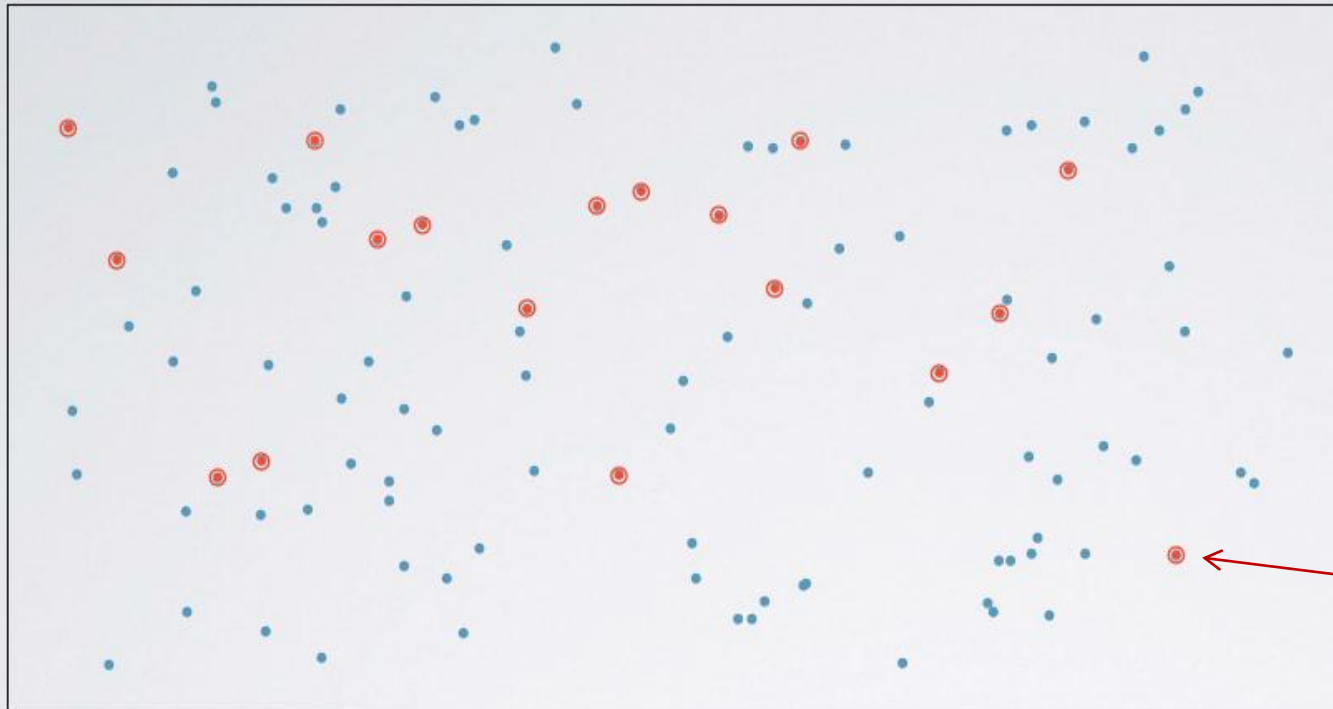


cluster sample



Sampling methods: random sampling

30

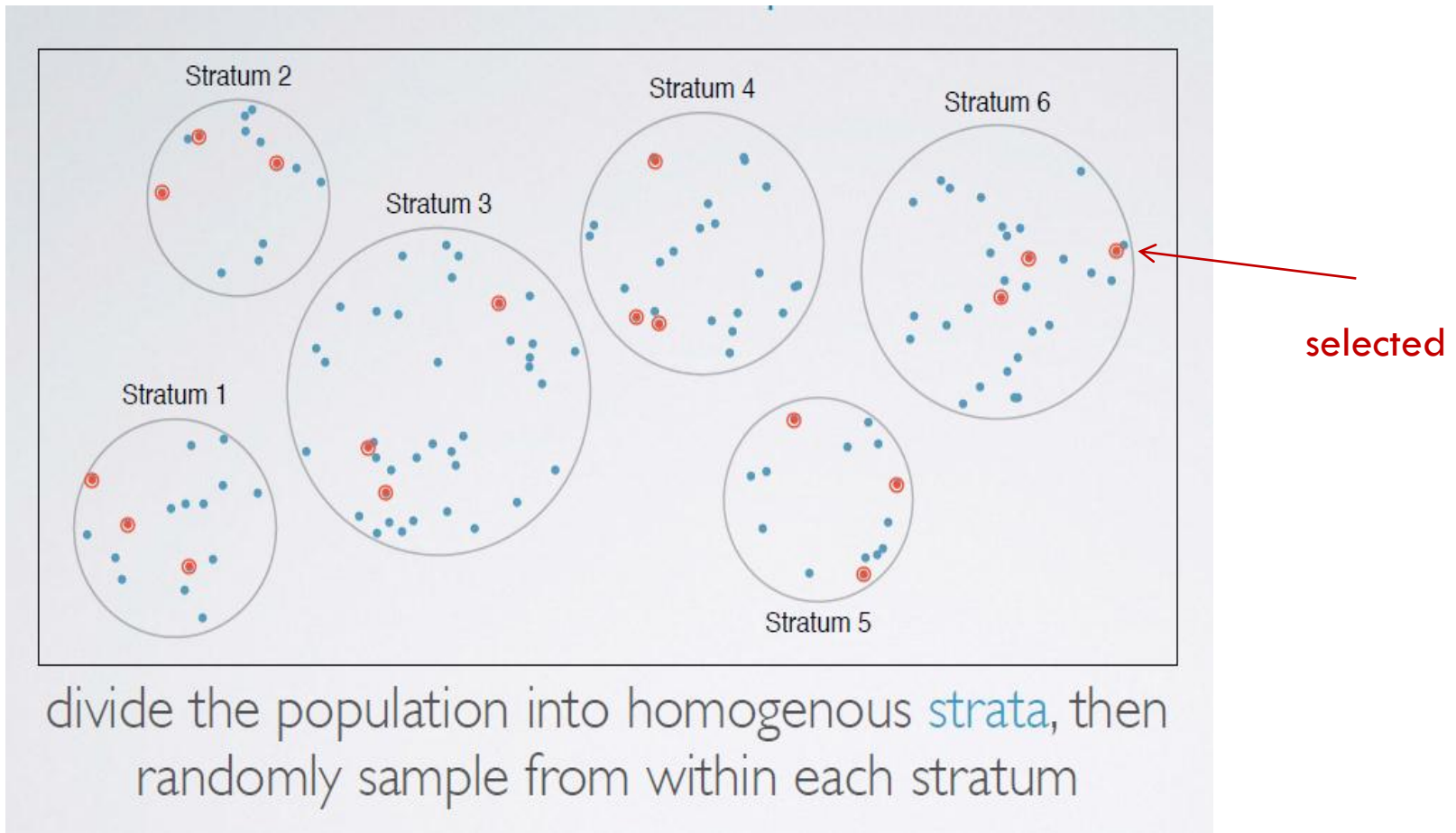


selected

each case is equally likely to be selected

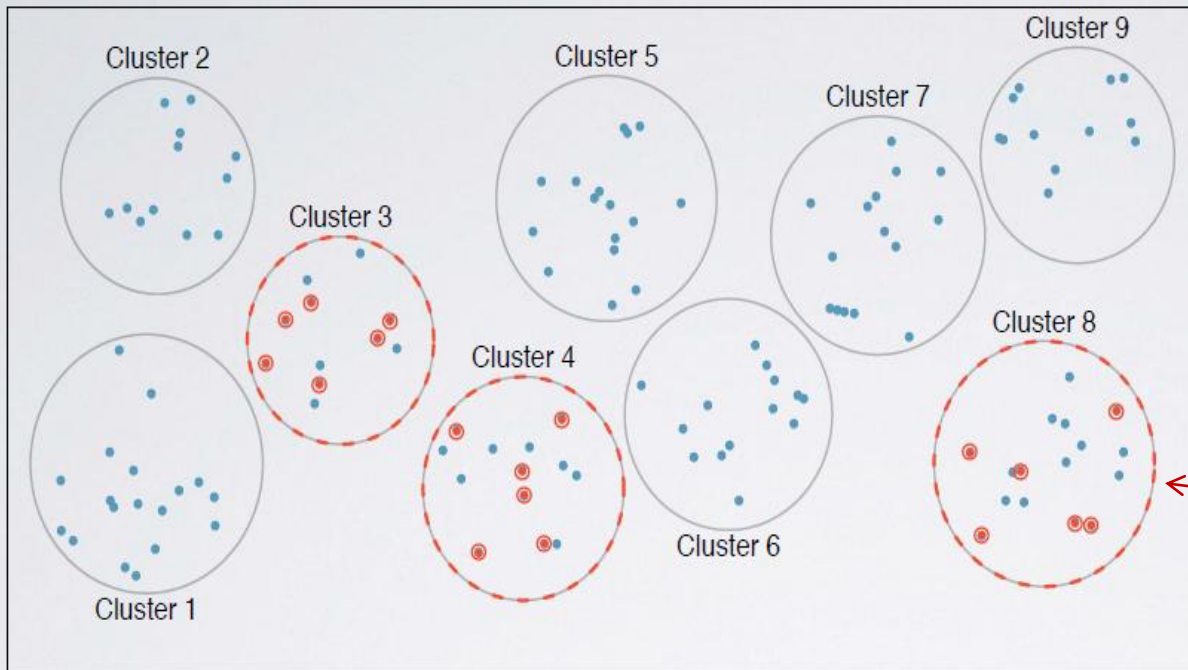
Sampling methods: stratified sample

31



Sampling methods: cluster

32



selected

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

Vizualizing numerical data

33

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

Data matrix

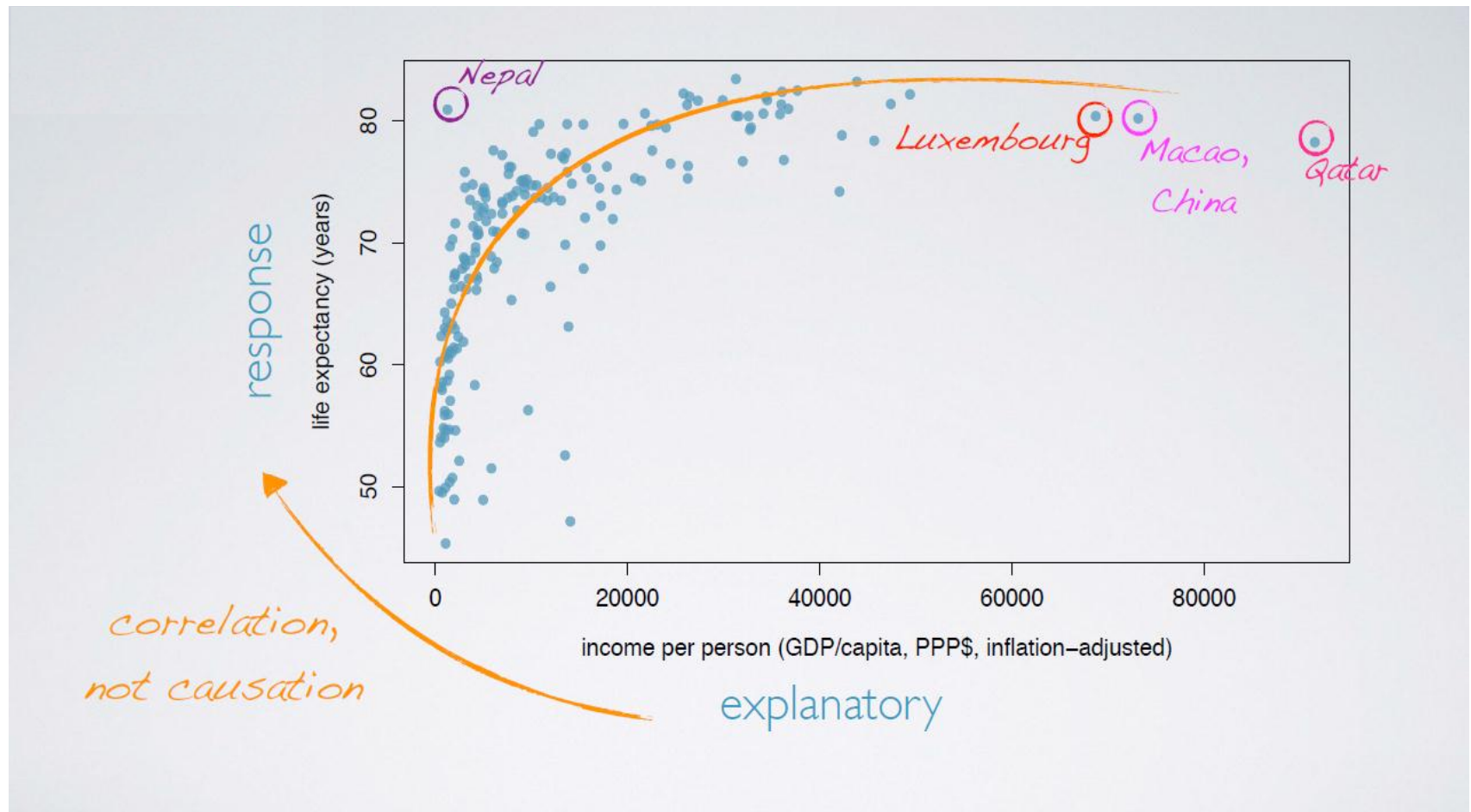
34

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

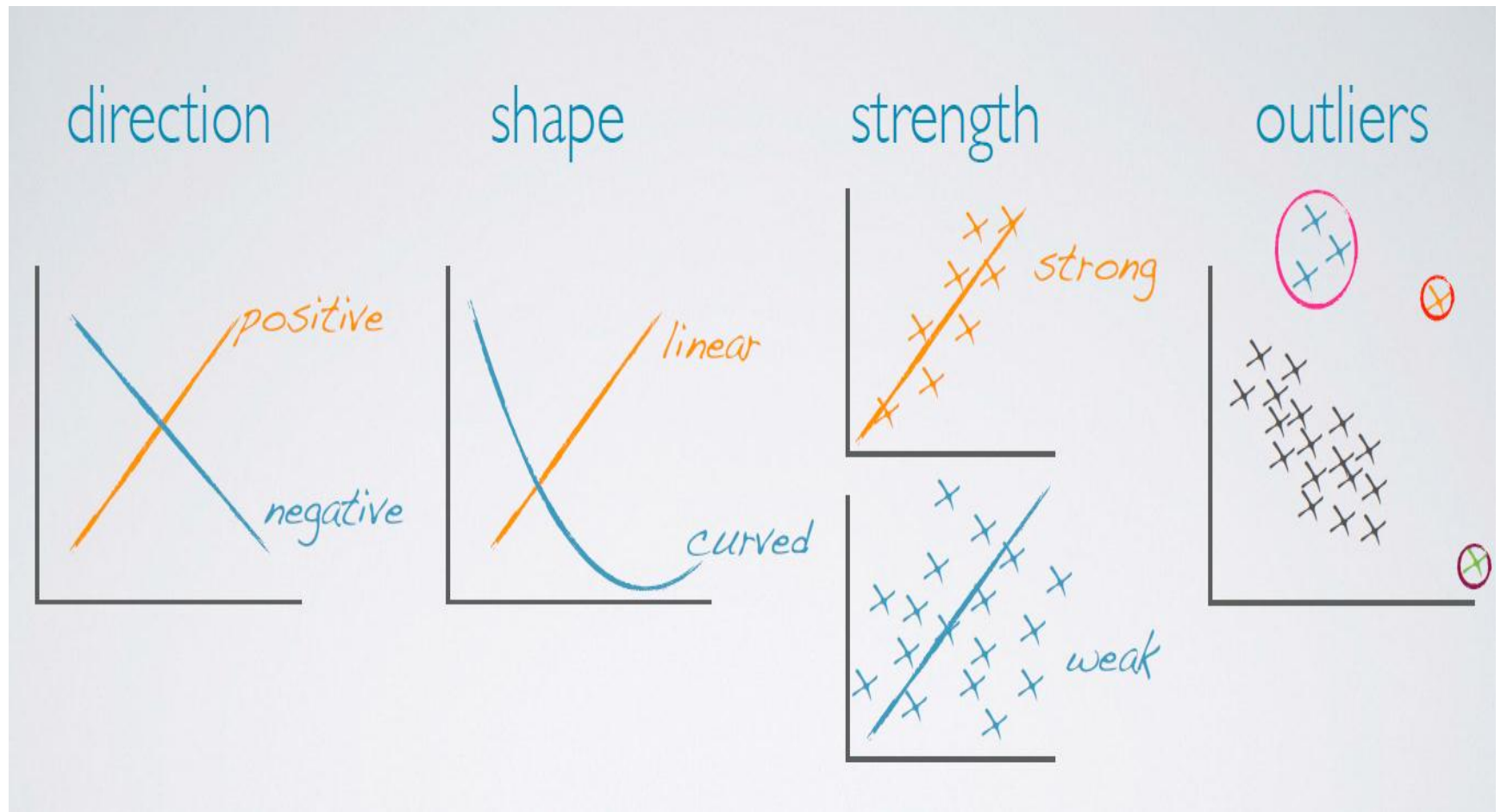
Scatterplots

35



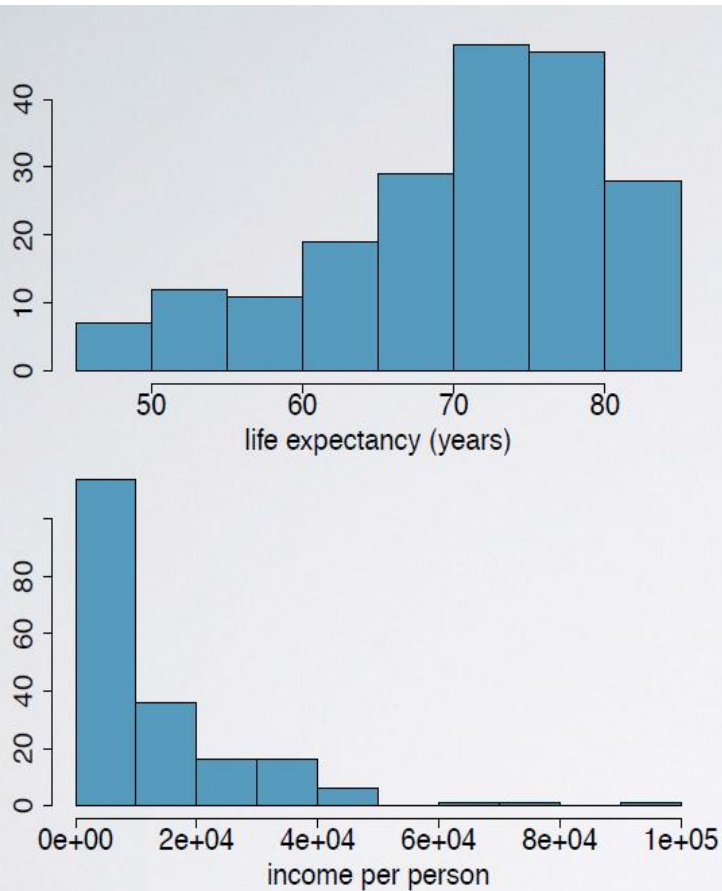
Evaluating their relationship

36



Histogram

37



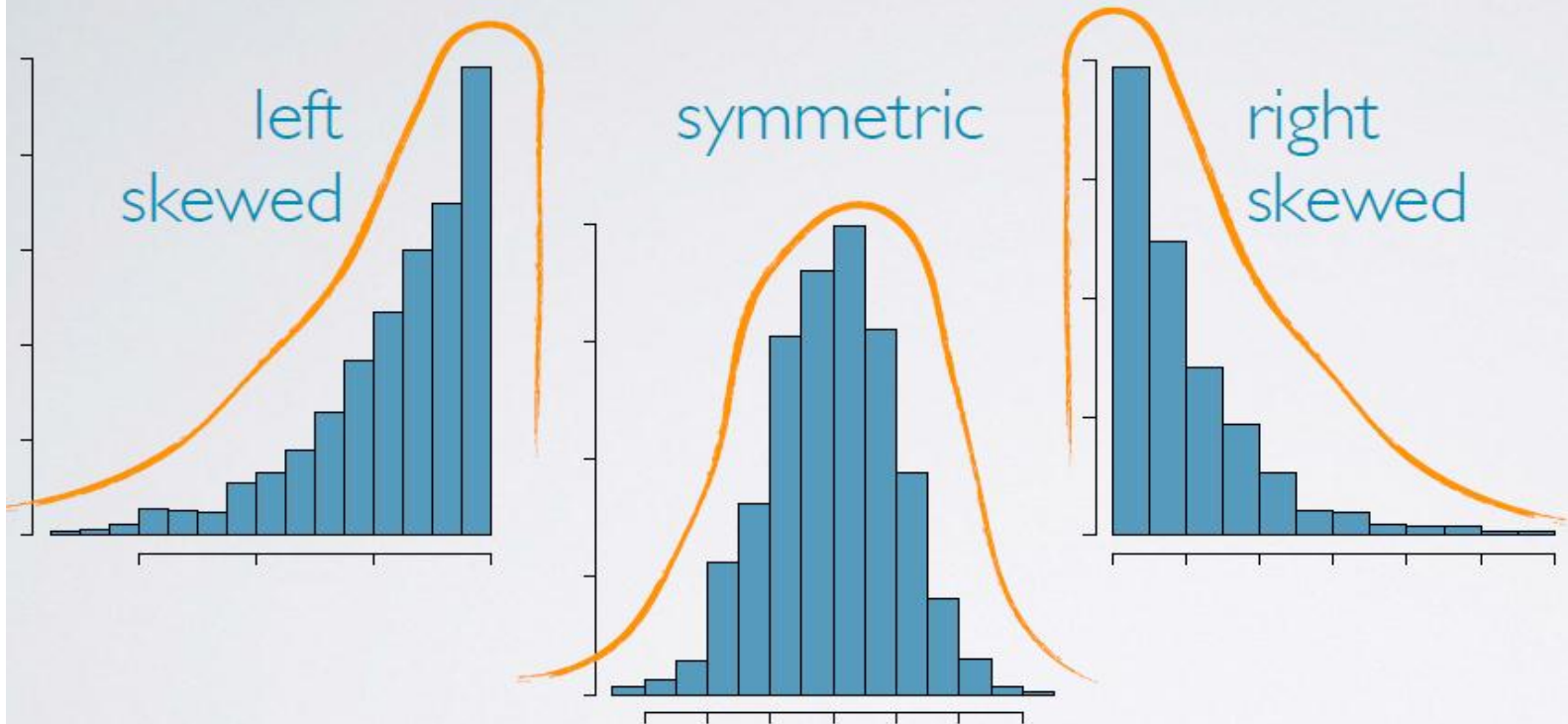
histogram

- ▶ provides a view of the [data density](#)
- ▶ especially useful for describing the shape of the distribution

Histogram

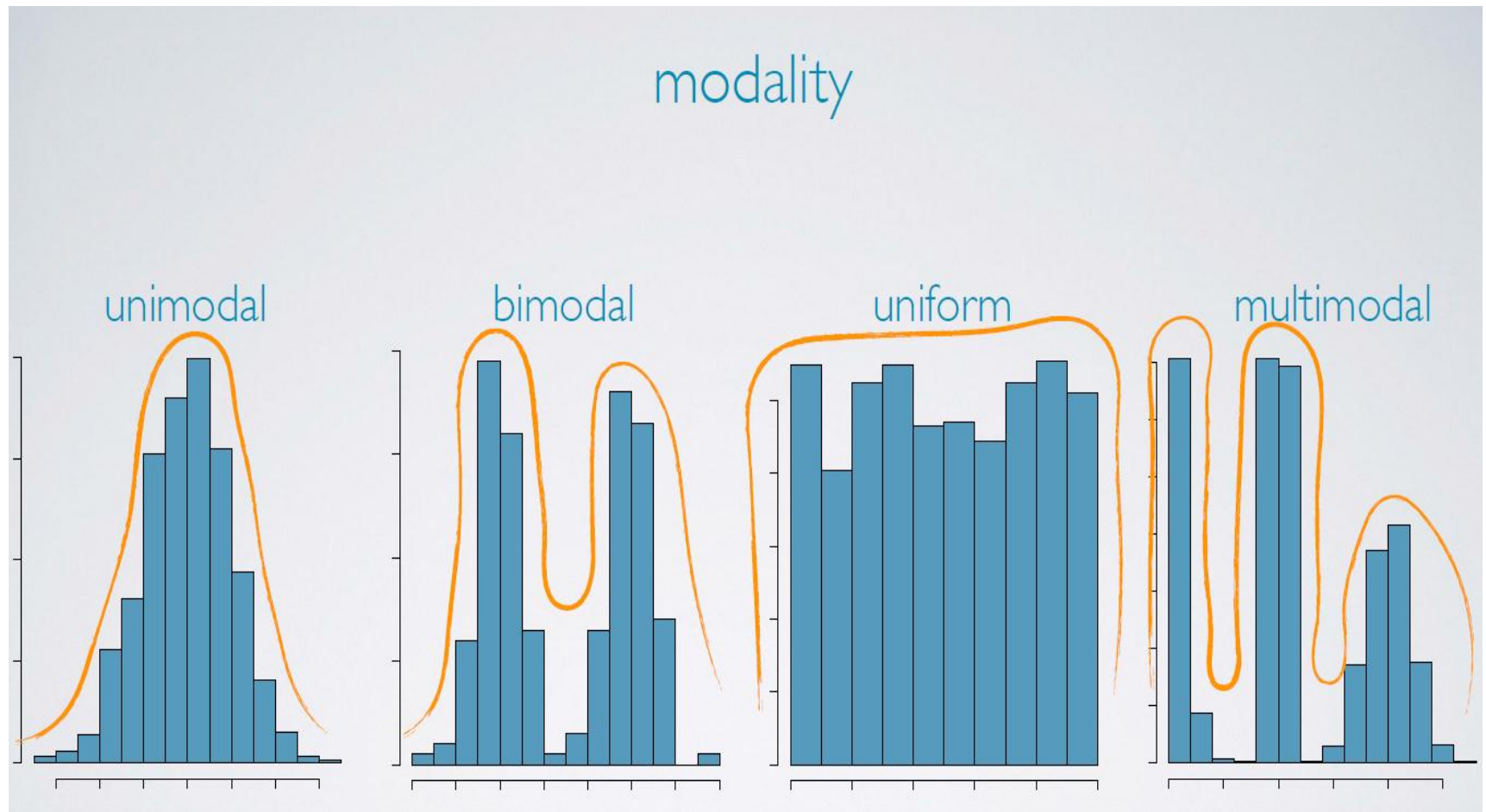
38

distributions are skewed to the side of the long tail



Histogram

39

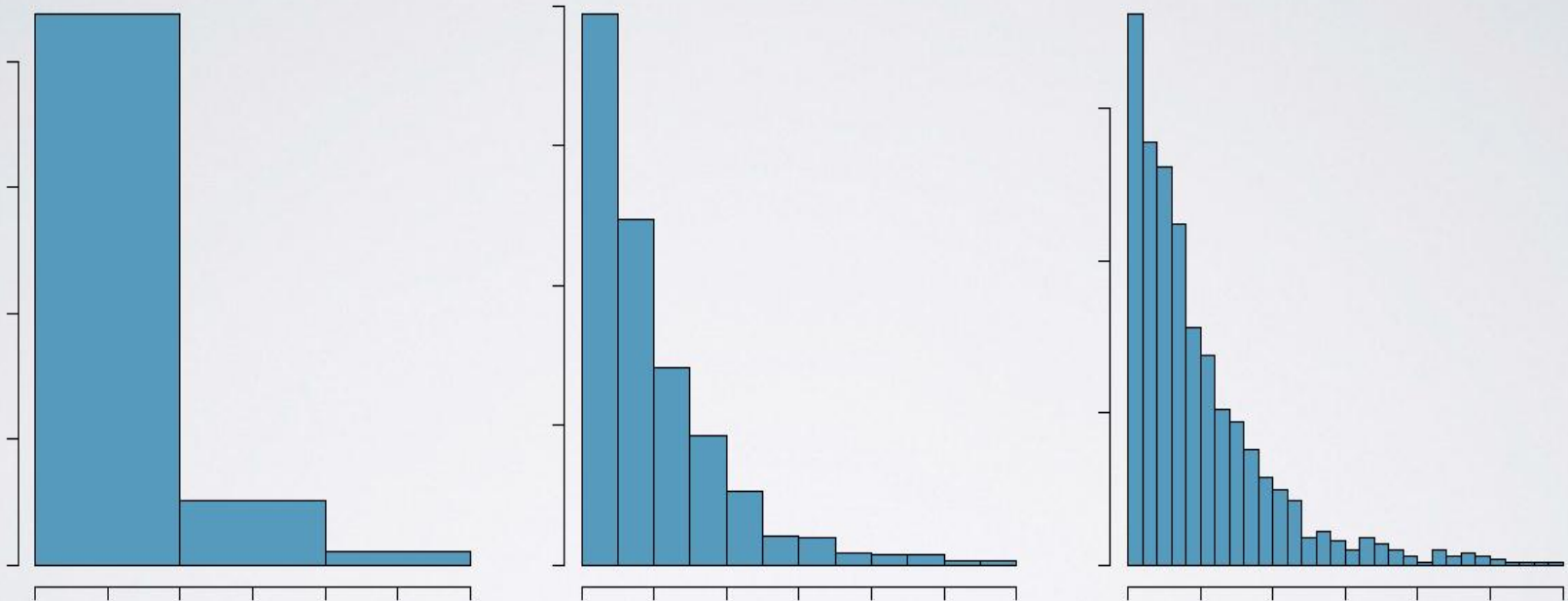


Histogram

40

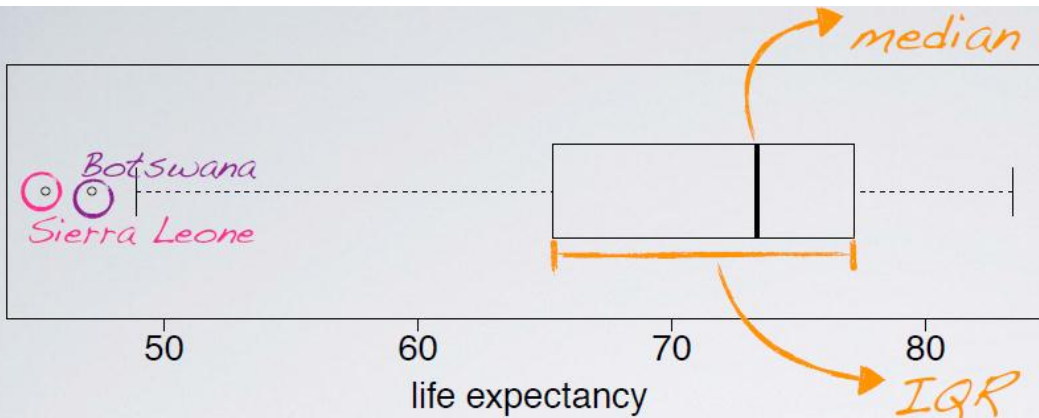
histogram & bin width

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



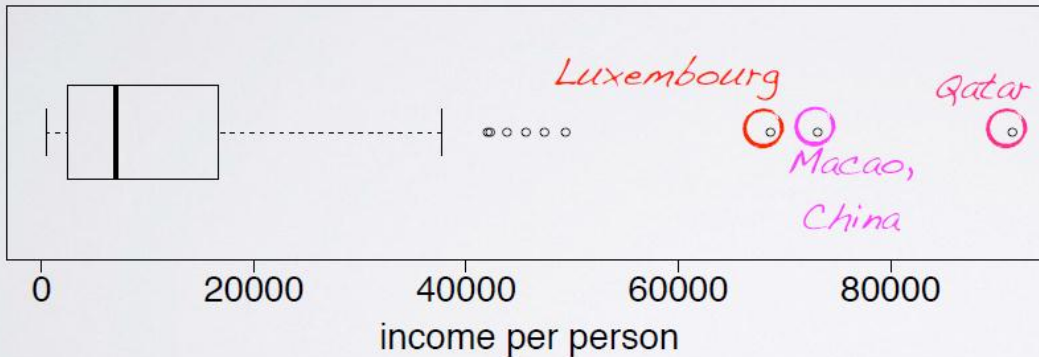
Box plot

41



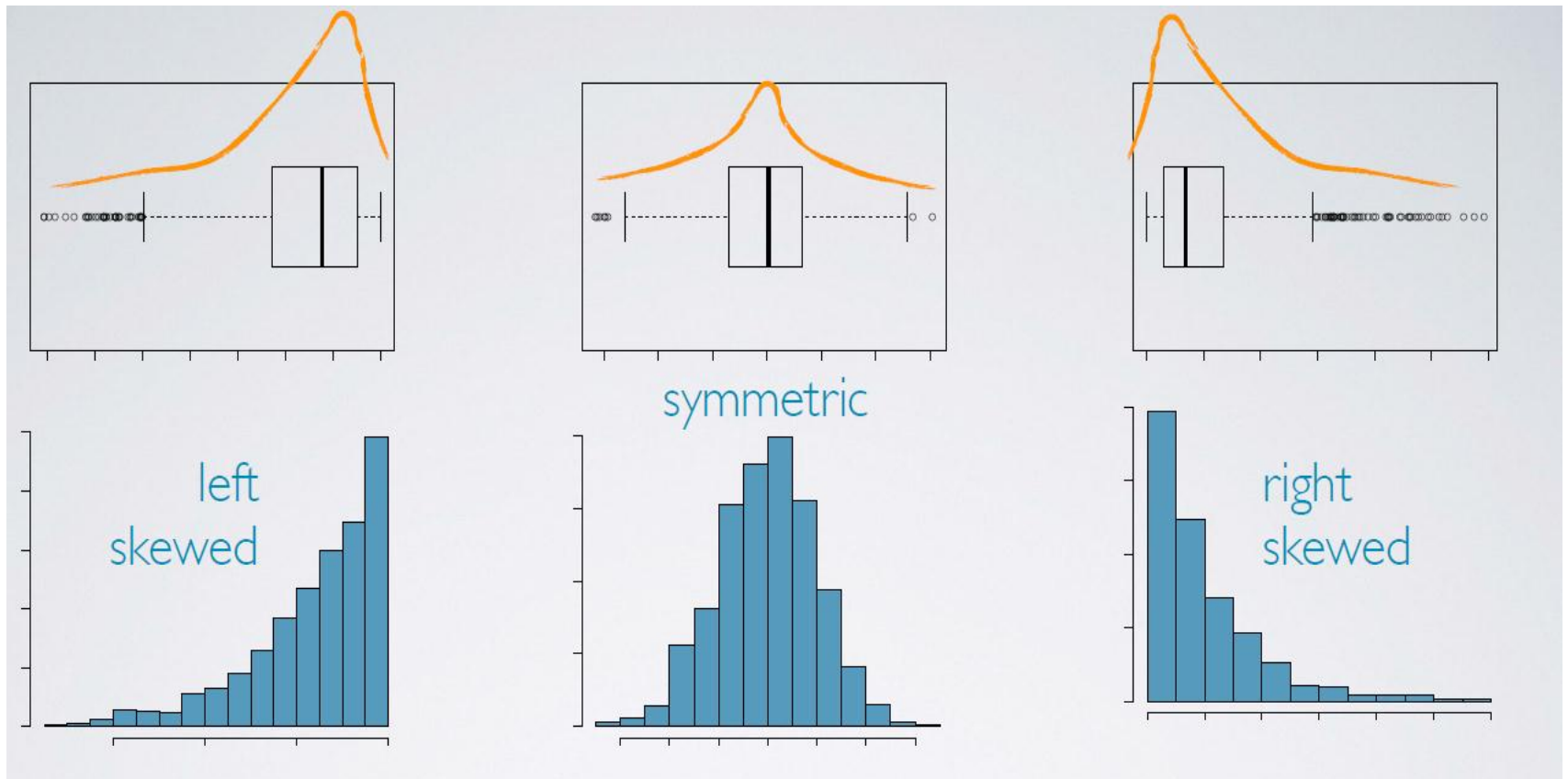
box plot

useful for highlighting outliers, median, IQR



Box plot

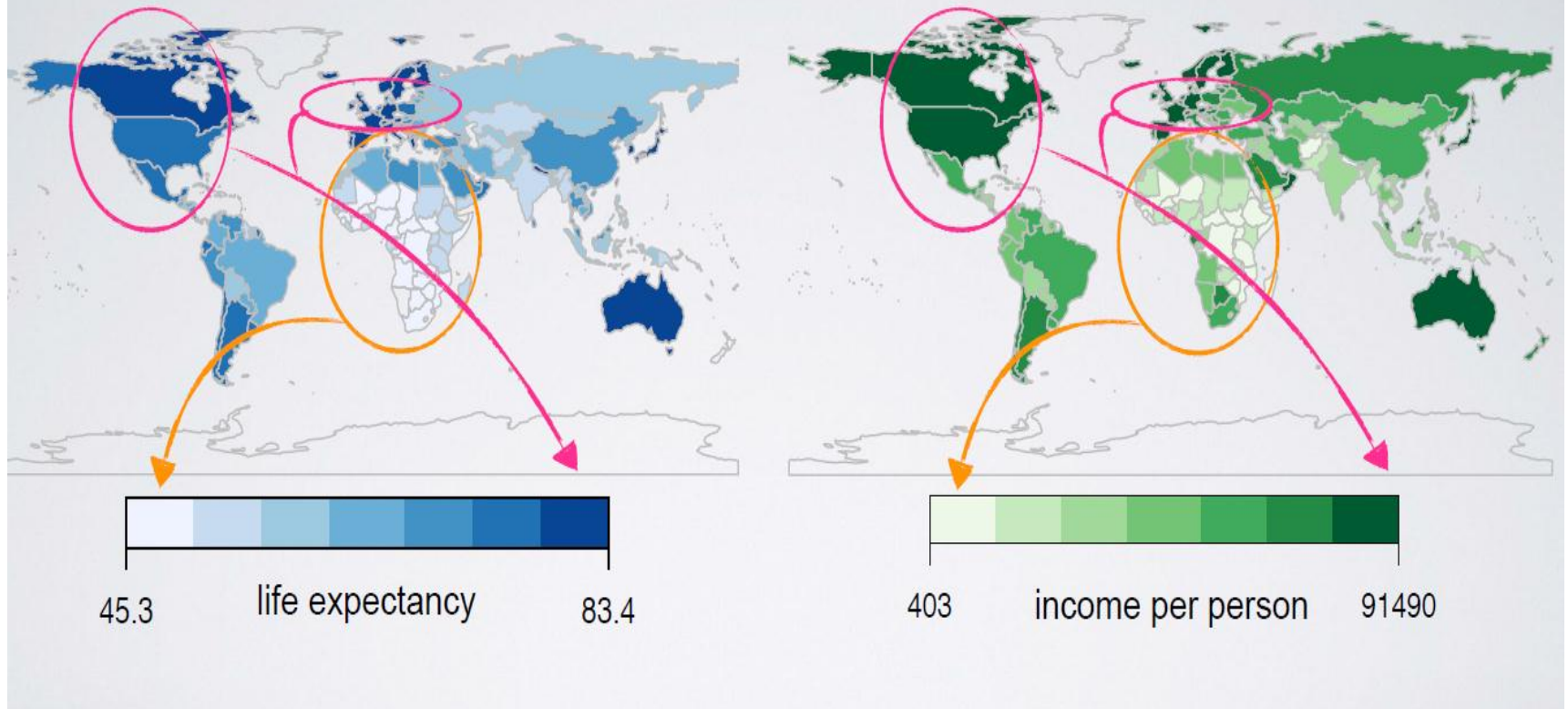
42



Intensity map

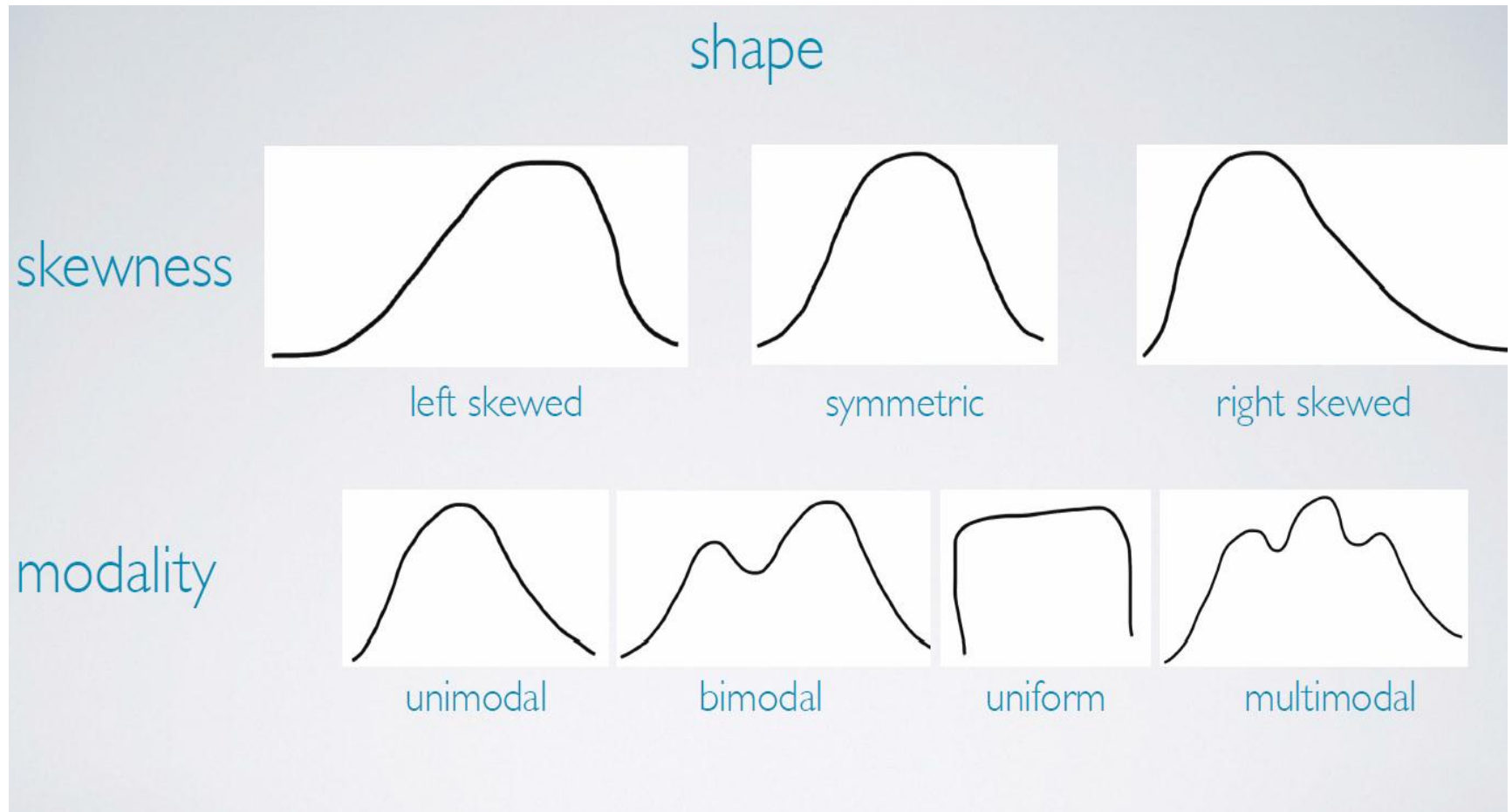
43

► Useful for highlighting the spatial distribution.



Measures of center

44



Measures of center

45

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

Measures of center

46

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

Data matrix

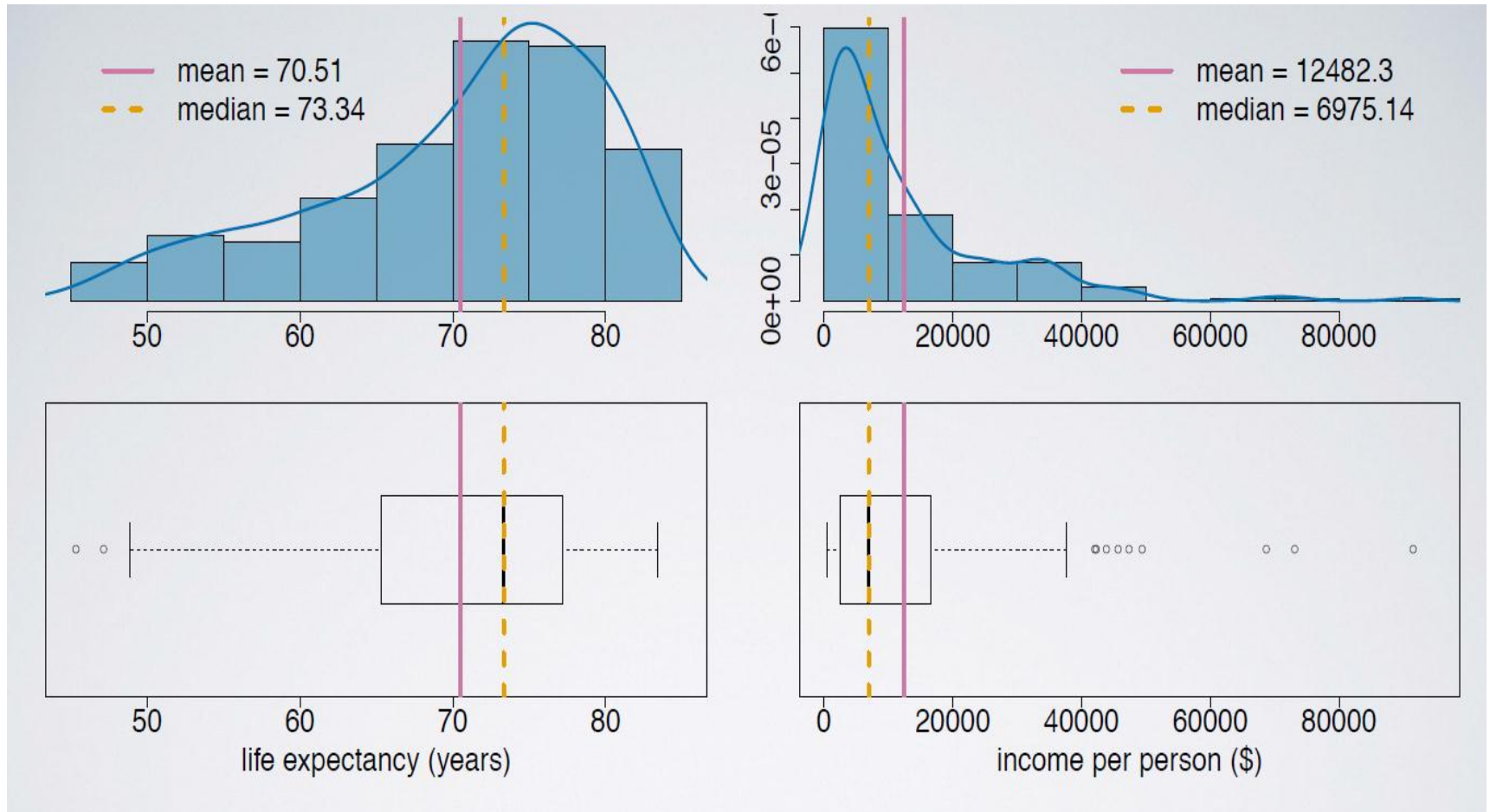
47

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

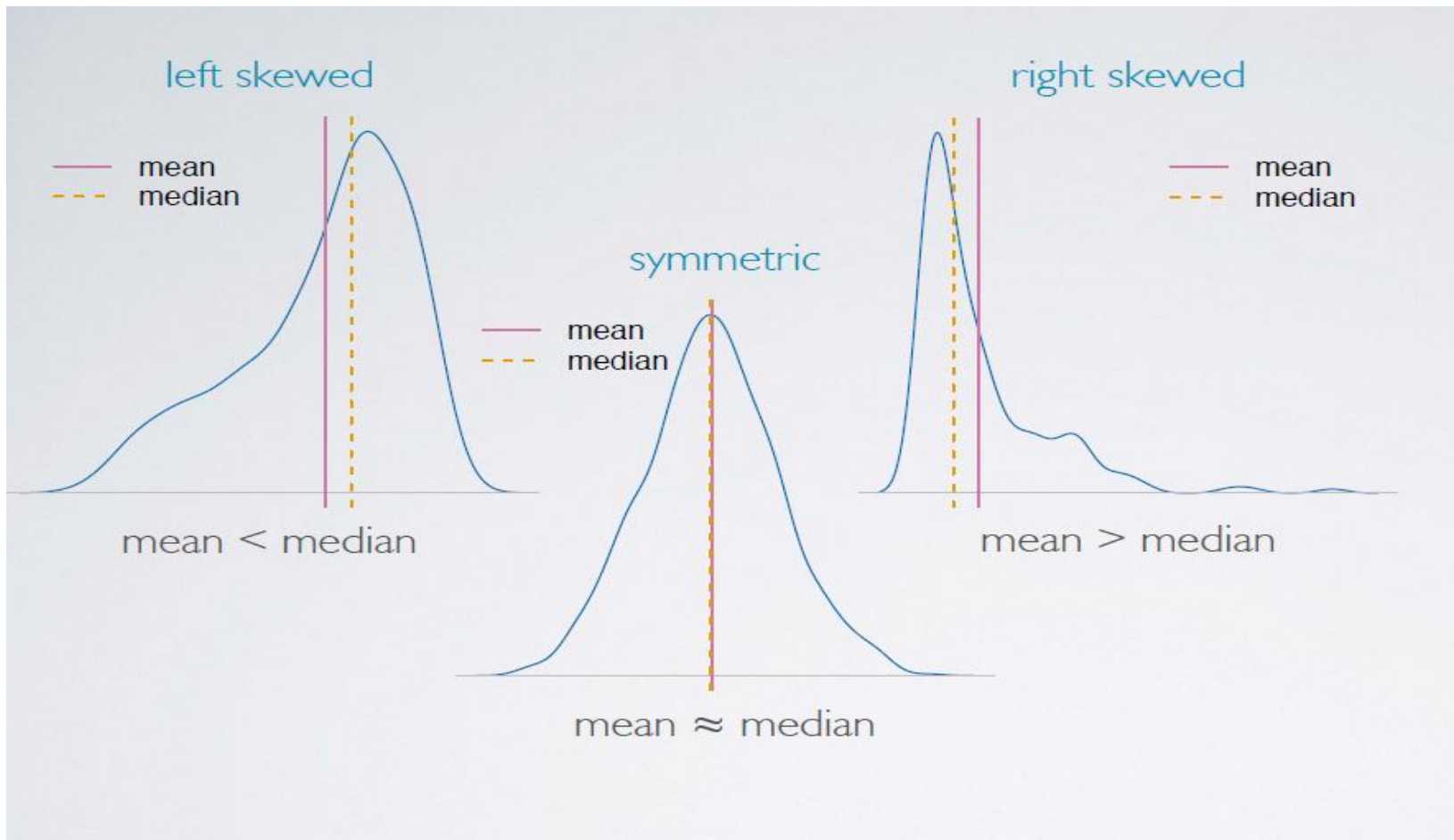
Measures of center

48



Skewness vs. measures of center

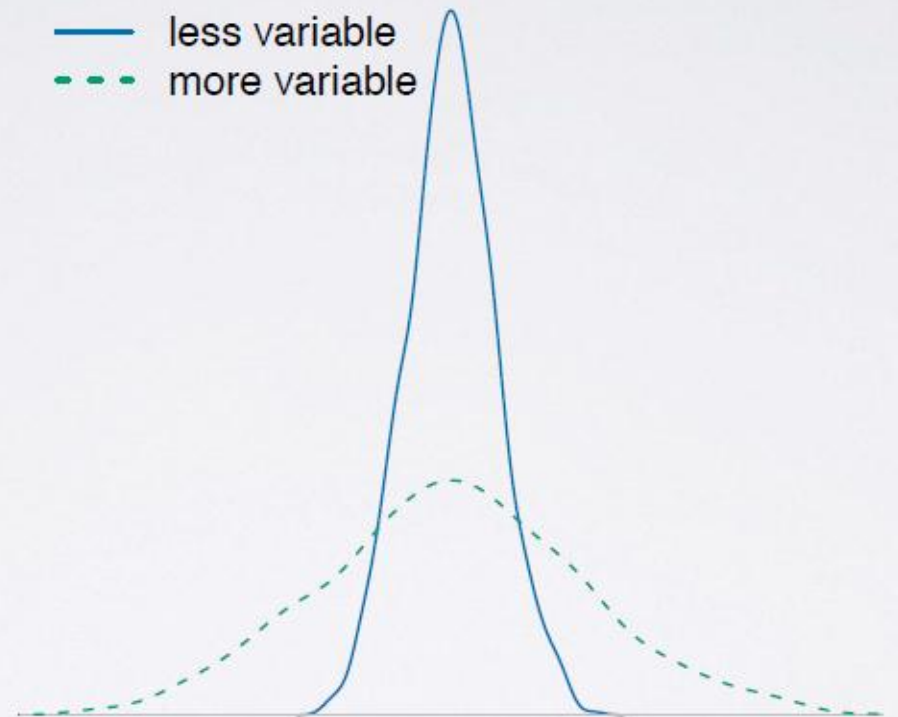
49



Measures of spread

50

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



Measures of spread

51

variance

sample
variance
 s^2
population
variance
 σ^2

roughly the average squared deviation from the mean

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

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

$$s^2 = \frac{(60.3 - 70.5)^2 + (77.2 - 70.5)^2 + \dots + (58.1 - 70.5)^2}{201 - 1}$$
$$= 83.06 \text{ years}^2$$

	country	life exp
1	Afghanistan	60.3
2	Albania	77.2
3	Algeria	70.9

201	Zimbabwe	58.1

Measures of spread

52

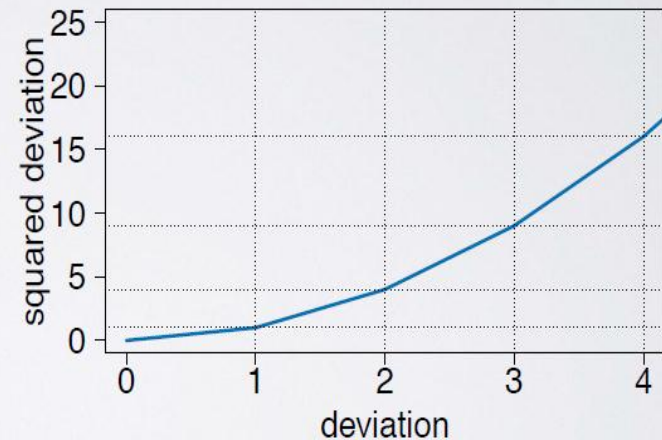
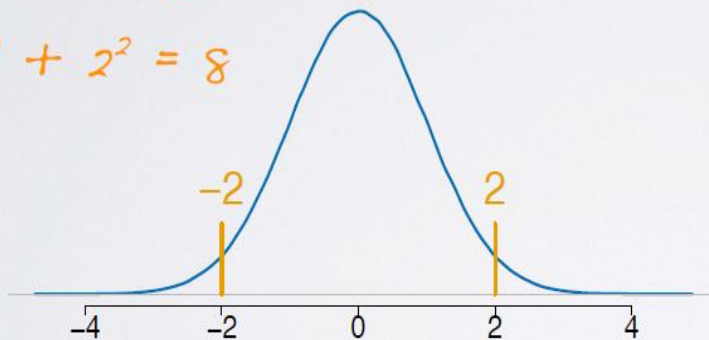
Why do we square the differences?

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

- ▶ 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

$$(-2) + 2 = 0$$

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



Measures of spread

53

standard deviation

sample sd
 s
population sd
 σ

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

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

*square root of
the variance*

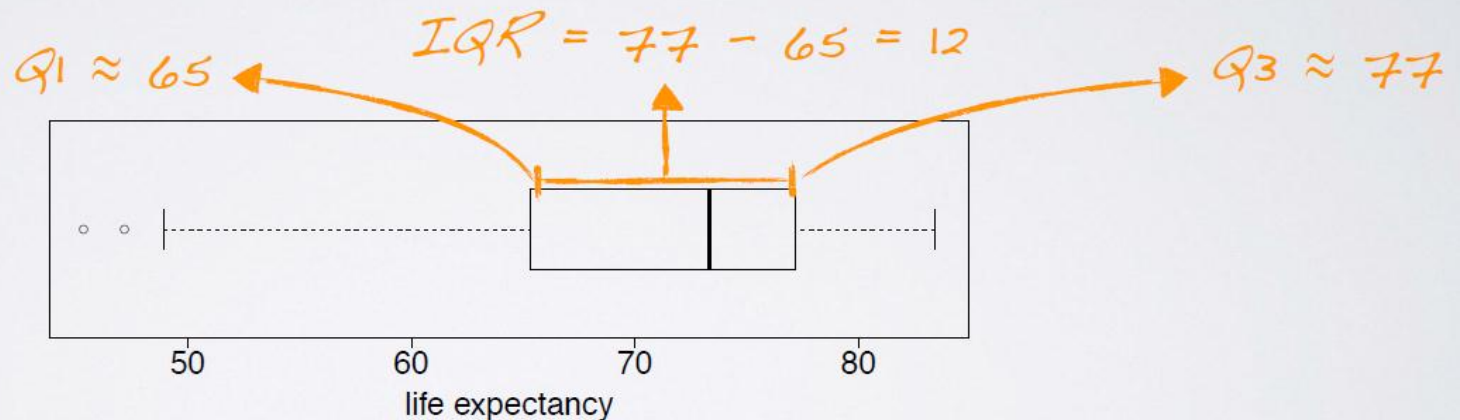
Measures of spread

54

interquartile range

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

$$IQR = Q3 - Q1$$



Robust statistics

55

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 statistics

56

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

*skewed,
with extreme
observations*

symmetric

Transforming data

57

- ▶ 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

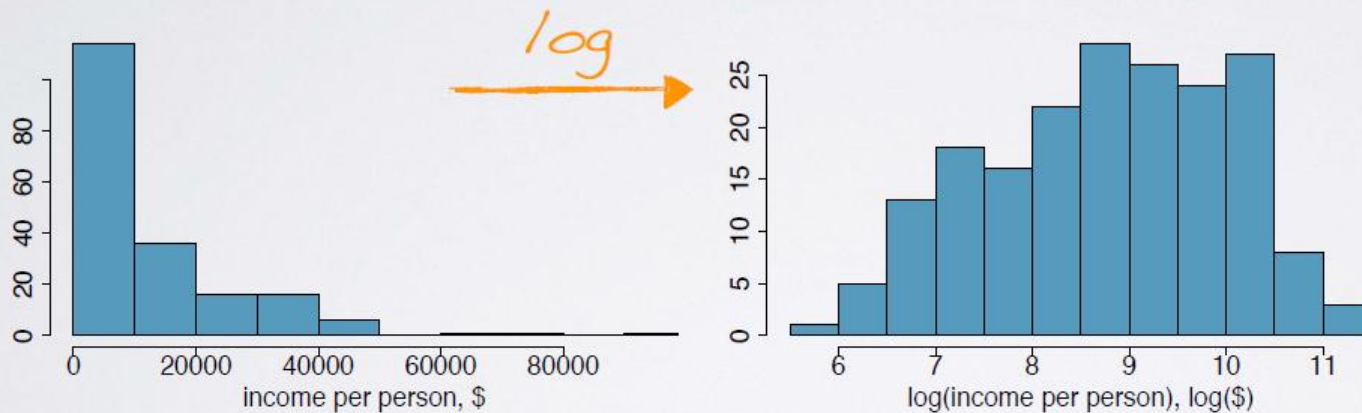
- ▶ to see the data structure differently
- ▶ to reduce skew assist in modeling
- ▶ to straighten a nonlinear relationship in a scatterplot

Transforming data

58

(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

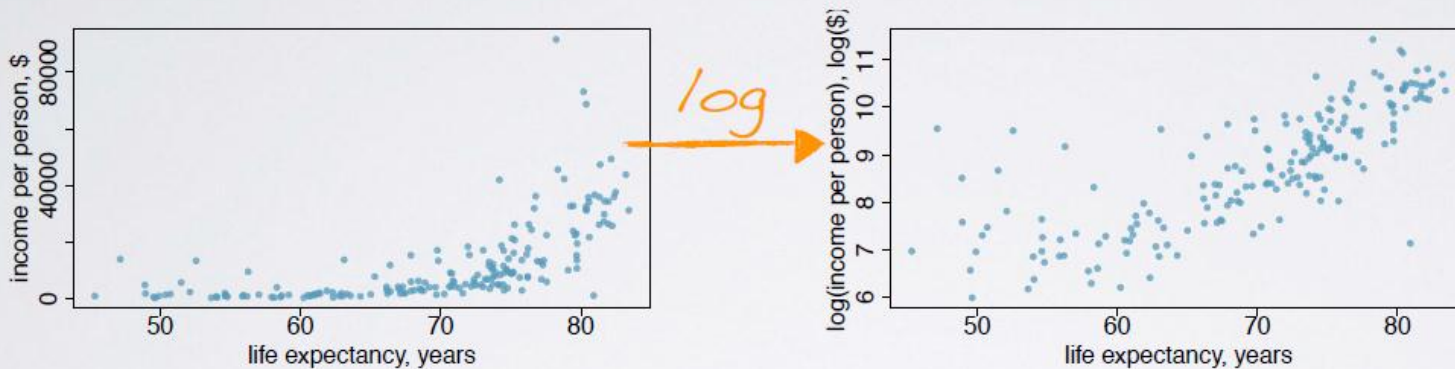


Transforming data

59

log transformation

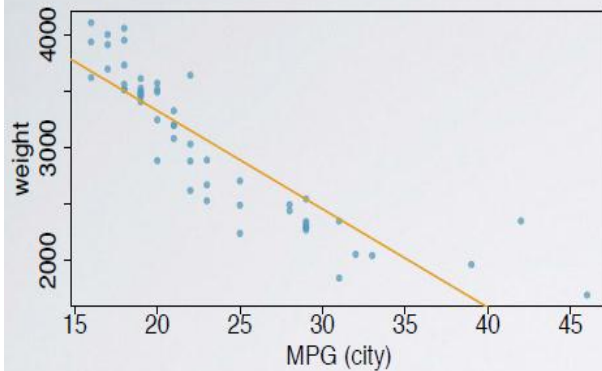
to make the relationship between the variables more linear, and hence easier to model with simple methods



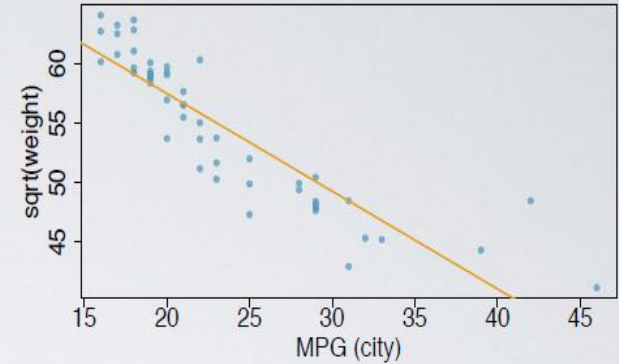
Transforming data

60

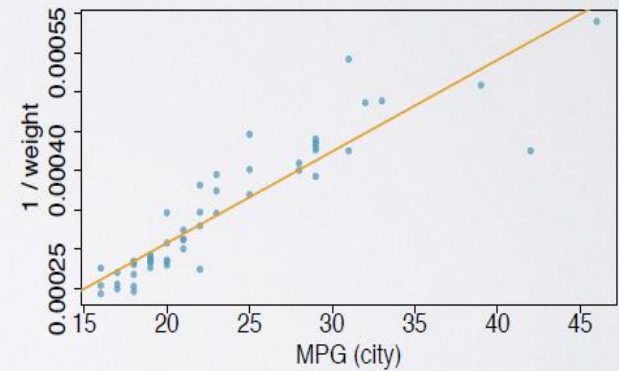
other transformations



square root



inverse



Exploring categorical variables

61

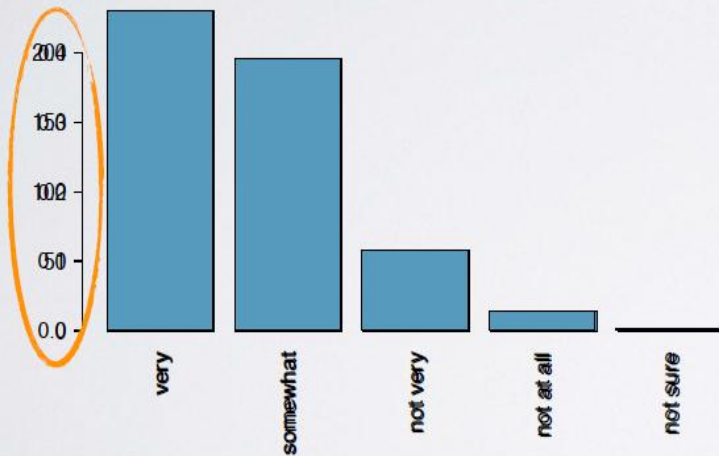
- ▶ describe distribution of a single categorical variable
- ▶ evaluate relationship between two categorical variables
- ▶ evaluate relationship between a categorical and a numerical variable

Exploring categorical variables

62

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

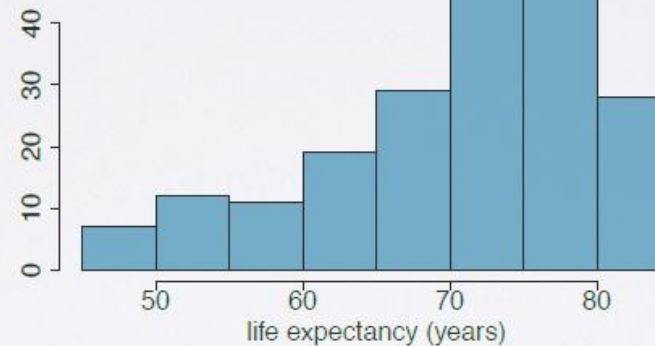
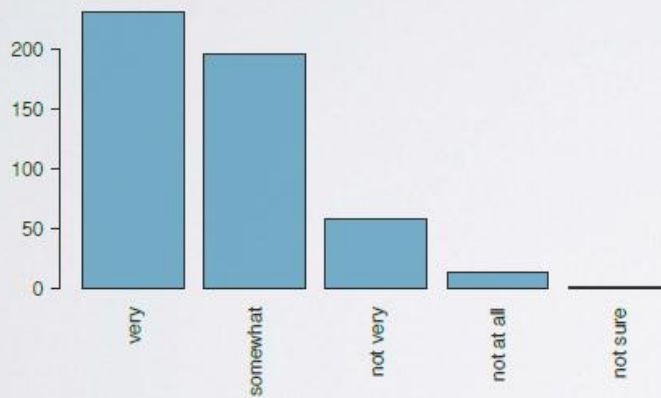


Exploring categorical variables

63

How are bar plots different than histograms?

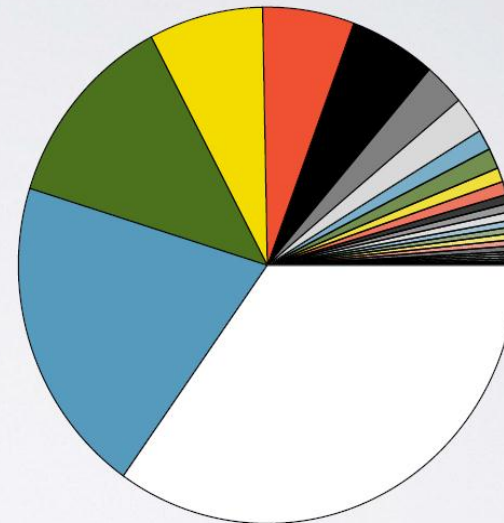
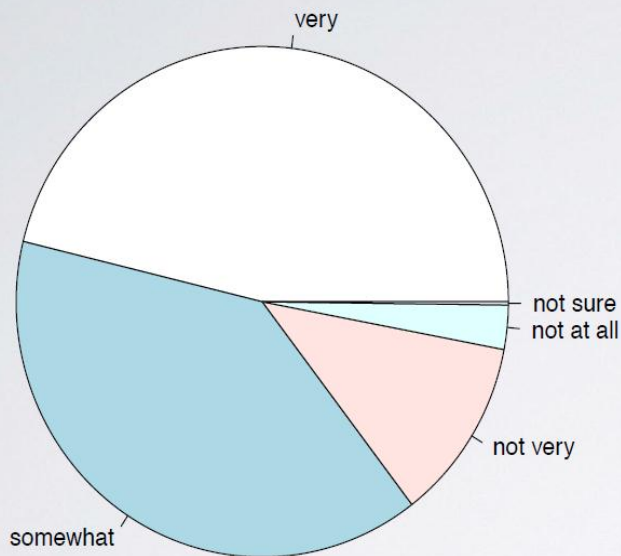
- ▶ barplots for categorical variables, histograms for numerical variables
- ▶ x-axis on a histogram is a number line, and the ordering of the bars are not interchangeable



Exploring categorical variables

64

~~pie chart?~~ no!



- RODENTIA
- CHIROPTERA
- CARNIVORA
- ARTIODACTYLA
- PRIMATES
- SORICOMORPHA
- LAGOMORPHA
- DIPROTODONTIA
- DIDELPHIMORPHA
- CETACEA
- DASYUROMORPHA
- AFROSORICIDA
- ERINACEOMORPHA
- SCANDENTIA
- PERISSODACTYLA
- HYRACOIDEA
- PERAMELEMORPHA
- CINGULATA
- PILOSA
- MACROSCELIDEA
- TUBULIDENTATA
- PHOLIDOTA
- MONOTREMATA
- PAUCITUBERCULATA
- SIRENIA
- PROBOSCIDEA
- DERMOPTERA
- NOTORYCTEMORPHA
- MICROBIOTHERIA

Exploring categorical variables

65

contingency table

		Income				
		< \$40K	\$40-80K	> \$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

Exploring categorical variables

66

□ 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: $31 / 124 = 25\%$

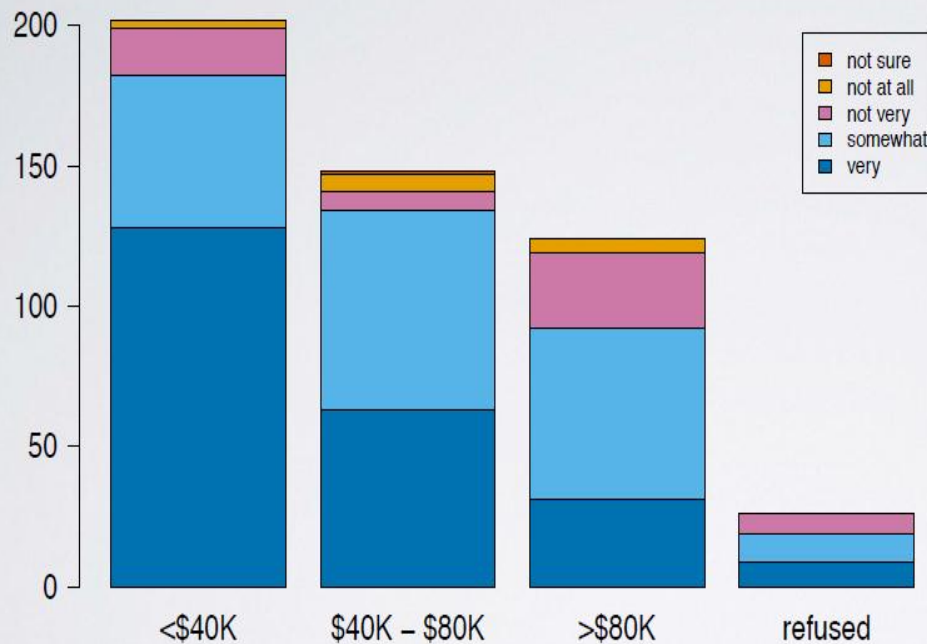
Refused: $9 / 26 = 35\%$

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

Exploring categorical variables

67

segmented bar plot

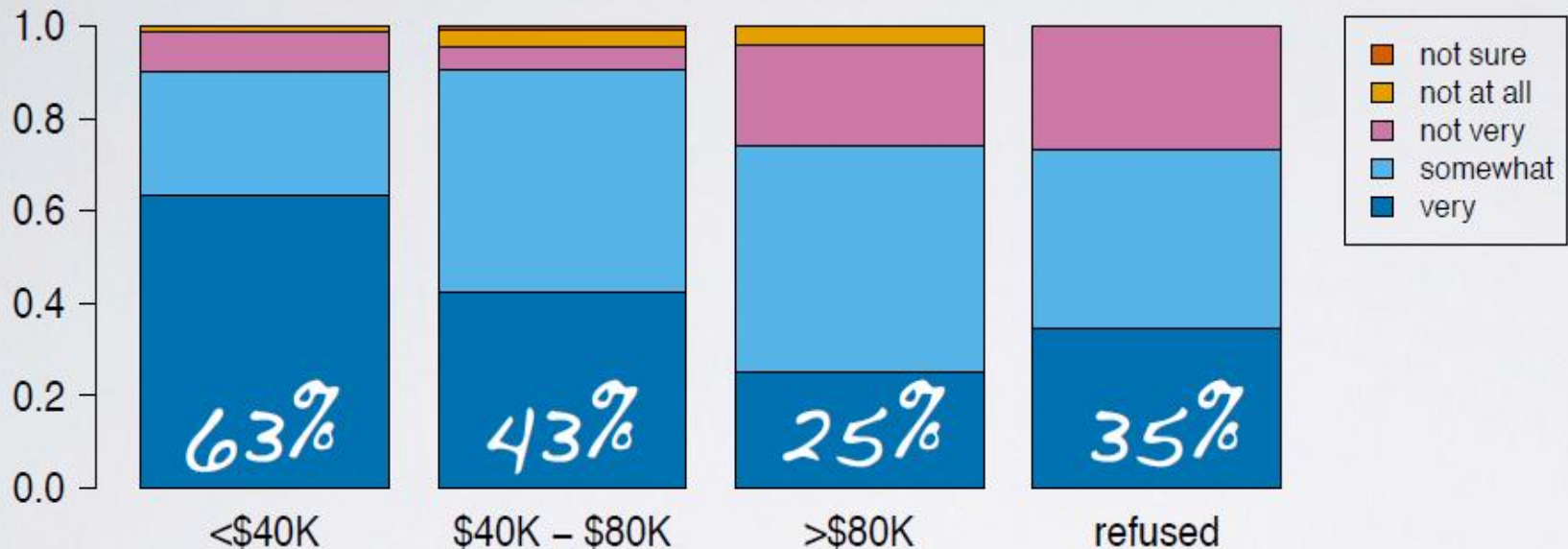


- ▶ useful for visualizing conditional frequency distributions
- ▶ compare relative frequencies to explore the relationship between the variables

Exploring categorical variables

68

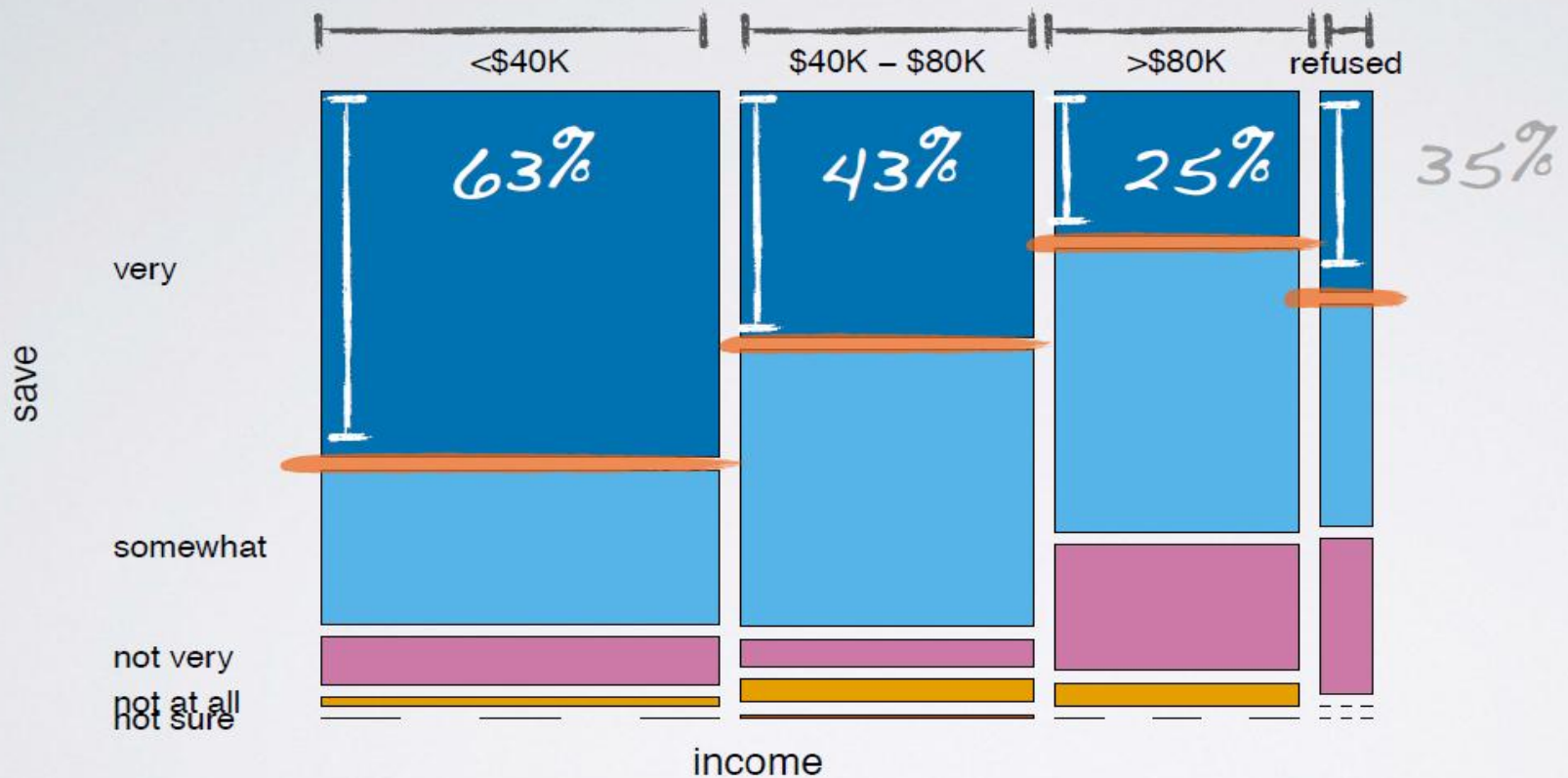
relative frequency segmented bar plot



Exploring categorical variables

69

mosaicplot



Exploring categorical variables

70

side-by-side box plots

