

19CSE304 - FOUNDATIONS OF DATA SCIENCE

Lecture: Introduction to Data Science

Dr. Venugopal K

Assoc Professor, CSE, ASE, Amritapuri

Introduction

Scientists seek to answer questions using rigorous methods and careful observations. These observations – collected from the likes of field notes, surveys, and experiments – form the backbone of a statistical investigation and are called data.

Statistics is the study of how best to collect, analyze, and draw conclusions from data, and in the initial sections, we focus on both the properties of data and on the collection of data.

What is Data Science?

- Data Science is about drawing useful conclusions from large and diverse data sets through exploration, prediction, and inference.
- Exploration involves identifying patterns in information.
- Prediction involves using information we know to make informed guesses about values we wish we knew.
- Inference involves quantifying our degree of certainty: will the patterns that we found in our data also appear in new observations? How accurate are our predictions?
- Our primary tools for exploration are visualizations and descriptive statistics, for prediction are machine learning and optimization, and for inference are statistical tests and models.

- Statistics is a central component of data science because statistics studies how to make robust conclusions based on incomplete information.
- Computing is a central component because programming allows us to apply analysis techniques to the large and diverse data sets that arise in real-world applications: not just numbers, but text, images, videos, and sensor readings.
- Through understanding a particular domain, data scientists learn to ask appropriate questions about their data and correctly interpret the answers provided by our inferential and computational tools.

- In summary, Data science is the discipline of drawing conclusions from data using computation. There are three core aspects of effective data analysis: exploration, prediction, and inference.
- A foundation in data science requires not only understanding statistical and computational techniques, but also recognizing how they apply to real scenarios.

Case study: using stents to prevent strokes

We consider an experiment that studies effectiveness of stents in treating patients at risk of stroke. Stents are devices put inside blood vessels that assist in patient recovery after cardiac events and reduce the risk of an additional heart attack or death.

Many doctors have hoped that there would be similar benefits for patients at risk of stroke. The principal question the researchers hope to answer is:

➤ ***Does the use of stents reduce the risk of stroke?***

The researchers conducted an experiment with 451 at-risk patients. Each volunteer patient was randomly assigned to one of two groups:

- ❑ **Treatment group.** Patients in the treatment group received a stent and medical management. The medical management included medications, management of risk factors, and help in lifestyle modification.

- ❑ **Control group.** Patients in the control group received the same medical management as the treatment group, but they did not receive stents.

- Researchers randomly assigned 224 patients to the treatment group and 227 to the control group. In this study, the control group provides a reference point against which we can measure the medical impact of stents in the treatment group.
- Researchers studied the effect of stents at two time points: 30 days after enrollment and 365 days after enrollment.
- The results of 5 patients are summarized in Figure 1.1. Patient outcomes are recorded as “stroke” or “no event”, representing whether or not the patient had a stroke at the end of a time period.

Patient	group	0-30 days	0-365 days
1	treatment	no event	no event
2	treatment	stroke	stroke
3	treatment	no event	no event
⋮	⋮	⋮	
450	control	no event	no event
451	control	no event	no event

Figure 1.1: Results for five patients from the stent study.

Considering data from each patient individually would be a long, cumbersome path towards answering the original research question. Instead, performing a statistical data analysis allows us to consider all of the data at once. Figure 1.2 summarizes the raw data in a more helpful way. In this table, we can quickly see what happened over the entire study.

	0-30 days		0-365 days	
	stroke	no event	stroke	no event
treatment	33	191	45	179
control	13	214	28	199
Total	46	405	73	378

Figure 1.2: Descriptive statistics for the stent study.

We can compute summary statistics from the table. A summary statistic is a single number summarizing a large amount of data. For instance, the primary results of the study after 1 year could be described by two summary statistics: the proportion of people who had a stroke in the treatment and control groups.

Proportion who had a stroke in the treatment (stent) group: **$45/224 = 0.20 = 20\%$** .

Proportion who had a stroke in the control group: **$28/227 = 0.12 = 12\%$** .

Summary conclusions

These two summary statistics are useful in looking for differences in the groups, and we are in for a surprise: an additional 8% of patients in the treatment group had a stroke! This is important for two reasons. First, it is contrary to what doctors expected, which was that stents would reduce the rate of strokes. Second, it leads to a statistical question: do the data show a “real” difference between the groups?

Is it possible that the 8% difference in the stent study is due to this natural variation. However, the larger the difference we observe (for a particular sample size), the less believable it is that the difference is due to chance.

The summary conclusions of the published analysis are: there is compelling evidence of harm by stents in this study of stroke patients.

The County Data set

We consider data for 3,142 counties in the United States, which includes each county's name, the state where it resides, its population in 2017, how its population changed from 2010 to 2017, poverty rate, and six additional characteristics.

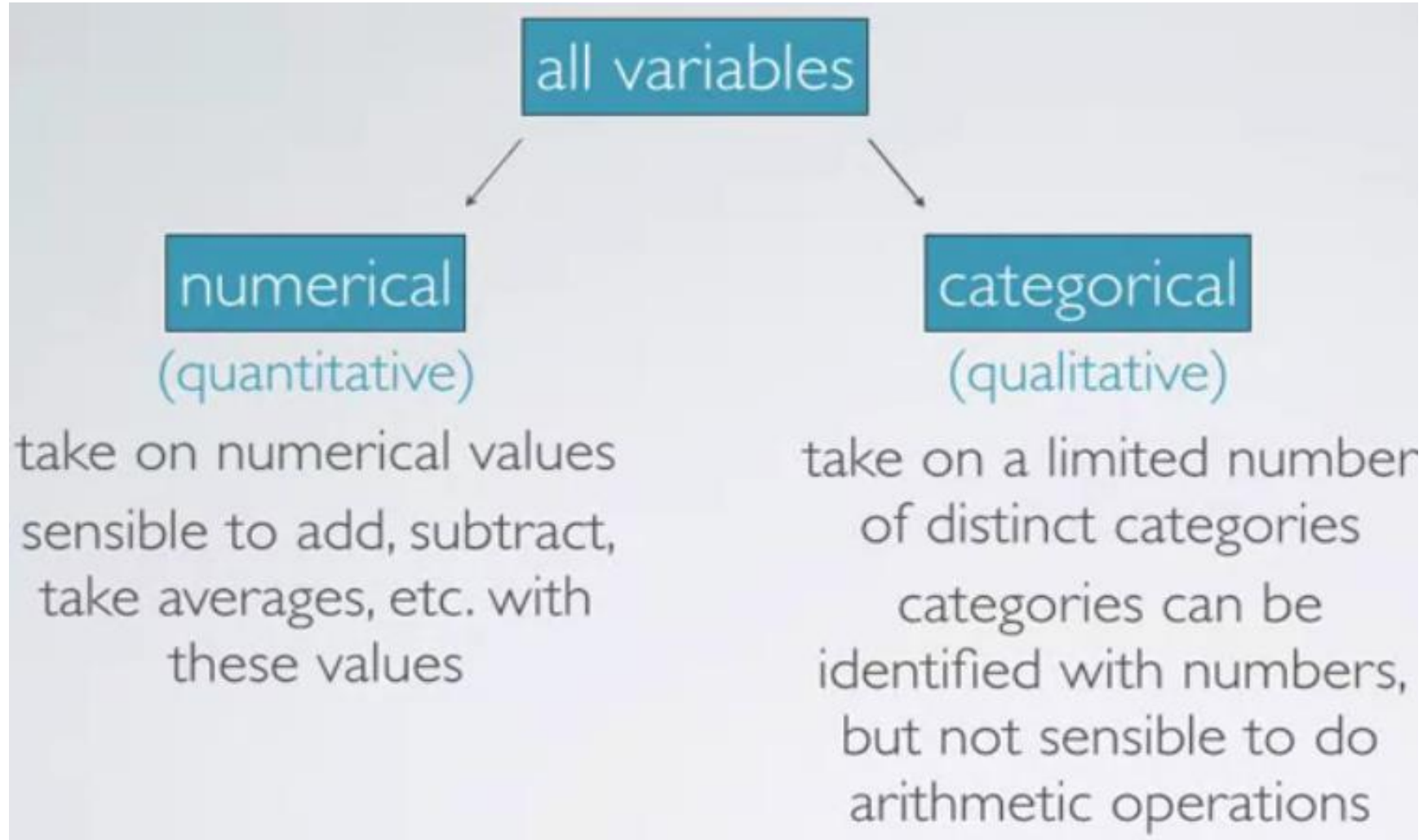
	name	state	pop	pop_change	poverty	homeownership	multi_unit	unemp_rate	metro	median_edu	median_hh_income
1	Autauga	Alabama	55504	1.48	13.7	77.5	7.2	3.86	yes	some_college	55317
2	Baldwin	Alabama	212628	9.19	11.8	76.7	22.6	3.99	yes	some_college	52562
3	Barbour	Alabama	25270	-6.22	27.2	68.0	11.1	5.90	no	hs_diploma	33368
4	Bibb	Alabama	22668	0.73	15.2	82.9	6.6	4.39	yes	hs_diploma	43404
5	Blount	Alabama	58013	0.68	15.6	82.0	3.7	4.02	yes	hs_diploma	47412
6	Bullock	Alabama	10309	-2.28	28.5	76.9	9.9	4.93	no	hs_diploma	29655
7	Butler	Alabama	19825	-2.69	24.4	69.0	13.7	5.49	no	hs_diploma	36326
8	Calhoun	Alabama	114728	-1.51	18.6	70.7	14.3	4.93	yes	some_college	43686
9	Chambers	Alabama	33713	-1.20	18.8	71.4	8.7	4.08	no	hs_diploma	37342
10	Cherokee	Alabama	25857	-0.60	16.1	77.5	4.3	4.05	no	hs_diploma	40041
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
3142	Weston	Wyoming	6927	-2.93	14.4	77.9	6.5	3.98	no	some_college	59605

Fig. Eleven rows of the county data set

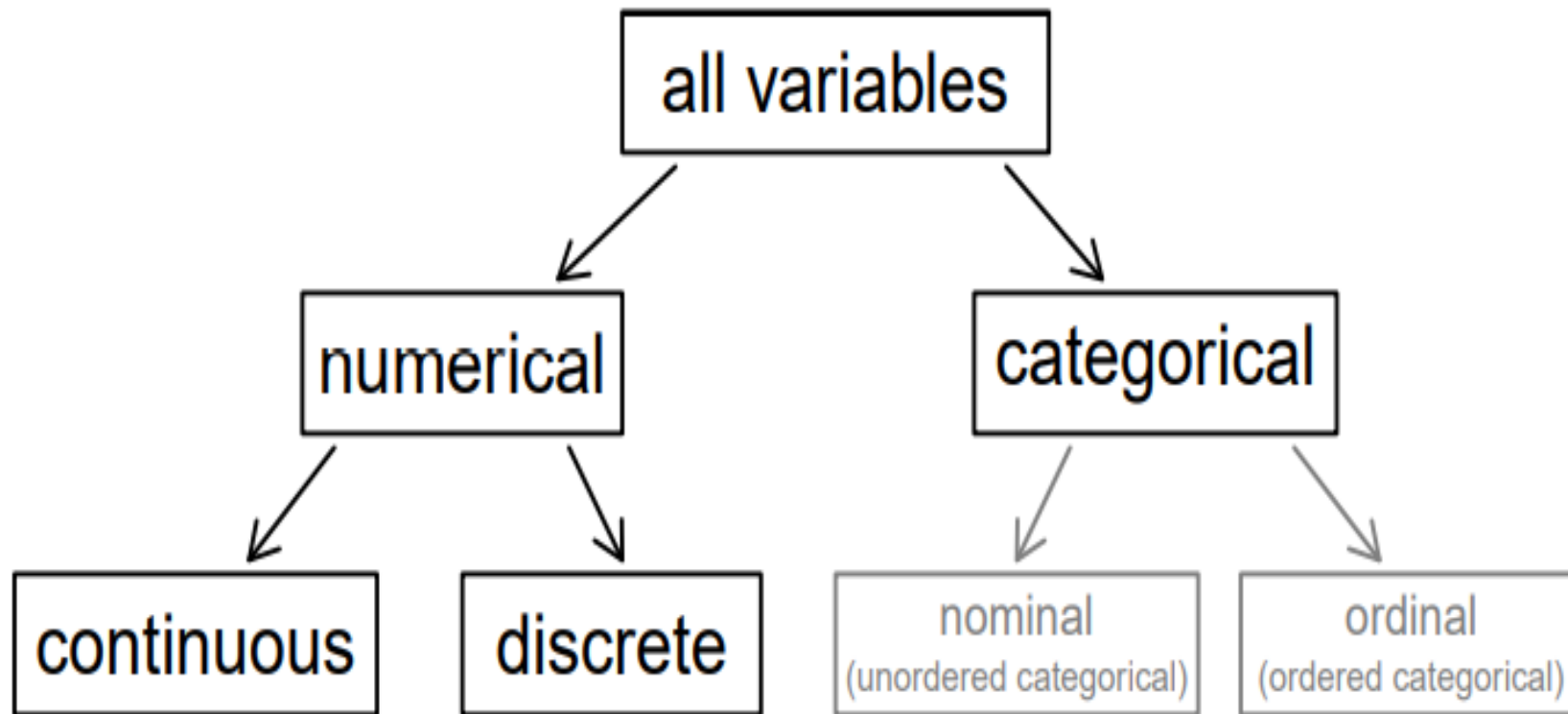
variable	description
name	County name.
state	State where the county resides, or the District of Columbia.
pop	Population in 2017.
pop_change	Percent change in the population from 2010 to 2017. For example, the value 1.48 in the first row means the population for this county increased by 1.48% from 2010 to 2017.
poverty	Percent of the population in poverty.
homeownership	Percent of the population that lives in their own home or lives with the owner, e.g. children living with parents who own the home.
multi_unit	Percent of living units that are in multi-unit structures, e.g. apartments.
unemp_rate	Unemployment rate as a percent.
metro	Whether the county contains a metropolitan area.
median_edu	Median education level, which can take a value among below_hs , hs_diploma , some_college , and bachelors .
median_hh_income	Median household income for the county, where a household's income equals the total income of its occupants who are 15 years or older.

Fig. Variables and their descriptions for the county data set.

Types of Variables



Types of Variables



Types of Variables in County dataset

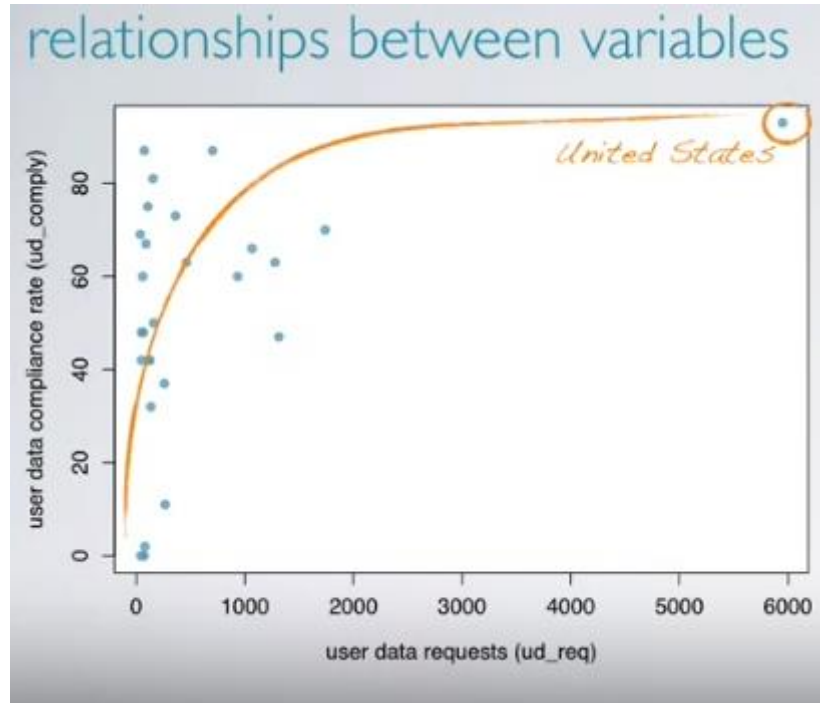
unemp_rate: numeric (continuous)

pop: numeric (whole numbers, no-negative, discrete)

state: categorical, **nominal variable** (possible values AL, AK, ... WY are called the variable's levels)

median_edu: categorical, **ordinal variable** (values below_hs, hs_diploma, some_college, bachelors have a natural ordering)

Relationship 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**

Explanatory and response variables

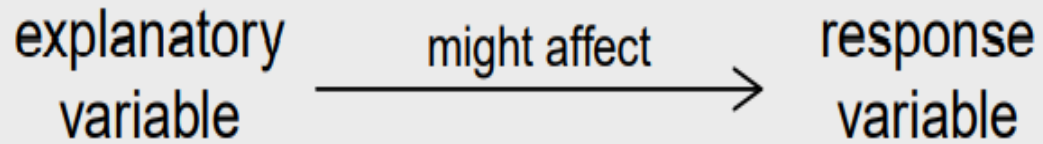
When we ask questions about the relationship between two variables, we sometimes also want to determine if the change in one variable causes a change in the other.

Consider the following question from the county data set:

- ***If there is an increase in the median household income in a county, does this drive an increase in its population?***

EXPLANATORY AND RESPONSE VARIABLES

When we suspect one variable might causally affect another, we label the first variable the explanatory variable and the second the response variable.



In the above example, the median household income is the explanatory variable and the population change is the response variable in the hypothesized relationship

Note: The act of labeling the variables in this way does nothing to guarantee that a causal relationship exists. A formal evaluation to check whether one variable causes a change in another requires an experiment.

Observational studies and experiments

There are two primary types of data collection: observational studies and experiments.

Researchers perform an **observational study** when they collect data in a way that does not directly interfere with how the data arise.

For instance, researchers may collect information via surveys, review medical or company records to form hypotheses about why certain diseases might develop.

In each of these situations, researchers merely observe the data that arise. In general, observational studies can provide evidence of a naturally occurring association between variables, but they cannot by themselves show a causal connection.

When researchers want to investigate the possibility of a causal connection, they conduct an experiment. To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are assigned a treatment. When individuals are randomly assigned to a group, the experiment is called a **randomized experiment**.

ASSOCIATION \neq CAUSATION

In general, association does not imply causation, and causation can only be inferred from a randomized experiment.

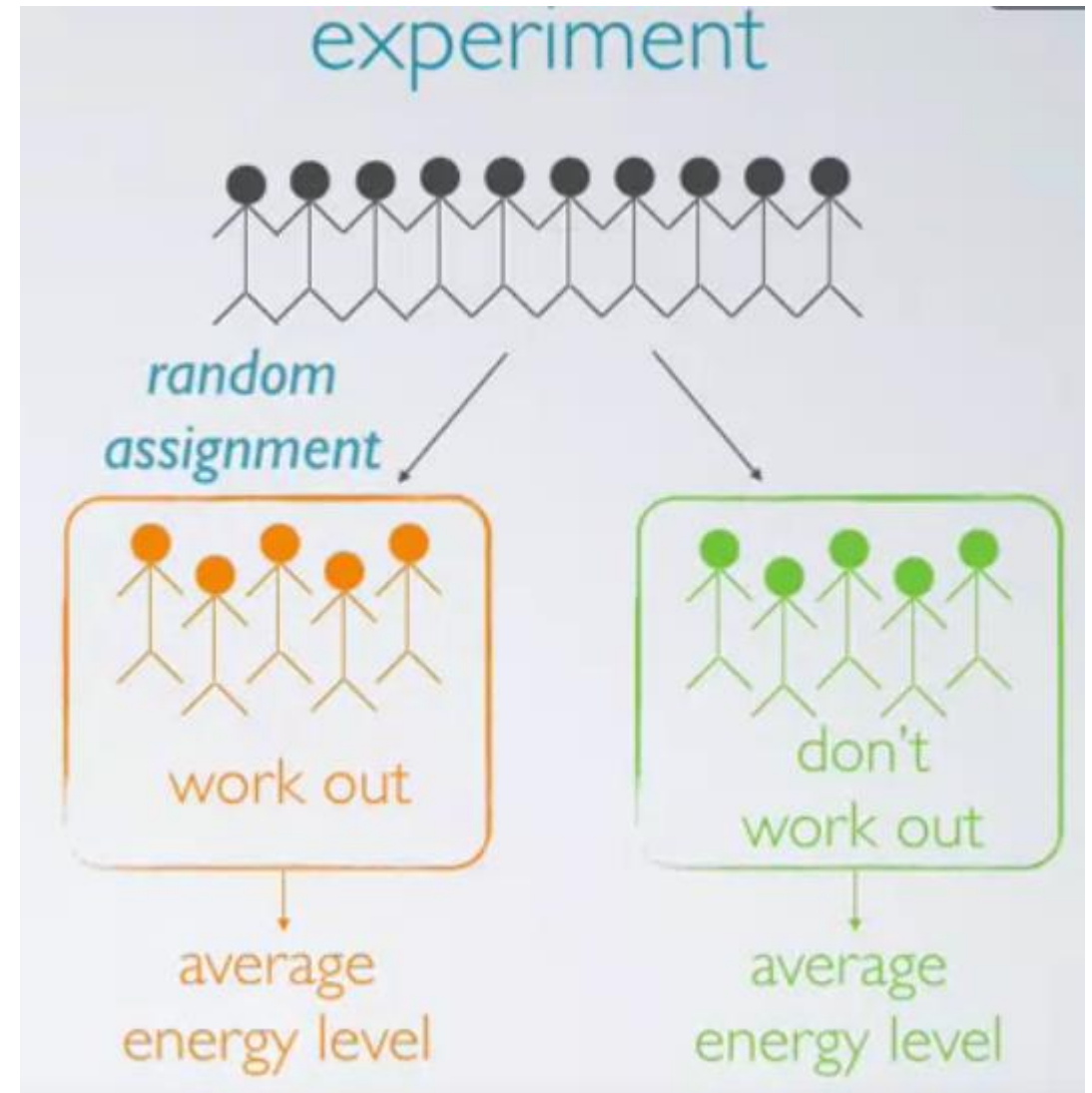
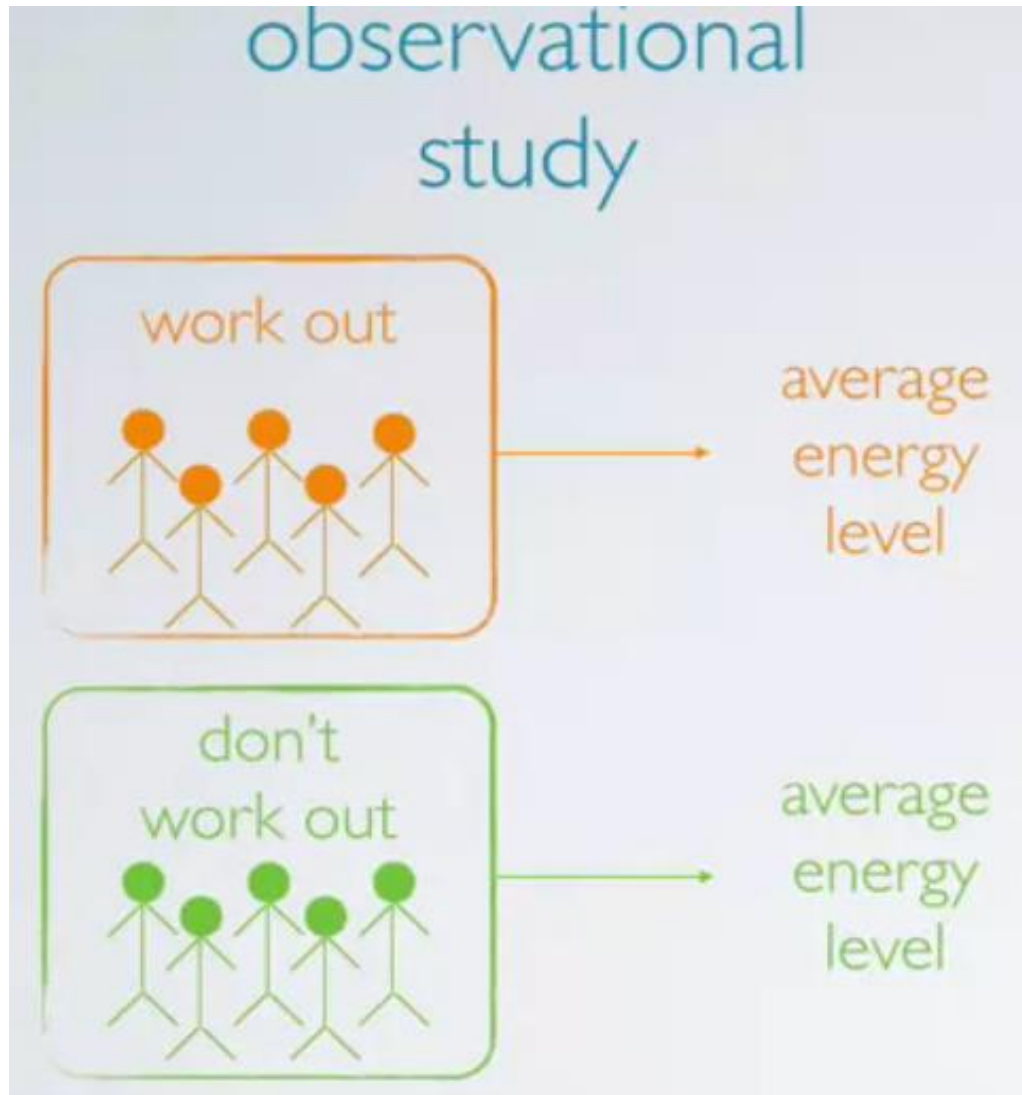
Observational studies and experiments



- ▶ 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

Example



Confounding factors

- If the treatment and control groups have systematic differences other than the treatment, then it might be difficult to identify causality
- Such differences are common in observational studies and referred to as confounding factors

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

Coffee and lung cancer. In the 1960's, studies showed that coffee drinkers had higher rates of lung cancer than those who did not drink coffee. Hence some people identified coffee as a cause of lung cancer.

But coffee does not cause lung cancer. The analysis contained a confounding factor—smoking.

In those days, coffee drinkers were also likely to have been smokers, and smoking does cause lung cancer.

Coffee drinking was associated with lung cancer, but it did not cause the disease.

Experiments

Studies where the researchers assign treatments to cases are called experiments. When this assignment includes randomization, eg. using a coin flip to decide which treatment a patient receives, it is called a randomized experiment. Randomized experiments are fundamentally important when trying to show a causal connection between two variables.

Randomized Controlled Trials (RCT)

- An excellent way to avoid confounding is to assign individuals to the treatment and control groups at random , and then administer the treatment to those who were assigned to the treatment group
- Randomization keeps the two groups similar apart from the treatment.
- Randomized Controlled Experiment, also known as a Randomized Controlled Trial (RCT).
- Such blind experiments using the trial drug/vaccine (for treatment group) and a placebo (for control group) is a standard practice in medical domain, especially in new drug/vaccine development.

In summary

- An observational study is one in which scientists make conclusions based on data that they have observed but had no hand in generating.
- Correlation \neq Causation
- Key to establishing Causality If the treatment and control groups are similar apart from the treatment , then differences between the outcomes in the two groups can be ascribed to the treatment.
- If the treatment and control groups have systematic differences other than the treatment, then it might be difficult to identify causality Confounding factors
- Randomized Controlled Experiment , also known as a Randomized Controlled Trial (RCT) is a scientific experiment aims to reduce the bias due to confounding factors

Sampling and sources of bias

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 [Share](#)

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.

Sampling and sources of 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

Populations and samples

Consider the research question: Does a new drug reduce the number of deaths in patients with severe heart disease?

The research question refers to a **target population**. The population includes all people with severe heart disease. A person with severe heart disease represents a case.

Often times, it is too expensive to collect data for every case in a population. Instead, a **sample** is taken. A sample represents a subset of the cases and is often a small fraction of the population. For instance, 50 people with severe heart disease (or some other number) in the population might be selected, and this sample data may be used to provide an estimate of the population average and answer the research question.

To investigate the possibility of a causal connection, the researchers conduct an experiment. Usually there will be both an explanatory and a response variable.

For instance, we may suspect administering a drug will reduce mortality in heart attack patients over the following year.

To check if there really is a causal connection between the explanatory variable and the response, researchers will collect a sample of individuals and split them into groups. The individuals in each group are assigned a treatment.

When individuals are randomly assigned to a group, the experiment is called a randomized experiment.

For example, each heart attack patient in the drug trial could be randomly assigned, perhaps by flipping a coin, into one of two groups:

the first group receives a placebo (fake treatment) and the second group receives the drug.

Anecdotal evidence

Consider the following possible responses to the research question: “My friend’s dad had a heart attack and died after they gave him a new heart disease drug, so the drug must not work”.

The conclusion is based on data. However, there are two problems. First, the data only represent one or two cases. Second, and more importantly, it is unclear whether these cases are actually representative of the population. Data collected in this haphazard fashion are called **anecdotal evidence**.

Experimental Design – terminology

placebo fake treatment, often used as the control group for medical studies	placebo effect showing change despite being on the placebo
blinding experimental units don't know which group they're in	double-blind both the experimental units and the researchers don't know the group assignment

Experimental Design – terminology

- Placebo: fake treatment, often used as the control group for medical studies
- Placebo effect: experimental units showing improvement simply because they believe they are receiving a special treatment
- Blinding: when experimental units do not know whether they are in the control or treatment group
- Double-blind: when both the experimental units and the researchers who interact with the patients do not know who is in the control and who is in the treatment group

Sampling methods

In a statistical study, sampling methods refer to how we select members from the population to be in the study.

If a sample isn't randomly selected, it will probably be biased in some way and the data may not be representative of the population.

There are many ways to select a sample—some good and some bad.

Bad ways to sample

Convenience sample: The researcher chooses a sample that is readily available in some non-random way.

- Example—A researcher polls people as they walk by on the street.

Why it's probably biased: The location and time of day and other factors may produce a biased sample of people.

Voluntary response sample: The researcher puts out a request for members of a population to join the sample, and people decide whether or not to be in the sample.

- Example—A TV show host asks his viewers to visit his website and respond to an online poll.

Why it's probably biased: People who take the time to respond tend to have similarly strong opinions compared to the rest of the population.

Good ways to sample

Simple random sample: Every member and set of members has an equal chance of being included in the sample.

Technology, random number generators, or some other sort of chance process is needed to get a simple random sample.

- Example—A teachers puts students' names in a hat and chooses without looking to get a sample of students.

Why it's good: Random samples are usually fairly representative since they don't favor certain members.

Stratified random sample: The population is first split into groups. The overall sample consists of some members from every group. The members from each group are chosen randomly.

Example—A student council surveys 100 students by getting random samples of 25 freshmen, 25 sophomores, 25 juniors, and 25 seniors.

Why it's good: A stratified sample guarantees that members from each group will be represented in the sample, so this sampling method is good when we want some members from every group.

Cluster random sample: The population is first split into **groups**. The overall sample consists of **every member** from some of the groups. **The groups are selected at random.**

- Example—An airline company wants to survey its customers one day, so they randomly select 5 flights that day and survey every passenger on those flights.

Why it's good: A cluster sample gets every member from some of the groups, so it's good when each group reflects the population as a whole.

A **multistage sample** is like a cluster sample, but rather than keeping all observations in each cluster, we collect a random sample within each selected cluster.

Systematic random sample: Members of the population are put in some order. A starting point is selected at random, and every n th member is selected to be in the sample.

- Example—A principal takes an alphabetized list of student names and picks a random starting point. Every 20th student is selected to take a survey.

Example

Suppose we are interested in estimating the malaria rate in a densely tropical portion of rural Indonesia. We learn that there are 30 villages in that part of the Indonesian jungle, each more or less similar to the next. Our goal is to test 150 individuals for malaria. What sampling method should be employed?

A simple random sample would likely draw individuals from all 30 villages, which could make data collection extremely expensive. Stratified sampling would be a challenge since it is unclear how we would build strata of similar individuals.

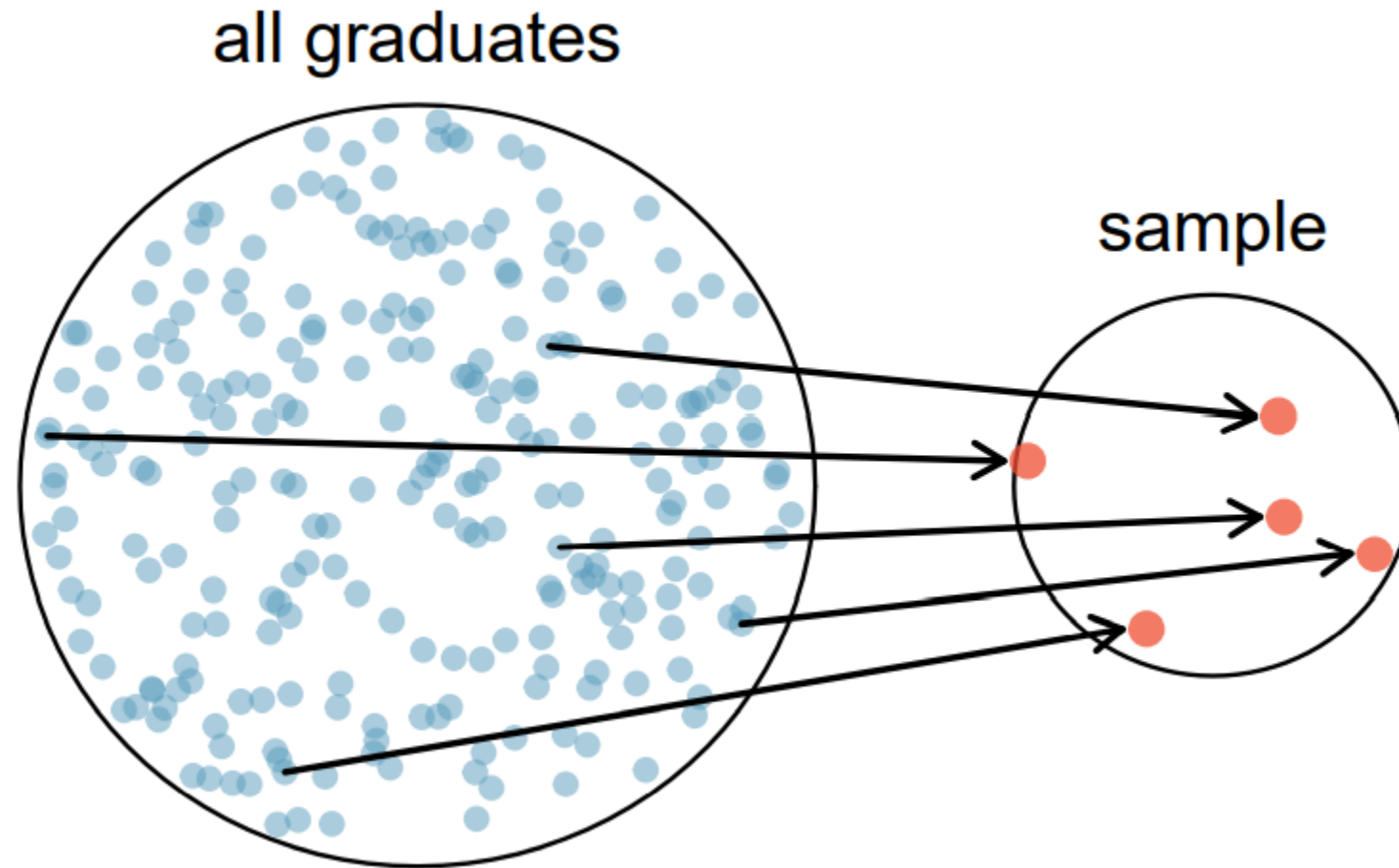
However, cluster sampling or multistage sampling seem like very good ideas. If we decided to use multistage sampling, we might randomly select half of the villages, then randomly select 10 people from each. This would probably reduce our data collection costs substantially in comparison to a simple random sample, and the cluster sample would still give us reliable information

Example: Sampling from a population

We might try to estimate the time to graduation for Duke undergraduates in the last 5 years by collecting a sample of students.

All graduates in the last 5 years represent the population, and graduates who are selected for review are collectively called the sample. In general, we always seek to randomly select a sample from a population. The most basic type of random selection is equivalent to how raffles are conducted.

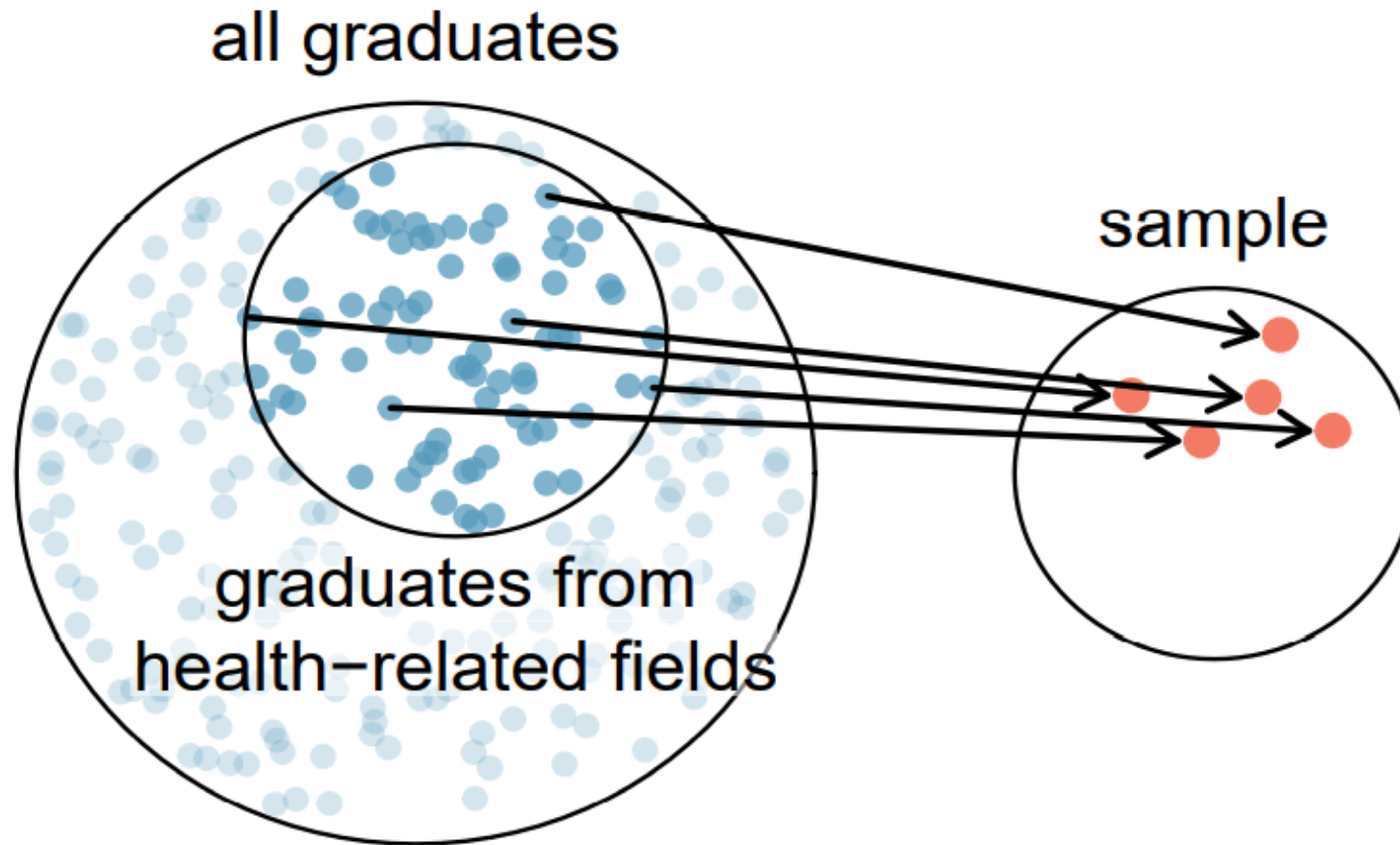
For example, in selecting graduates, we could write each graduate's name on a raffle ticket and draw 100 tickets. The selected names would represent a random sample of 100 graduates. We pick samples randomly to reduce the chance we introduce biases.



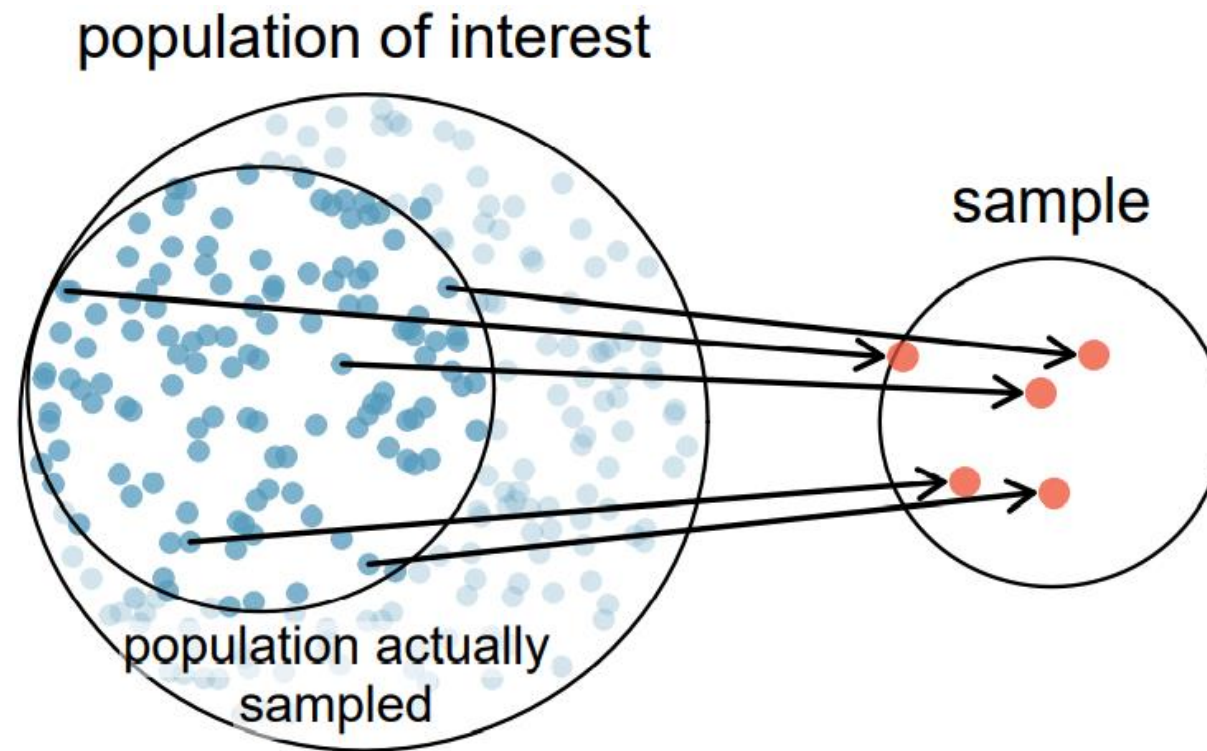
In the figure, five graduates are randomly selected from the population to be included in the sample.

Examples of Biased Samples

1) Asked to pick a sample of graduates, a nutrition major might inadvertently pick a disproportionate number of graduates from health-related majors.



2) Even when people are picked at random, eg. for surveys, caution must be exercised if the non-response rate is high. For instance, if only 30% of the people randomly sampled for a survey actually respond, then the results may not be representative of the entire population. This **non-response bias** can skew results.



3) Another common downfall is a **convenience sample**, where individuals who are easily accessible are more likely to be included in the sample. For instance, if a political survey is done by stopping people walking in the Bronx, this will not represent all of New York City. It is often difficult to discern what sub-population a convenience sample represents.

Summarizing data

The Figure below displays rows 1, 2, 3, and 50 of a data set for 50 randomly sampled loans offered through Lending Club, which is a peer-to-peer lending company. These observations will be referred to as the **loan50 data set**. Each row in the table represents a single loan. The columns represent characteristics, called variables, for each of the loans.

	loan_amount	interest_rate	term	grade	state	total_income	homeownership
1	7500	7.34	36	A	MD	70000	rent
2	25000	9.43	60	B	OH	254000	mortgage
3	14500	6.08	36	A	MO	80000	mortgage
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	3000	7.96	36	A	CA	34000	rent

variable	description
loan_amount	Amount of the loan received, in US dollars.
interest_rate	Interest rate on the loan, in an annual percentage.
term	The length of the loan, which is always set as a whole number of months.
grade	Loan grade, which takes a values A through G and represents the quality of the loan and its likelihood of being repaid.
state	US state where the borrower resides.
total_income	Borrower's total income, including any second income, in US dollars.
homeownership	Indicates whether the person owns, owns but has a mortgage, or rents.

Examining numerical data

Here we will explore techniques for summarizing numerical variables. For example, consider the loan amount variable from the **loan50 data set**, which represents the loan size for all 50 loans in the data set.

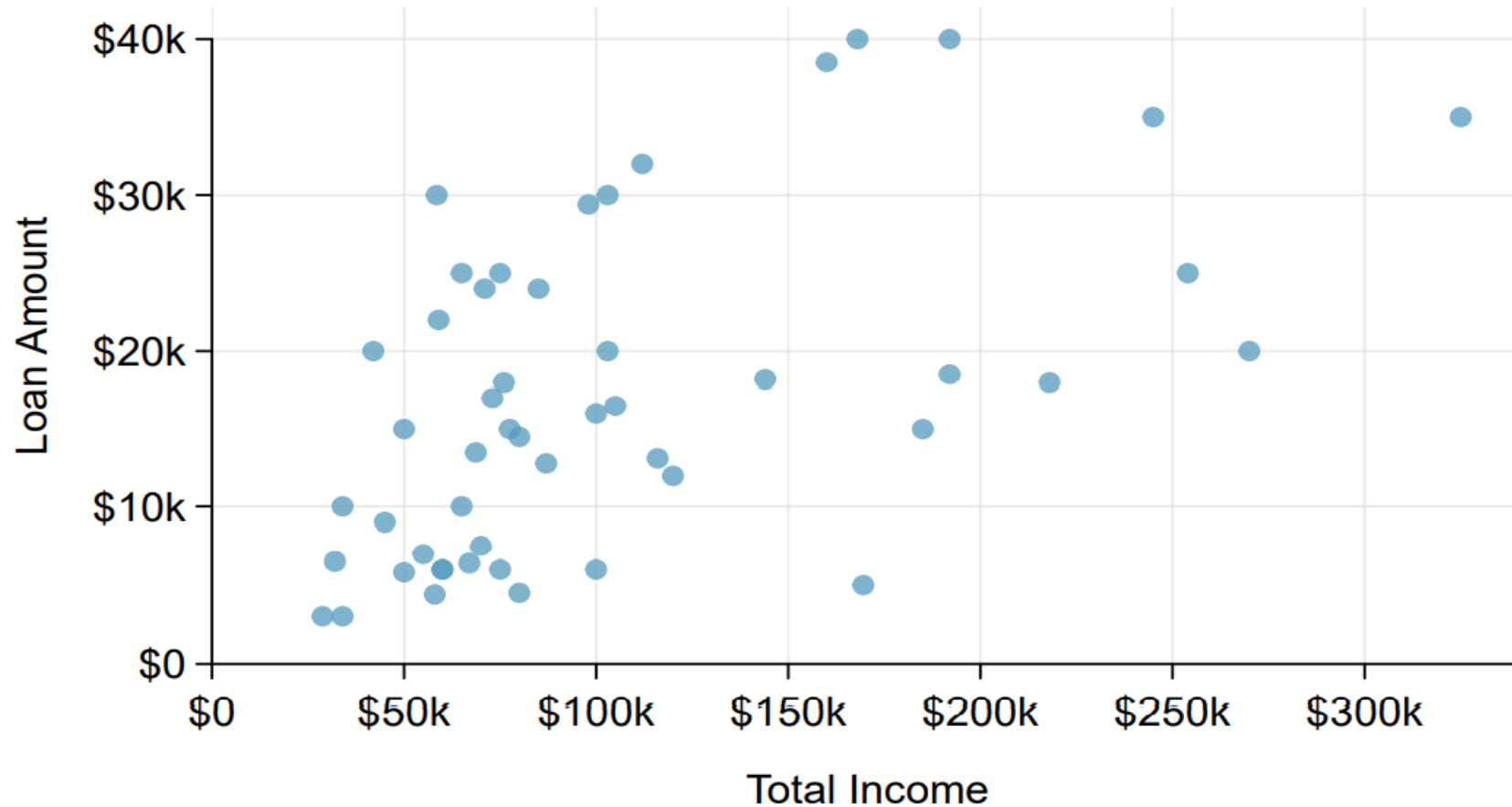
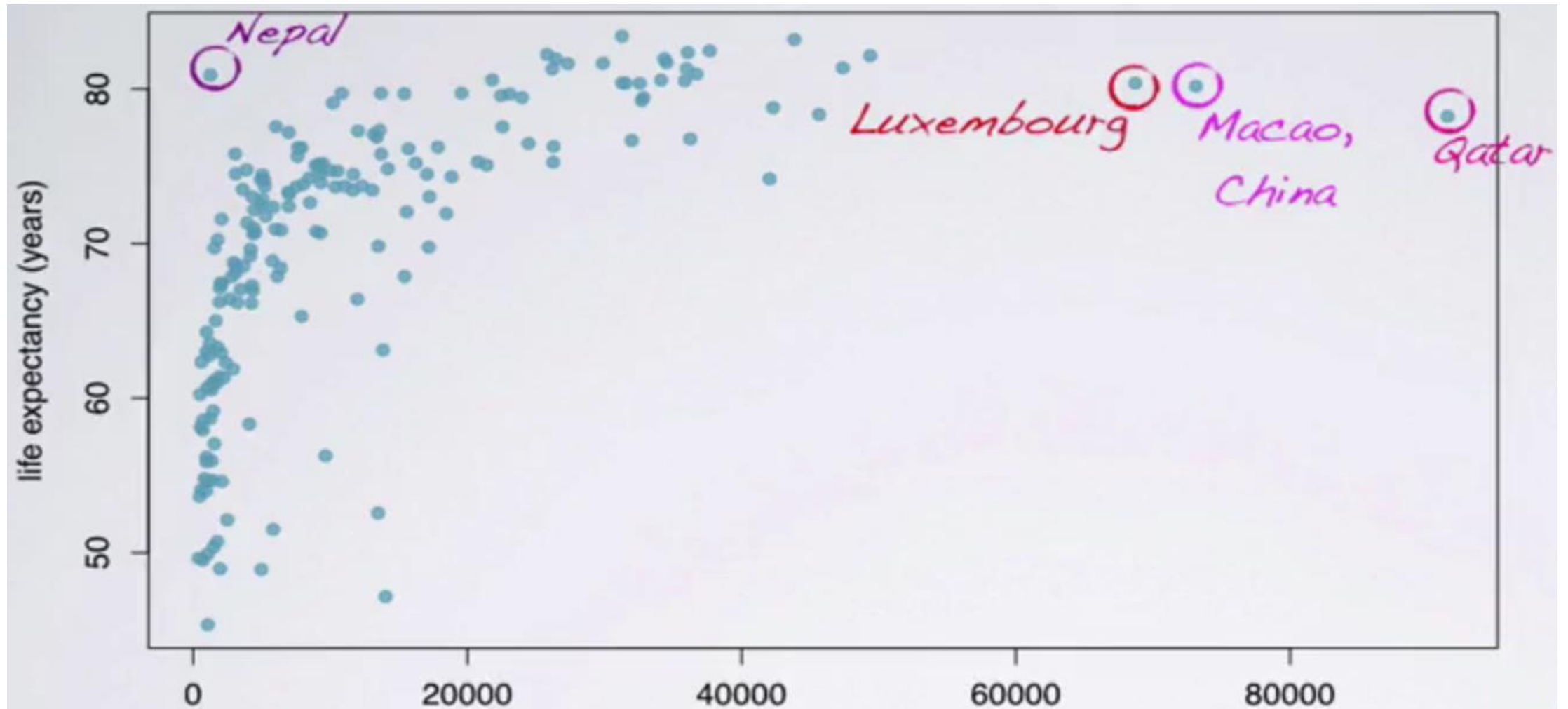


Fig. A scatterplot of total income versus loan amount for the loan50 data set.

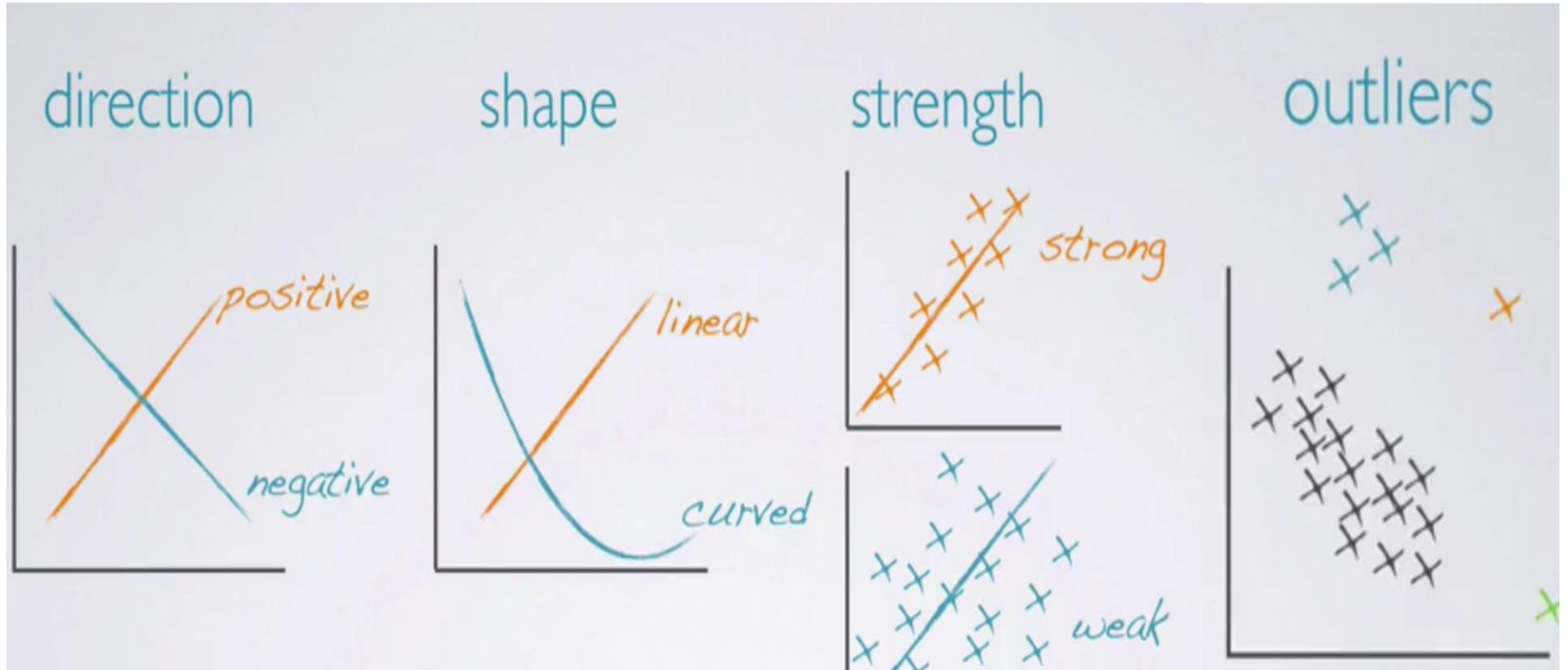
Visualizing numerical data

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



Income per person (GDP/capita PPP\$ inflation-adjusted)

Evaluate the relationship



Histograms and shape

With larger data samples rather than showing the value of each observation, we prefer to think of the value as belonging to a bin. For example, in the **loan50 data set**, we create a table of counts for the number of loans with interest rates between 5.0% and 7.5%, then the number of loans with rates between 7.5% and 10.0%, and so on.

Interest Rate	5.0% - 7.5%	7.5% - 10.0%	10.0% - 12.5%	12.5% - 15.0%	...	25.0% - 27.5%
Count	11	15	8	4	...	1

Figure: Counts for the binned interest rate data.

Histogram

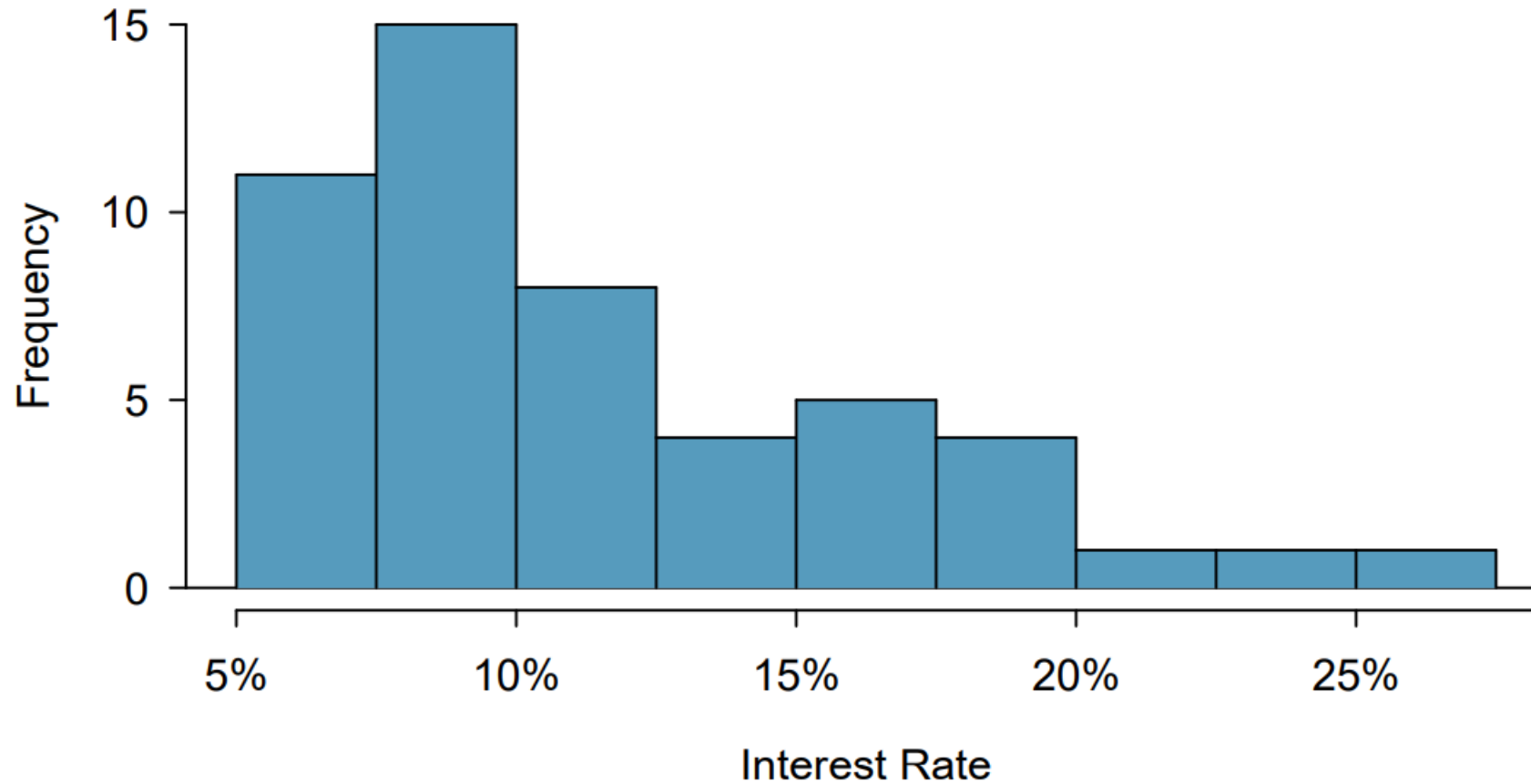


Figure: A histogram of interest rate. This distribution is strongly skewed to the right.

Histograms provide a view of the data density. Higher bars represent where the data are relatively more common.

For instance, there are many more loans with rates between 5% and 10% than loans with rates between 20% and 25% in the data set.

The bars make it easy to see how the density of the data changes relative to the interest rate. Histograms are especially convenient for understanding the shape of the data distribution. The figure suggests that most loans have rates under 15%, while only a handful of loans have rates above 20%.

When data trail off to the right in this way and has a longer right tail, the shape is said to be **right skewed**.

Data sets with the reverse characteristic – a long, thinner tail to the left – are said to be **left skewed**. Data sets that show roughly equal trailing off in both directions are called **symmetric**.

In addition to looking at whether a distribution is skewed or symmetric, histograms can be used to identify modes. A mode defined as the value with the most occurrences in the data set.

A mode is represented by a prominent peak in the distribution. For example, there is only one prominent peak in the histogram of interest rate, and hence the distribution is **unimodal**.

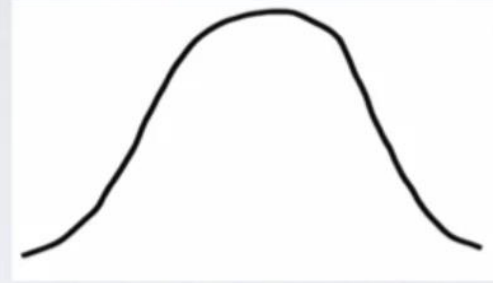
Shapes of numerical Distribution

shape

skewness



left skewed



symmetric



right skewed

modality



unimodal



bimodal

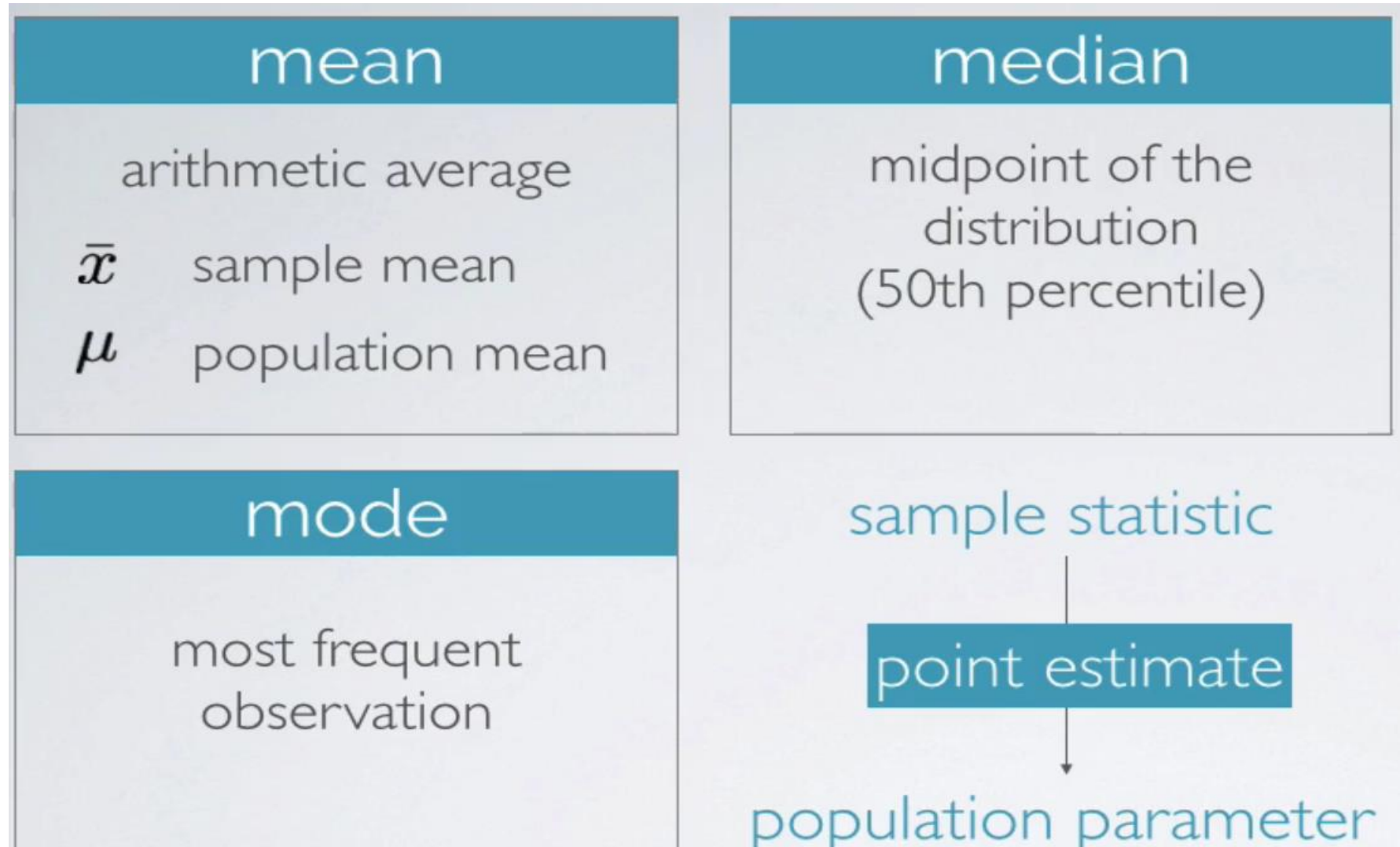


uniform



multimodal

Measures of Center



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


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

Elements of a Boxplot

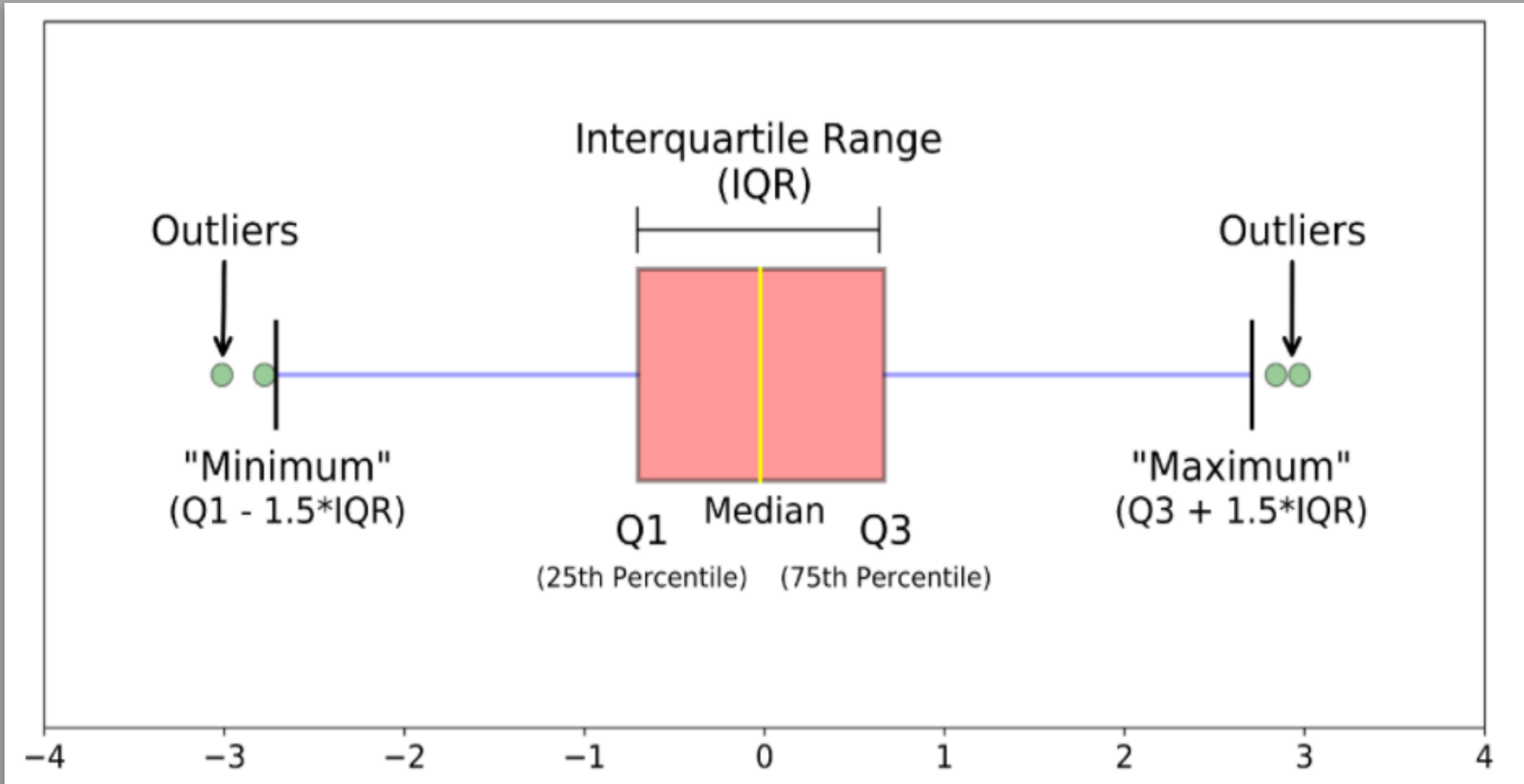
A boxplot is a standardized way of displaying the dataset based on a **five-number summary**: the minimum, the maximum, the sample median, and the first and third quartiles.

- **Minimum** (Q_0 or 0th **percentile**): the lowest data point excluding any outliers.
- **Maximum** (Q_4 or 100th percentile): the largest data point excluding any outliers.
- **Median** (Q_2 or 50th percentile): the middle value of the dataset.
- **First quartile** (Q_1 or 25th percentile): also known as the *lower quartile* $q_n(0.25)$, is the median of the lower half of the dataset.
- **Third quartile** (Q_3 or 75th percentile): also known as the *upper quartile* $q_n(0.75)$, is the median of the upper half of the dataset.^[4]

- **Interquartile range (IQR)** : is the distance between the upper and lower quartiles.

$$\text{IQR} = Q_3 - Q_1 = q_n(0.75) - q_n(0.25)$$

Box plot

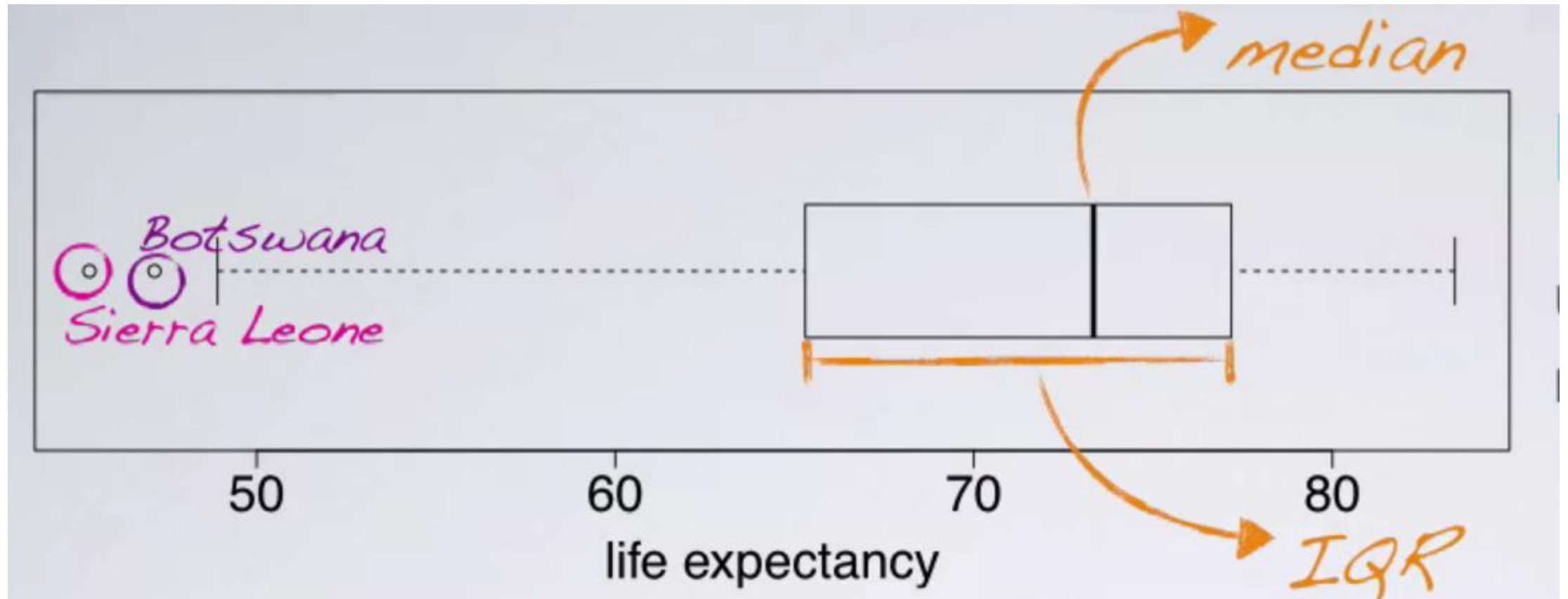


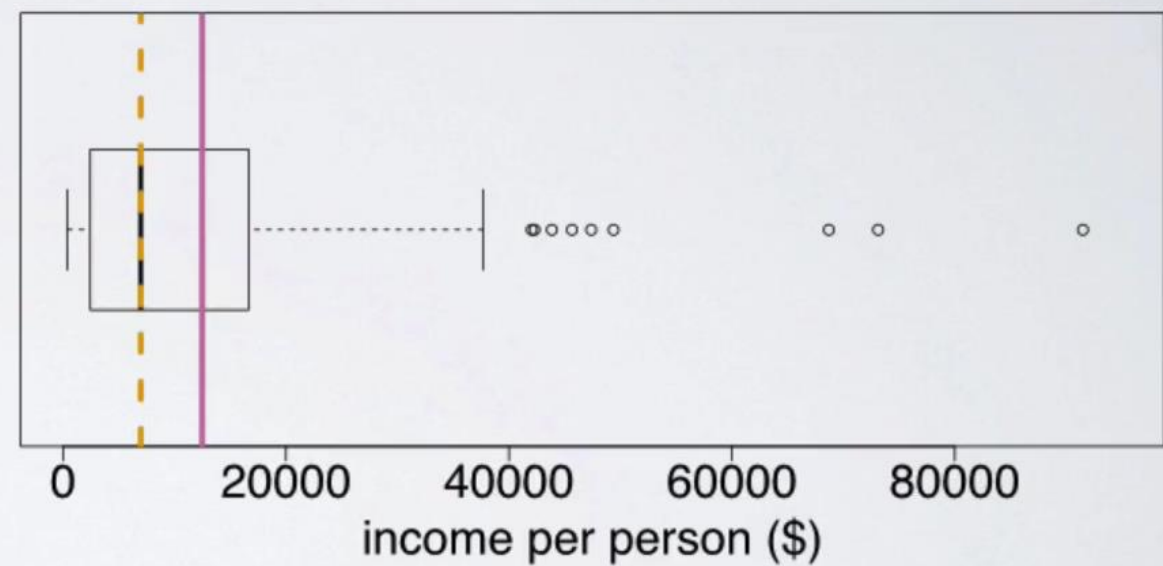
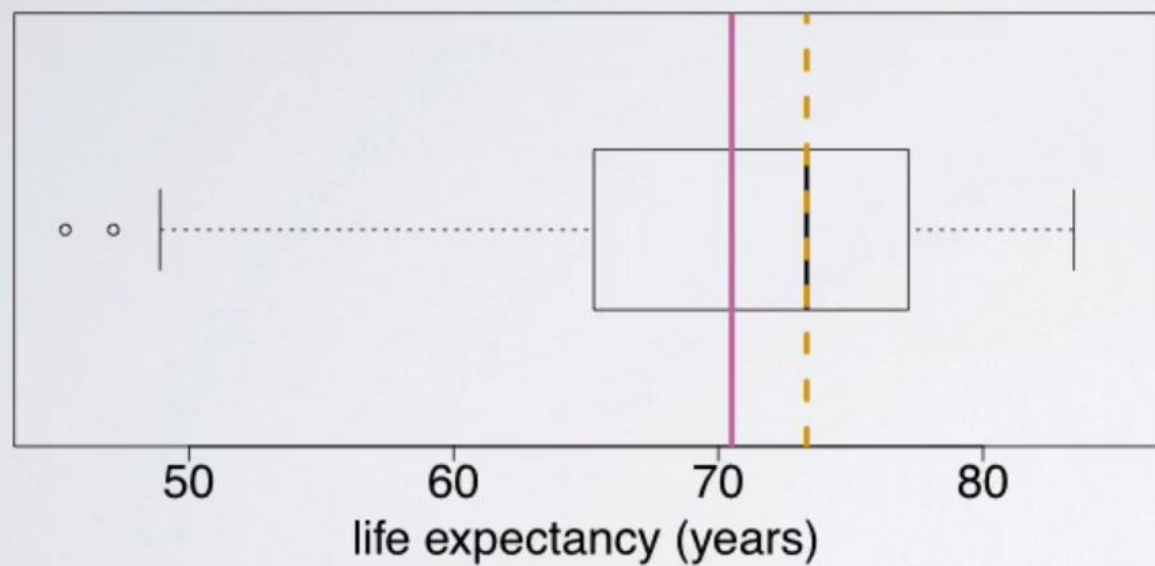
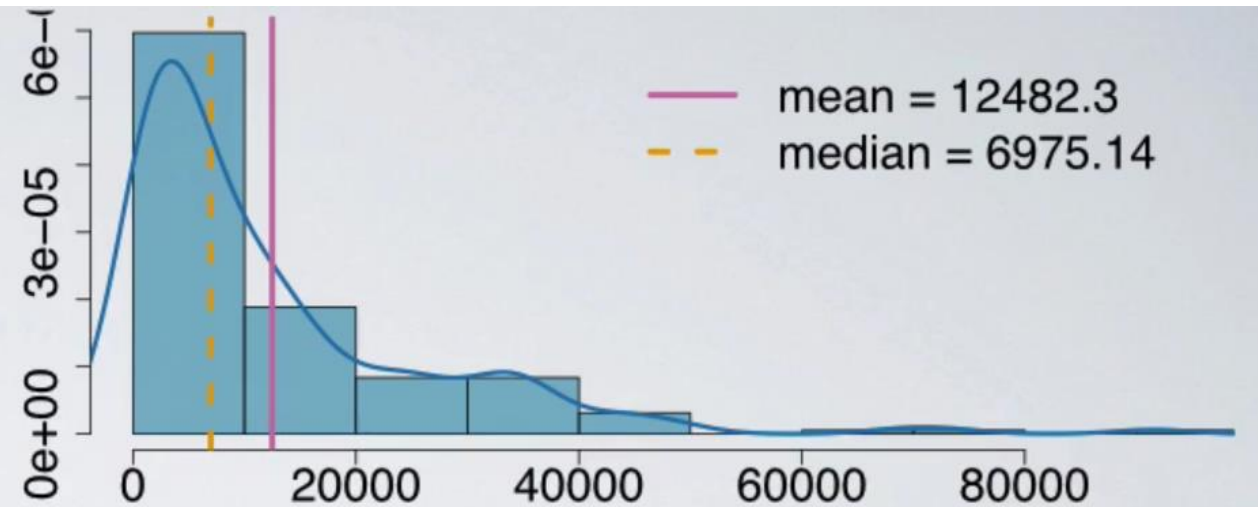
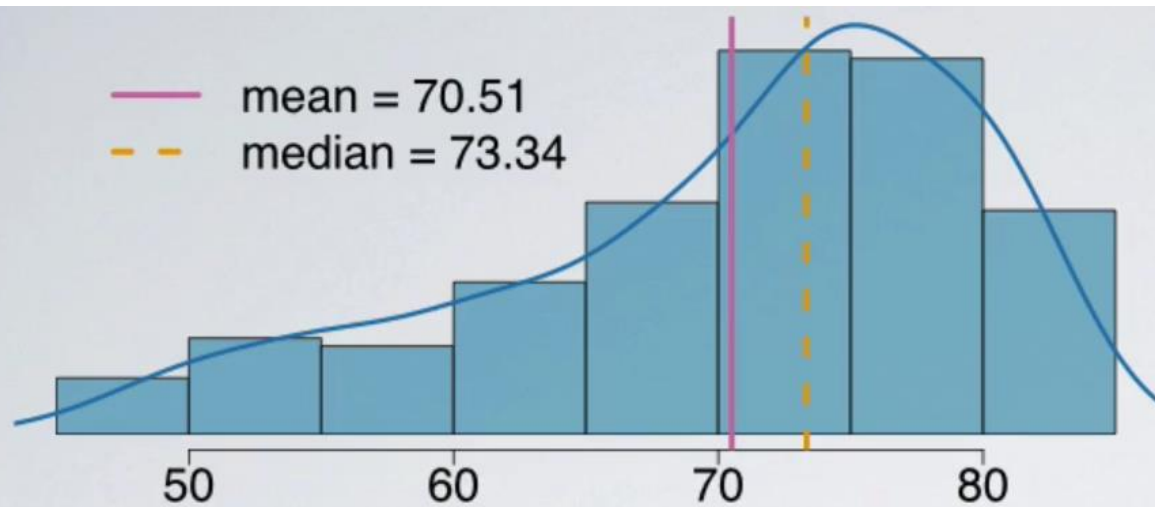
Example: Gapminder dataset

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

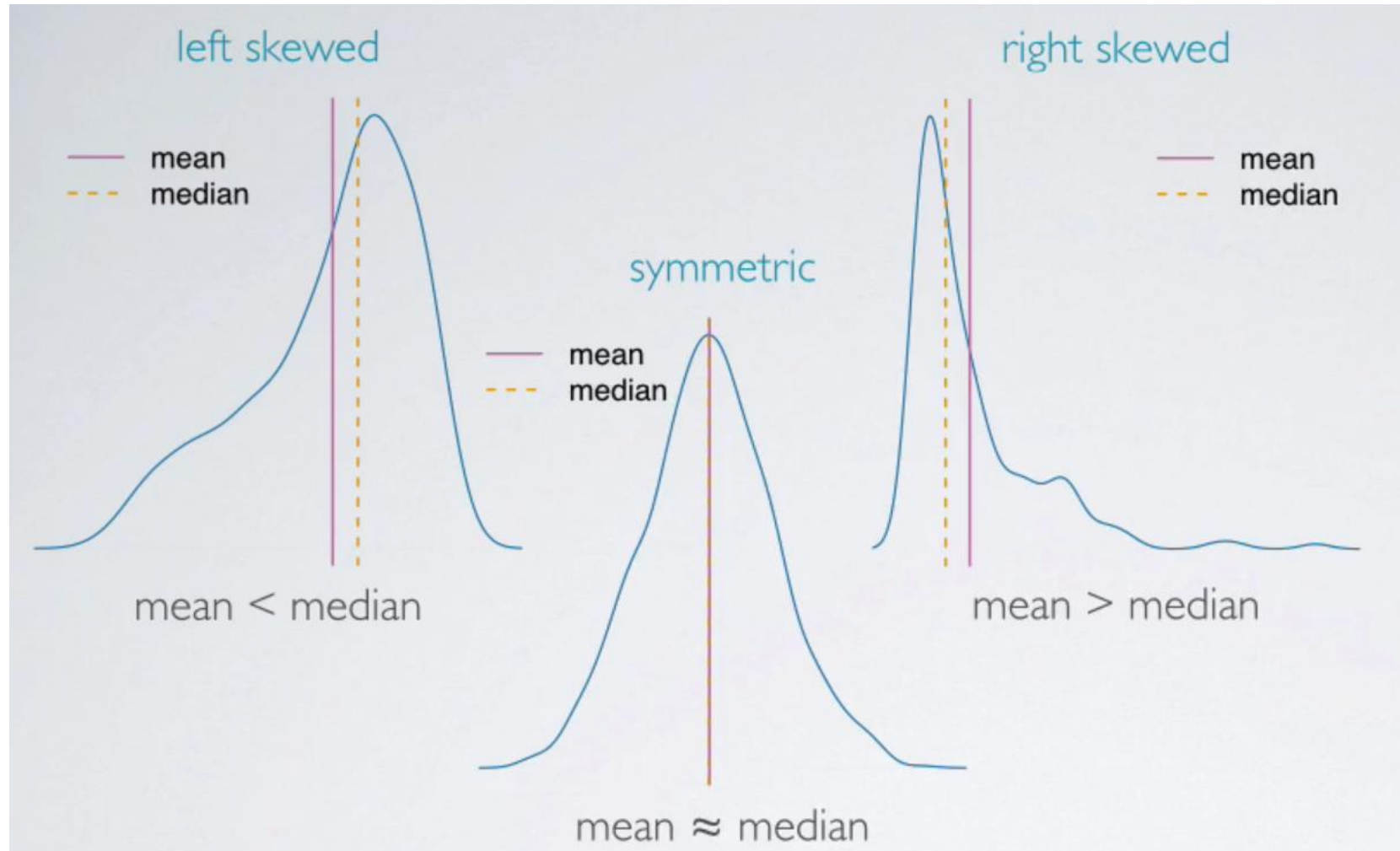
Source: gapminder.com

Box plot example



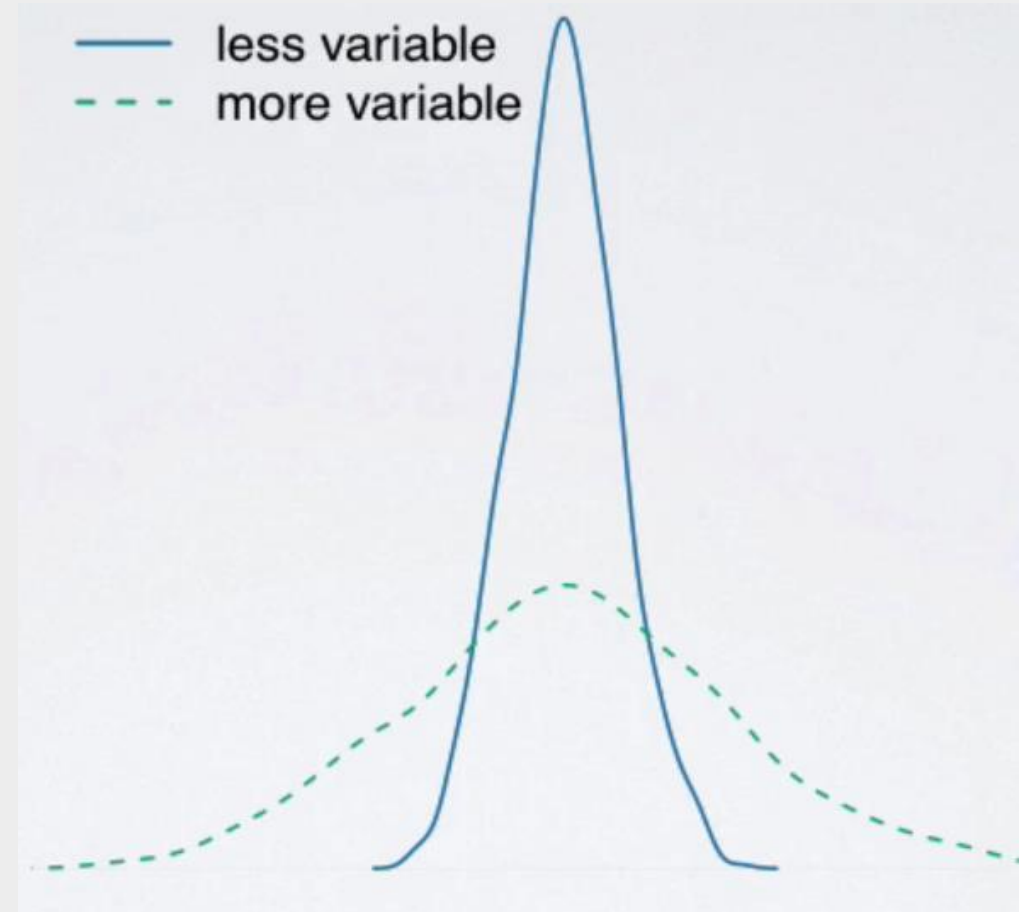


Skewness vs. Measures of center



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:

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

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

201	Zimbabwe	58.1

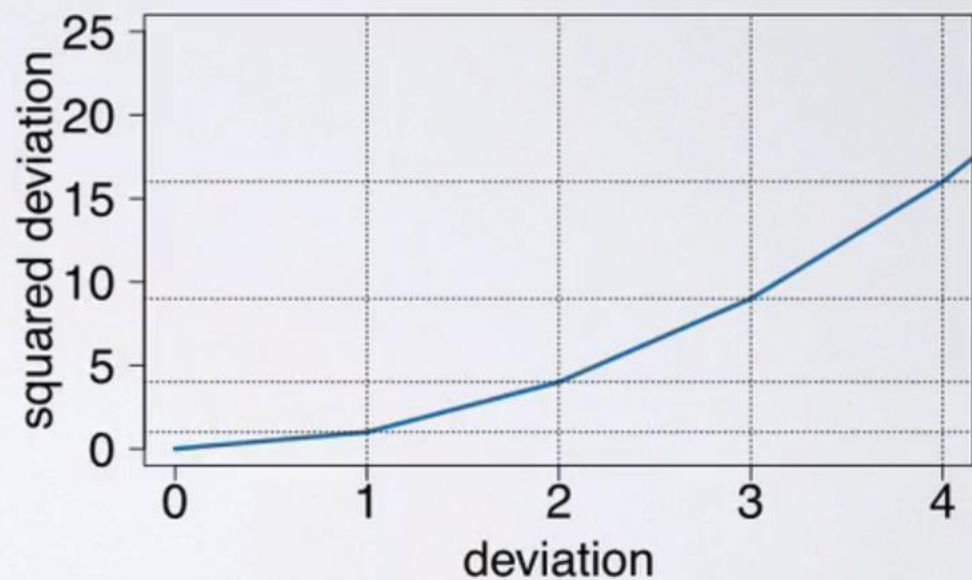
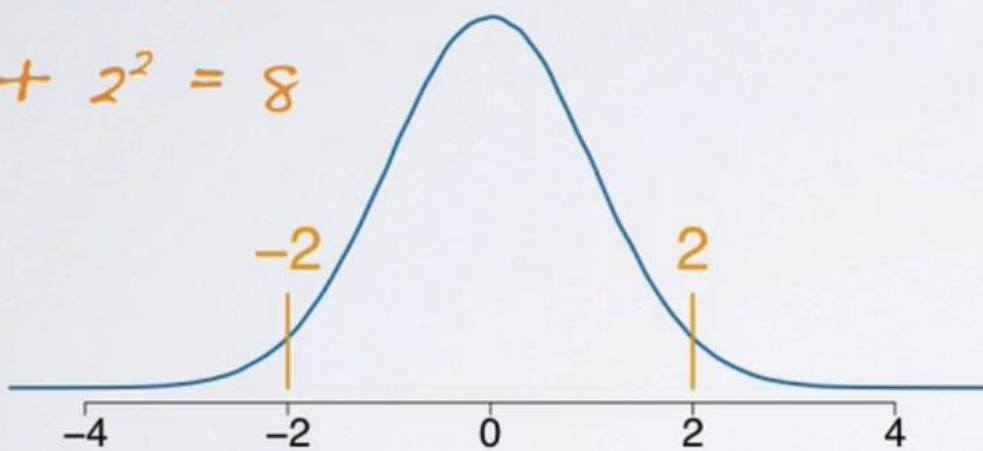
Why do we square the differences?

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

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

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

$$(-2) + 2 = 0$$
$$(-2)^2 + 2^2 = 8$$



Standard Deviation

standard deviation

sample sd
 s
population sd
 σ

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

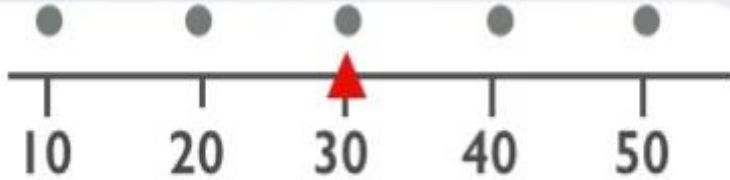
$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1}}$$

square root of the variance

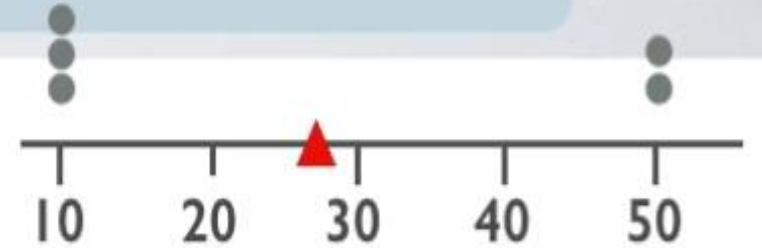
Example: Variability

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

☐ SET 1



☒ SET 2

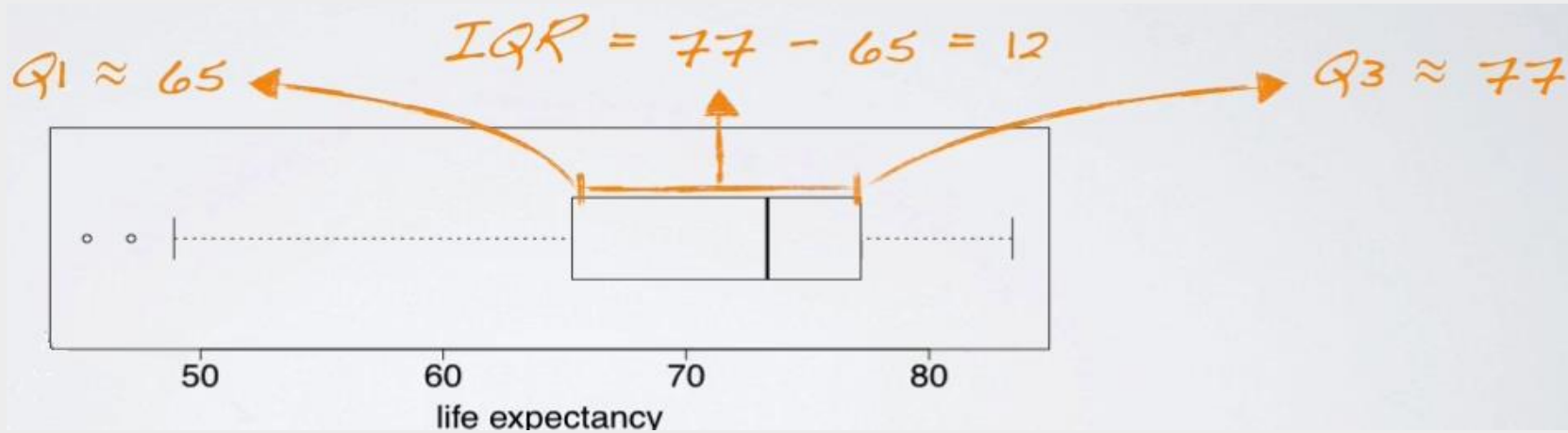


Interquartile Range

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

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

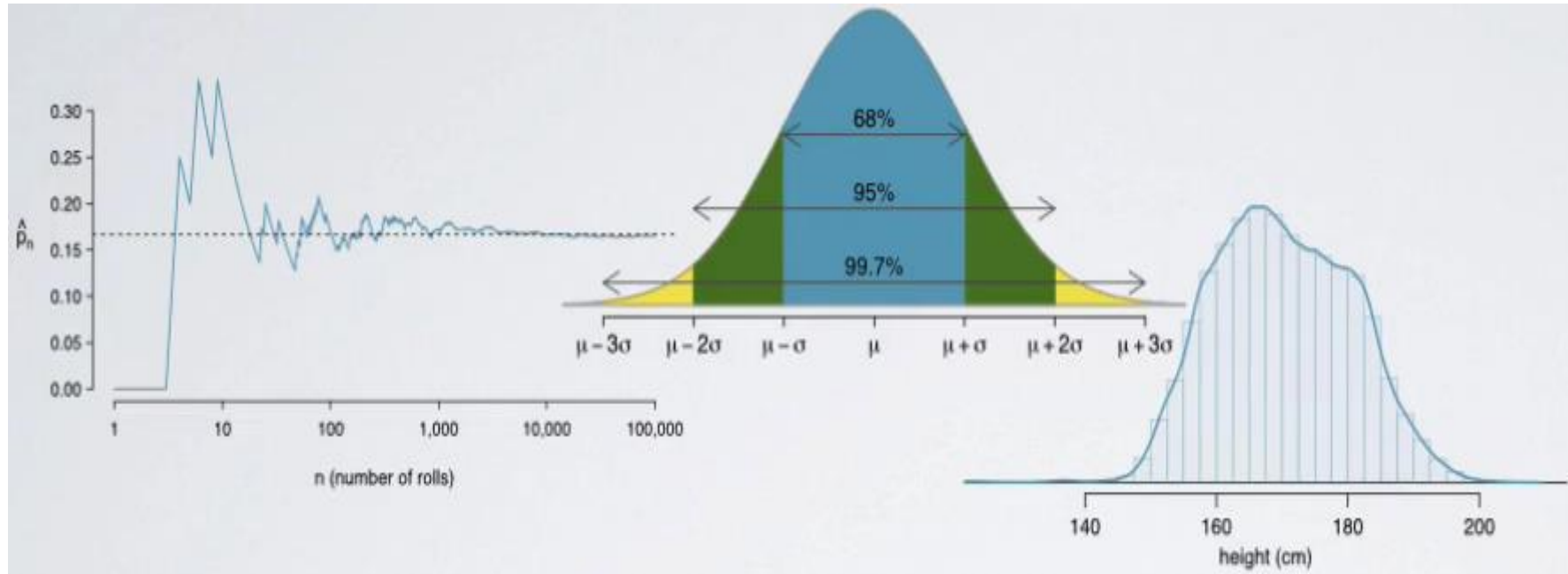
	robust	non-robust
center	median	mean
spread	IQR	SD, range

*skewed,
with extreme*

symmetric

- The median and IQR are called robust statistics because extreme observations have little effect on their values: moving the most extreme value generally has little influence on these statistics.
- On the other hand, the mean and standard deviation are more heavily influenced by changes in extreme observations, which can be important in some situations.

Probability and Distributions



1. Important in inferential statistics, a branch of statistics that relies on sample information to make decisions about a population.
2. Used to make decisions in the face of uncertainty.

Random process

In a **random process** we know what outcomes could happen, but we don't know which particular outcome will happen.



Characterizing random phenomena

Sources of errors in observed outcomes

- Lack of knowledge of generating process (model error)

We may not know all the rules that govern the data generating process, i.e., we may not know all the laws, and the knowledge of all the causes that affects the outcomes. This is called modeling error.

- Errors in sensors used for measuring outcomes (**measurement error**)

Even if we know everything the sensor that we use for observing these outcomes may themselves contain errors. Such errors are called measurement errors.

- These two errors are modeled using probability density functions and therefore, the outcomes are also predicted with certain confidence intervals, which we derive.

Characterizing random phenomena

- **Deterministic phenomena:** Phenomena whose outcome can be predicted with the high level of confidence can be considered as deterministic.
 - Ex. **the conversion between Celsius and Kelvin** is deterministic, and It is an exact formula that will always give you the correct answer (assuming you perform the calculations correctly):
 $\text{Kelvin} = \text{Celsius} + 273.15$
- **Stochastic phenomena:** phenomena which can have many possible outcomes for the same experimental conditions. The outcomes can be predicted with limited confidence.
 - Ex: **Outcome of a coin toss**

Types of random phenomena

Types of random phenomena

- **Discrete:** Outcomes are finite
 - Coin toss : $\{H, T\}$
 - Throw of a dice : $\{1, 2, 3, 4, 5, 6\}$
- **Continuous:** infinite number of outcomes
 - Body temperature measurement in deg Celsius

probability

$P(A) =$
Probability
of event A

There are several possible interpretations of probability but they (almost) completely agree on the mathematical rules probability must follow:

$$0 \leq P(A) \leq 1$$

frequentist interpretation

The probability of an outcome is the proportion of times the outcome would occur if we observed the random process an infinite number of times.

bayesian interpretation

A Bayesian interprets probability as a subjective degree of belief.

Largely popularized by revolutionary advance in computational technology and methods during the last twenty years.

Photo by dahlstroms on Flickr (<http://www.flickr.com/photos/dahlstroms/527634847/>)

Sample space, Event

Consider a random experiment. The set of all the possible outcomes is called the **sample space** of the experiment, and is denoted by **S**. Any subset **E** of the sample space **S** is called an **event**.

- Tossing a coin. The sample space is **S = {H, T}**. **E = {H}** is an event.
- Two coin tosses: **S = {HH, HT, TH, TT}**. **E**: "Occurrence of Head in first toss" = **{HH, HT}**
- Tossing a die. The sample space is **S = {1, 2, 3, 4, 5, 6}**. **E = {2, 4, 6}** is an event, which can be described in words as "the number is even".

Probability measure, Axioms of Probability

Probability measure is a function that assigns a real value to every outcome of a random phenomena that satisfies the following axioms:

- $0 \leq P(A) \leq 1$ (Probabilities are non-negative and less than 1 for any event A)
- $P(S) = 1$ (one of the outcomes should occur)
- For two mutually exclusive events A and B
 - $P(A \cup B) = P(A) + P(B)$
- Interpretation of probability as a frequency :
 - Conduct an experiment (coin toss) N times. If N_A is number of times outcome A occurs then $P(A) = N_A/N$

Disjoint and non-disjoint outcomes

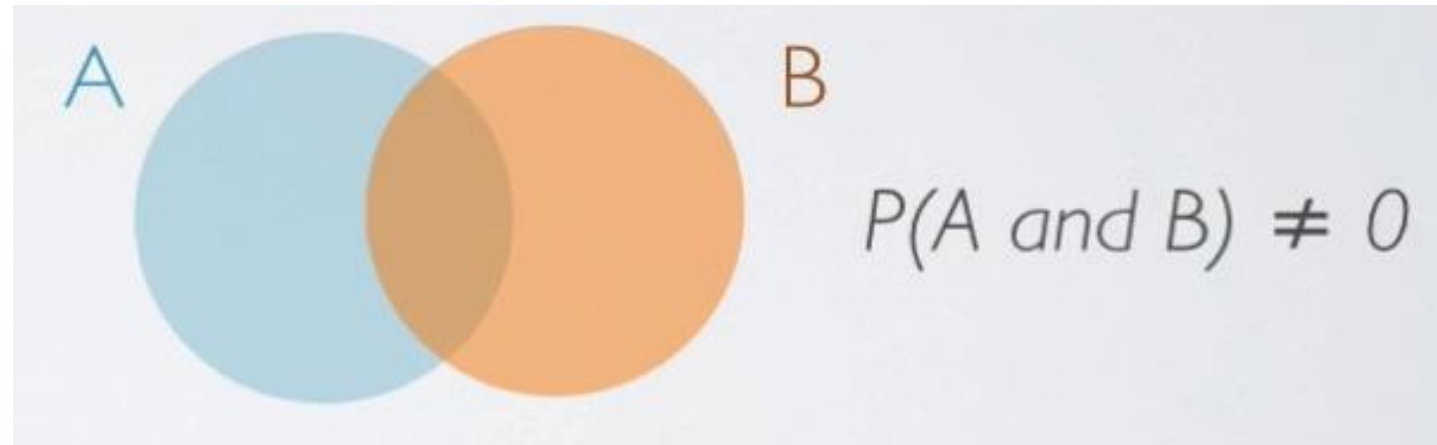
Disjoint (mutually exclusive) outcomes: Cannot happen at the same time.

- The outcome of a single coin toss cannot be a head and a tail.
- A student both cannot fail and pass a class.
- A single card drawn from a deck cannot be an ace and a queen.



Non-disjoint outcomes: Can happen at the same time.

- A student can get an A in Stats and A in Econ in the same semester.



Example: Disjoint event

In a two coin toss experiment the events **{HH}** and **{HT}** are mutually exclusive =>

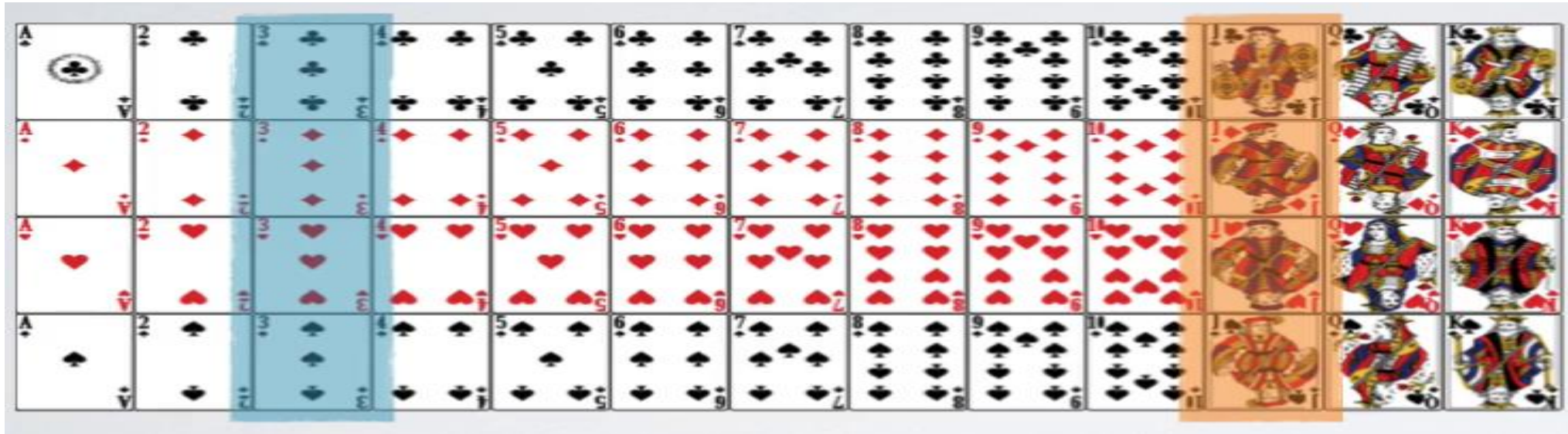
$$P(\text{HH or HT}) = P(\text{HH}) + P(\text{HT}) = 0.25 + 0.25 = 0.5$$

Note: The above is the same as the event **{H}** appears first in a two coin tossing experiment

Disjoint Events + General Addition Rule

Union of disjoint events

What is the probability of drawing a Jack or three from a well shuffled full deck of cards?

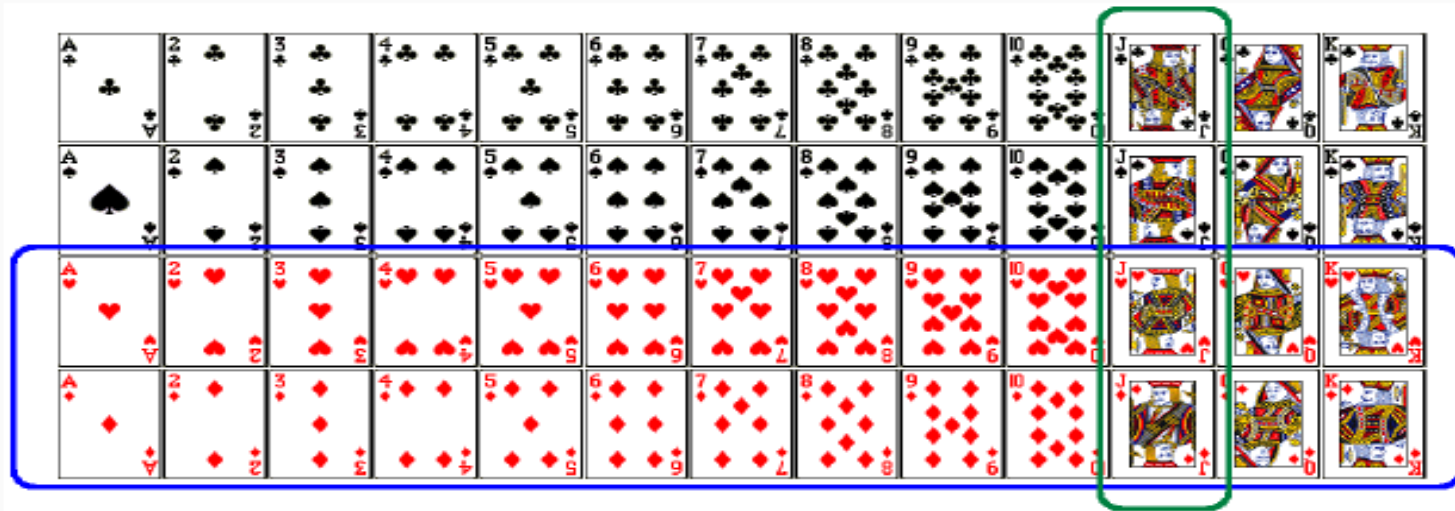


$$\begin{aligned} P(J \text{ or } 3) \\ &= P(J) + P(3) \\ &= (4/52) + (4/52) \\ &\approx 0.154 \end{aligned}$$

For disjoint events A and B,
 $P(A \text{ or } B) = P(A) + P(B)$

Union of non-disjoint events

What is the probability of drawing a jack or a red card from a well shuffled full deck?

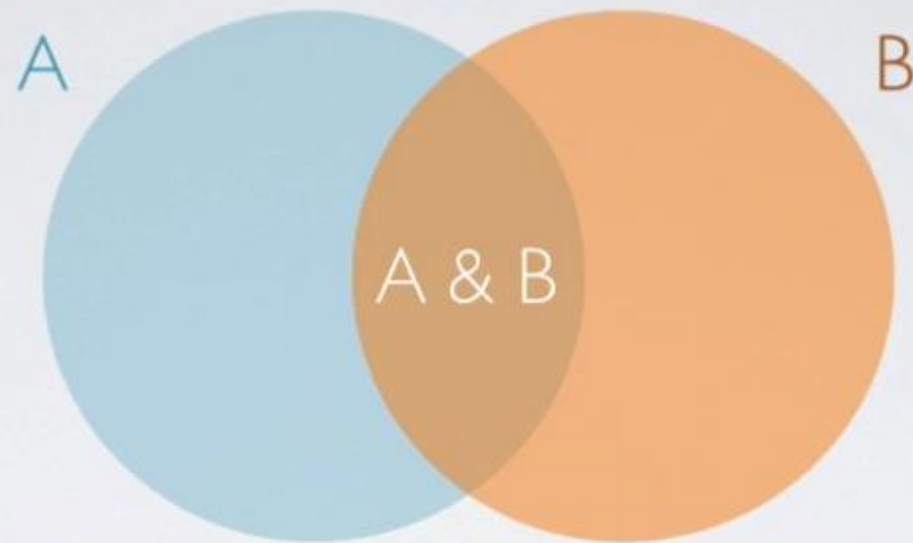


$$\begin{aligned} P(\text{jack or red}) &= P(\text{jack}) + P(\text{red}) - P(\text{jack and red}) \\ &= \frac{4}{52} + \frac{26}{52} - \frac{2}{52} = \frac{28}{52} \approx 0.538 \end{aligned}$$

For non-disjoint events A and B,
 $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

General addition rule:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



Note: When A and B are disjoint, $P(A \text{ and } B) = 0$, so the formula simplifies to $P(A \text{ or } B) = P(A) + P(B)$.

Complementary events

Complementary events are two mutually exclusive events whose probabilities that add up to 1.

- A couple has one kid. If we know that the kid is not a boy, what is gender of this kid? {F} as boy and girl are complementary outcomes.

The diagram illustrates complementary events for one and two coin tosses. For one toss, 'head' and 'tail' are complementary. For two tosses, 'head-head' and 'tail-head' are complementary, as indicated by a bracket and the word 'complementary' written above them.

one toss	head	tail
probability	0.5	0.5

two tosses	head - head	tail - tail	head - tail	tail - head
probability	0.25	0.25	0.25	0.25

Some rules of probability

- Following important probability rules can be proved using Venn diagrams

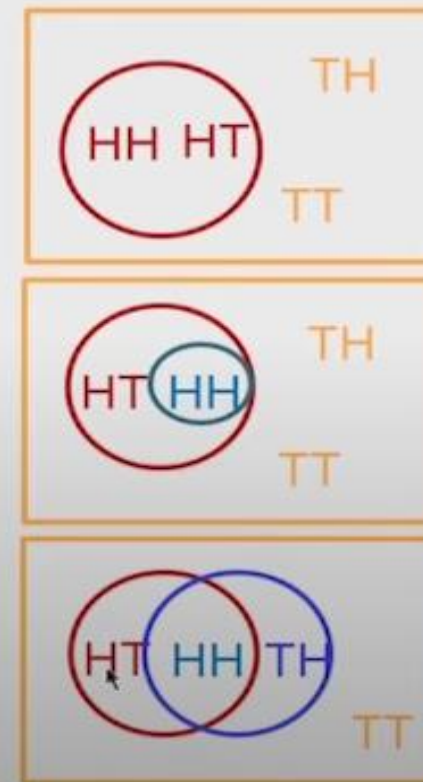
$S = \square$ $A = \bigcirc$ $B = \bigcirc$

All outcomes are equally likely

If A^c is the complement of event A ,
 $P(A^c) = P(S) - P(A) = 1 - P(A) = 0.5$

If $B \subseteq A$, $P(B) \leq P(A)$; $0.25 < 0.5$

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= 0.5 + 0.5 - 0.5 \cdot 0.5 = 0.75 \end{aligned}$$



Addition Rule of Disjoint Outcomes

If A_1 and A_2 represent two disjoint outcomes, then the probability that one of them occurs is given by

$$P(A_1 \text{ or } A_2) = P(A_1) + P(A_2)$$

If there are many disjoint outcomes A_1, \dots, A_k , then the probability that one of these outcomes will occur is

$$P(A_1) + P(A_2) + \dots + P(A_k)$$

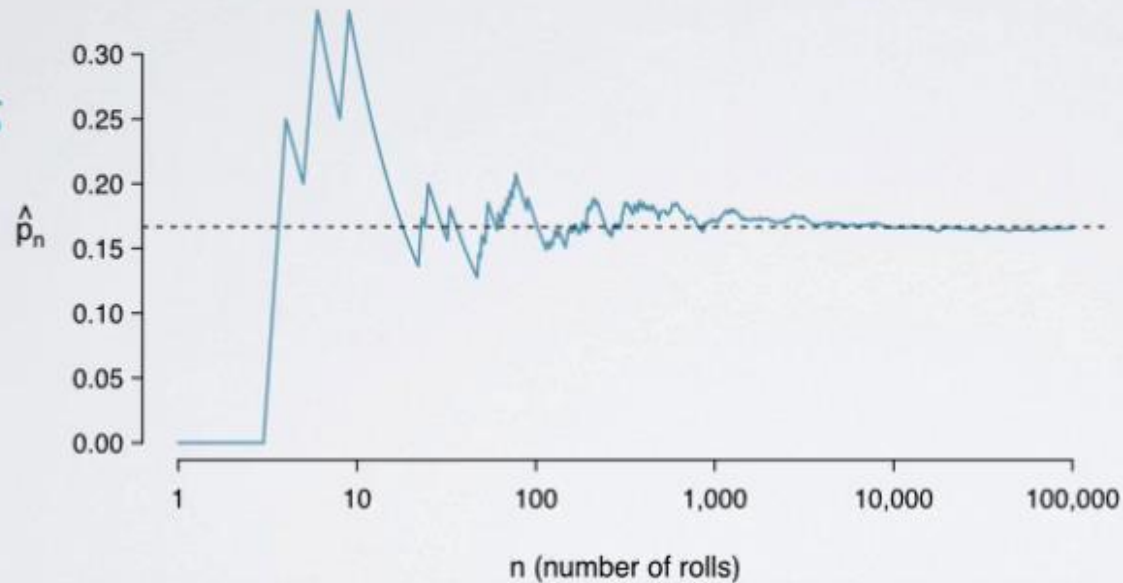
When rolling a die, the outcomes 1 and 2 are disjoint, and we compute the probability that one of these outcomes will occur by adding their separate probabilities:

$$P(1 \text{ or } 2) = P(1) + P(2) = 1/6 + 1/6 = 1/3$$

Law of large numbers

law of large numbers states that as more observations are collected, the proportion of occurrences with a particular outcome converges to the probability of that outcome.

examples



Say you toss a coin 10 times, and it lands on Heads each time. What do you think the chance is that another head will come up on the next toss? 0.5, less than 0.5, or more than 0.5?

H H H H H H H H H H ?

The probability is still
50%:
 $P(\text{H on the 11th toss})$
 $= P(\text{H on the 10th toss})$
 $= 0.50$

The coin is
not
due for a tail.

Common
misunderstanding of law
of large numbers:
gambler's fallacy
(law of averages)

Independent Events

Two processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other.

- Knowing that the coin landed on a head on the first toss does not provide any useful information for determining what the coin will land on in the second toss. \Rightarrow Outcomes of two tosses of a coin are independent.
- Ex. In a two coin toss experiment **$P(HH) = P(\text{Head in first toss}) * P(\text{Head in second toss}) = 0.5 * 0.5 = 0.25$**
- Knowing that the first card drawn from a deck is an ace does provide useful information for determining the probability of drawing an ace in the second draw. \Rightarrow Outcomes of two draws from a deck of cards (**without replacement**) are dependent.

Independent Events

two processes are **independent** if knowing the outcome of one provides no useful information about the outcome of the other.

1st toss



2nd toss



$$P(H) = 0.5$$

$$P(T) = 0.5$$

outcomes of two tosses of a coin are
independent

1st draw



2nd draw



$$P(A) = 3/51$$

$$P(J) = 4/51$$

outcomes of two draws from a deck of
cards (without replacement) are **dependent**

Product Rule for Independent Events

$$P(\mathbf{A \text{ and } B}) = P(\mathbf{A}) \times P(\mathbf{B})$$

Or more generally, **$P(\mathbf{A1 \text{ and } \dots \text{ and } Ak}) = P(\mathbf{A1}) \times \dots \times P(\mathbf{Ak})$**

You toss a coin twice, what is the probability of getting two tails in a row?

$$P(\text{T on the first toss}) \times P(\text{T on the second toss}) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Checking for Independence

If $P(\text{A occurs, given that B is true}) = P(\text{A} | \text{B}) = P(\text{A})$, then A and B are independent.

In 2013, SurveyUSA interviewed a random sample of 500 NC residents asking them whether they think widespread gun ownership protects law abiding citizens from crime, or makes society more dangerous.

- 58% of all respondents said it protects citizens.
- 67% of White respondents,
- 28% of Black respondents,
- and 64% of Hispanic respondents shared this view.

Opinion on gun ownership and race ethnicity are most likely _____?

- (a) complementary
- (b) mutually exclusive
- (c) independent
- ☒ (d) dependent
- (e) disjoint

$$P(\text{protects citizens}) = 0.58$$

$$P(\text{protects citizens} | \text{White}) = 0.67$$

$$P(\text{protects citizens} | \text{Black}) = 0.28$$

$$P(\text{protects citizens} | \text{Hispanic}) = 0.64$$

Example

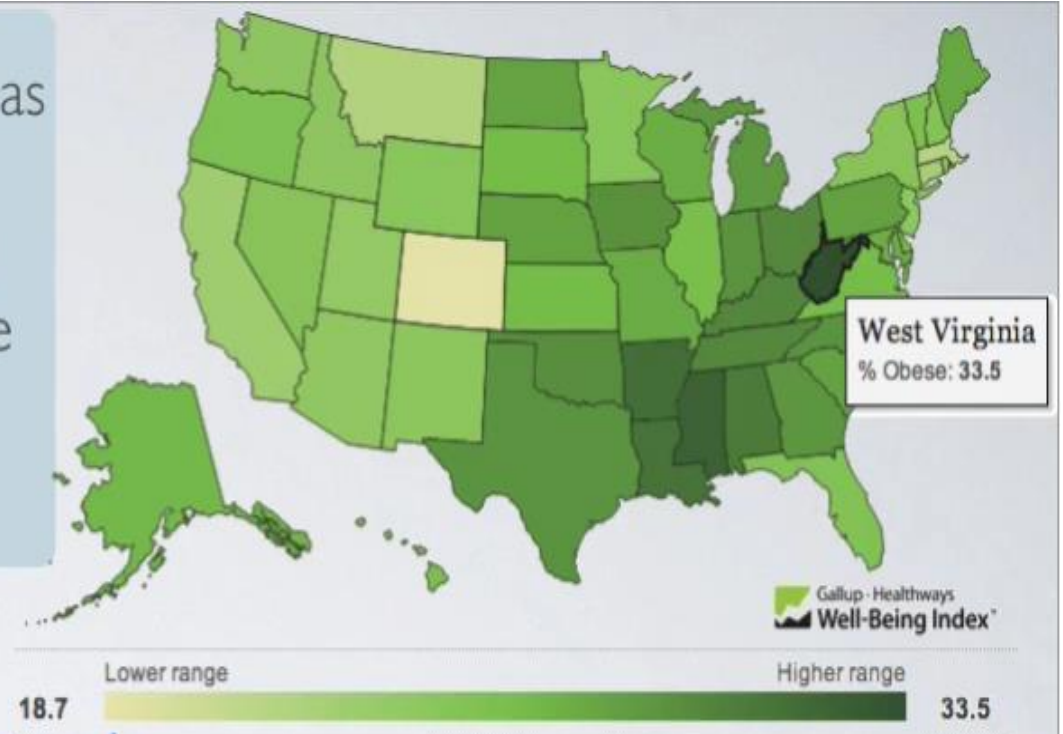
A 2012 Gallup poll suggests that West Virginia has the highest obesity rate among US states, with 33.5% of West Virginians being obese. Assuming that the obesity rate stayed constant, what is the probability that two randomly selected West Virginians are both obese? *independent*

$$P(\text{obese}) = 0.335$$

$$P(\text{both obese}) = P(\text{1st obese}) \times P(\text{2nd obese})$$

$$= 0.335 \times 0.335$$

$$\approx 0.11$$



Disjoint vs. Independent

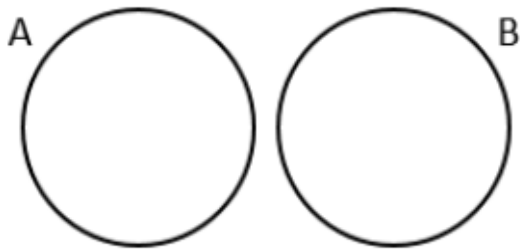
two events that are
disjoint
(mutually exclusive)
cannot happen
at the same time

$$P(A \text{ and } B) = 0$$

two processes are
independent
if knowing the outcome
of one
provides no useful
information about the
outcome of the other

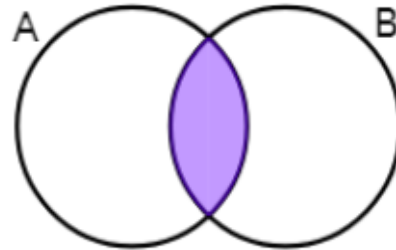
$$P(A | B) = P(A)$$

Mutually Exclusive Events



$$P(A \text{ or } B) = P(A) + P(B)$$

Non-Mutually Exclusive Events



$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

The Sum of the Probabilities of an **Event** and its **Complement** add to '1'.

Example: For a dice

$$P(>4) = \frac{2}{6} = \frac{1}{3}$$



$$P(\text{Complement } >4) = P(\leq 4) = \frac{4}{6} = \frac{2}{3}$$

$$P(>4) + P(\text{Complement } >4) = \frac{1}{3} + \frac{2}{3} = 1$$

Mutually Exclusive Example

What is the probability of a dice showing a 2 or 5?

$$P(2) = \frac{1}{6} \quad P(5) = \frac{1}{6}$$

$$P(2 \text{ or } 5) = P(2) + P(5)$$

$$= \frac{1}{6} + \frac{1}{6}$$

$$= \frac{2}{6} = \frac{1}{3}$$

The probability of a dice showing 2 or 5 is $\frac{1}{3}$

Disjoint vs. complementary

Q. ***Do the sum of probabilities of two disjoint events always add up to 1?***

A. Not necessarily, there may be more than 2 events in the sample space, e.g. party affiliation.

Q. ***Do the sum of probabilities of two complementary events always add up to 1?***

A. Yes, that's the definition of complementary, e.g. heads and tails.



References

- 1) Ani Adhikari. John DeNero, Computational and Inferential Thinking: The Foundations of Data Science. GitBook, 2019.
- 2) OpenIntro Statistics online book: OpenIntro Statistics