

# Introductory Statistics

Instructor: Guenther Walther

We will look at six main topics:

- ▶ Descriptive statistics for exploring data, especially visualization
- ▶ Some elementary probability
- ▶ Sampling distributions and the central limit theorem
- ▶ Regression
- ▶ Confidence intervals and tests of significance
- ▶ Multiple comparisons, reproducibility

There won't be many formulas, rather we will look at the important statistical ideas behind these methods. Once you understand these ideas, then it's not difficult to look up detailed formulas and more advanced methodology.

Good introductory textbooks are *Statistics* by Freedman, Pisani, and Purves , or *Introduction to Probability & Statistics* by Mendenhall and Beaver (which is a more formal exposition).

## Descriptive statistics - why is it important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29° F at launch.

## Descriptive statistics - why is it important?

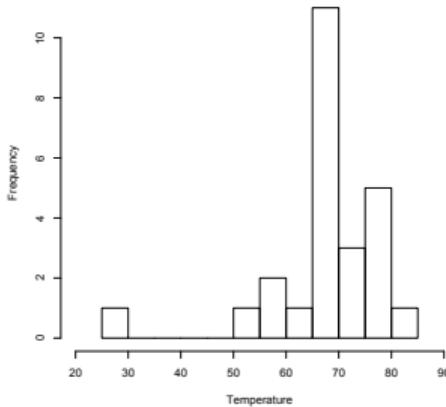
In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29° F at launch.

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F):  
66,70,69,80,68,67,72,70,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58,29

## Descriptive statistics - why is it important?

In January 1986, the space shuttle Challenger broke apart shortly after liftoff. The accident was caused by a part that was not designed to fly at the unusually cold temperature of 29° F at launch.

Here are the launch-temperatures of the first 25 shuttle missions (in degrees F):  
66,70,69,80,68,67,72,70,70,57,63,70,78,67,53,67,75,70,81,76,79,75,76,58,29



The two most important functions of descriptive statistics are:

- ▶ Communicate information
- ▶ Support reasoning about data

The two most important functions of descriptive statistics are:

- ▶ Communicate information
- ▶ Support reasoning about data

When exploring data of large size, it becomes essential to use summaries.

## Graphical summaries of data

It is best to use a graphical summary to communicate information, because people prefer to look at pictures rather than at numbers.

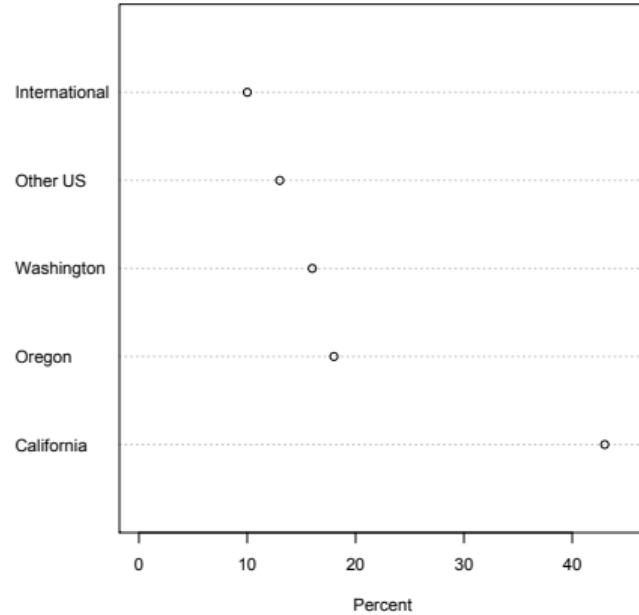
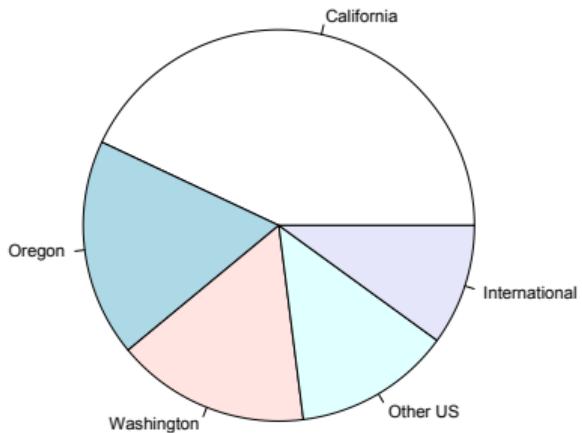
## Graphical summaries of data

It is best to use a graphical summary to communicate information, because people prefer to look at pictures rather than at numbers.

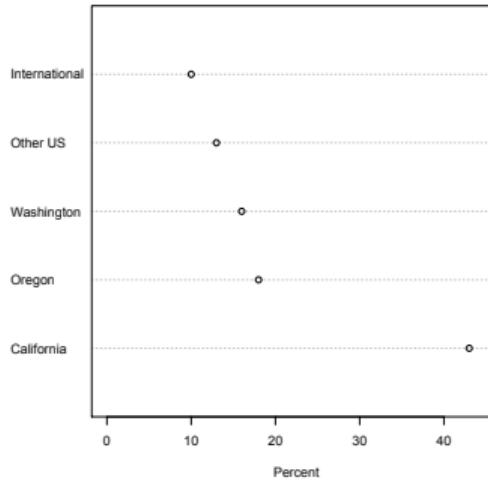
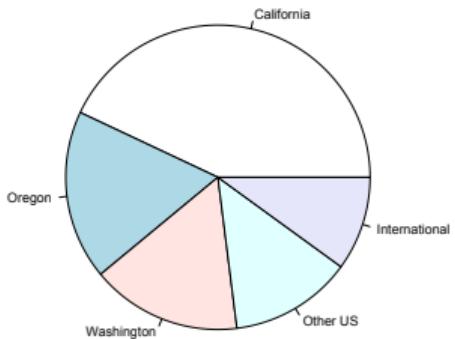
There are many ways to visualize data. The nature of the data and the goal of the visualization determine which method to choose.

## Pie chart and dot plot

For data that is *qualitative* (e.g. colors, car types,...), use a **pie chart** or a **dot plot**.



## Pie chart and dot plot



The dot plot makes it easier to compare frequencies of various categories, while the pie chart allows more easily to eyeball what fraction of the total a category corresponds to.

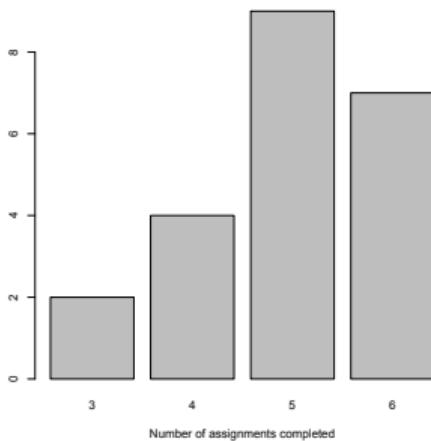
## Bar graph

When the data are *quantitative* (i.e. numbers), then they should be put on a number line. This is because the ordering and the distance between the numbers convey important information.

## Bar graph

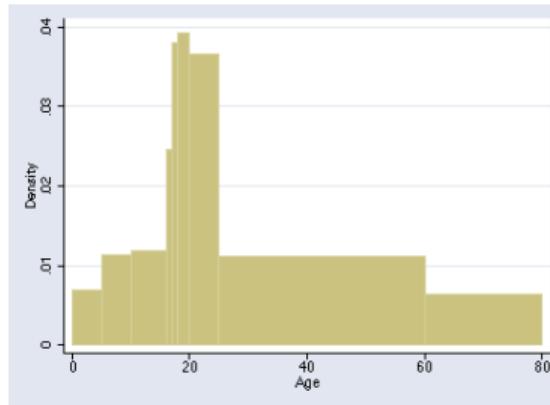
When the data are *quantitative* (i.e. numbers), then they should be put on a number line. This is because the ordering and the distance between the numbers convey important information.

The **bar graph** is essentially a dot plot put on its side.



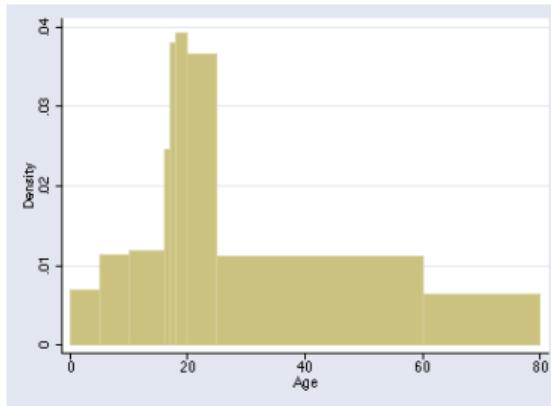
## The histogram

The **histogram** allows to use blocks with different widths.



## The histogram

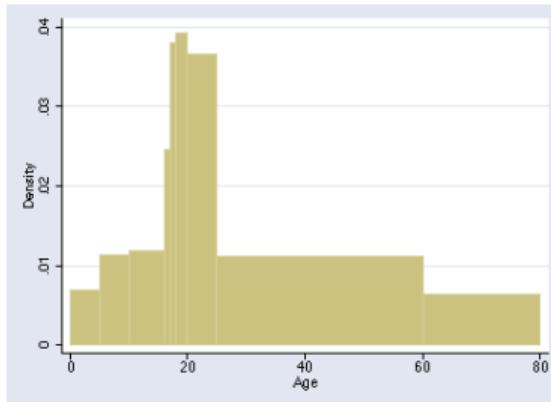
The **histogram** allows to use blocks with different widths.



Key point: The areas of the blocks are proportional to frequency.

## The histogram

The **histogram** allows to use blocks with different widths.

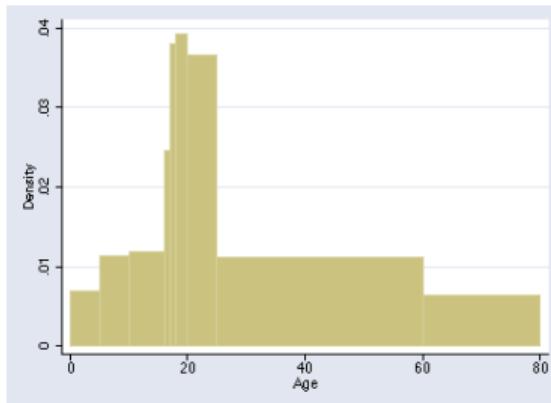


Key point: The areas of the blocks are proportional to frequency.

So the percentage falling into a block can be figured without a vertical scale since the total area equals 100%.

## The histogram

The **histogram** allows to use blocks with different widths.

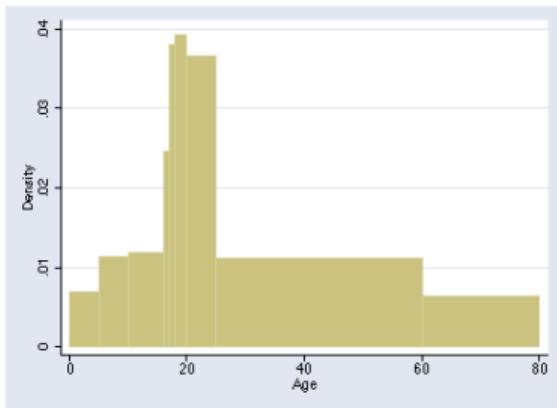


Key point: The areas of the blocks are proportional to frequency.

So the percentage falling into a block can be figured without a vertical scale since the total area equals 100%.

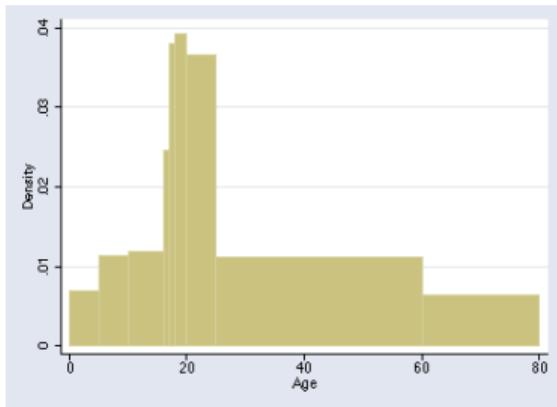
But it's helpful to have a vertical scale (*density scale*). Its unit is '% per unit', so in the above example the vertical unit is '% per year'.

The histogram gives two kinds of information about the data:

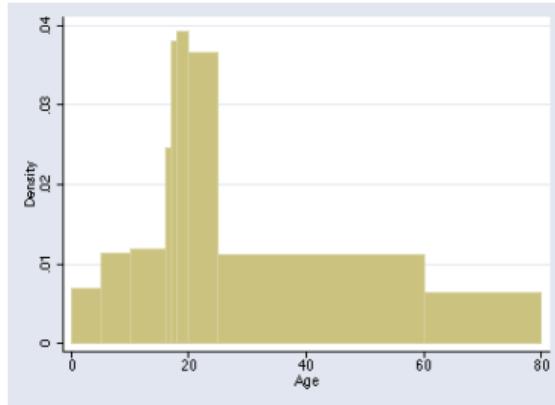


1. **Density (crowding):** The height of the bar tells how many subjects there are for one unit on the horizontal scale.

The histogram gives two kinds of information about the data:

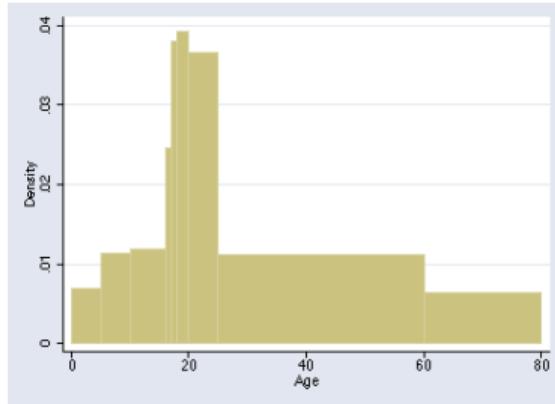


1. **Density (crowding):** The height of the bar tells how many subjects there are for one unit on the horizontal scale. For example, the highest density is around age 19 as  $.04 = 4\%$  of all subjects are age 19. In contrast, only about  $0.7\%$  of subjects fall into each one year range for ages 60–80.



2. Percentages (relative frequencies): Those are given by

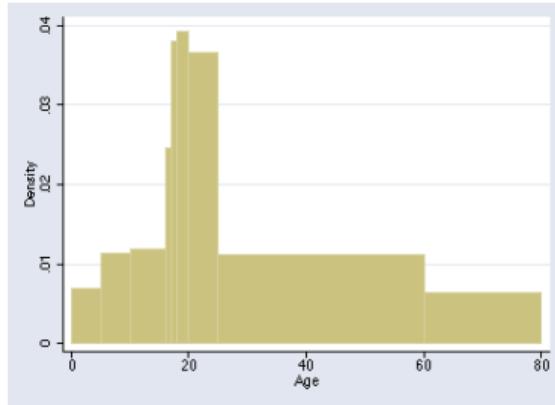
$$\text{area} = \text{height} \times \text{width}.$$



2. Percentages (relative frequencies): Those are given by

$$\text{area} = \text{height} \times \text{width}.$$

For example, about 14% of all subjects fall into the age range 60–80, because the corresponding area is  $(20 \text{ years}) \times (0.7 \% \text{ per year})=14 \text{ \%}$ .



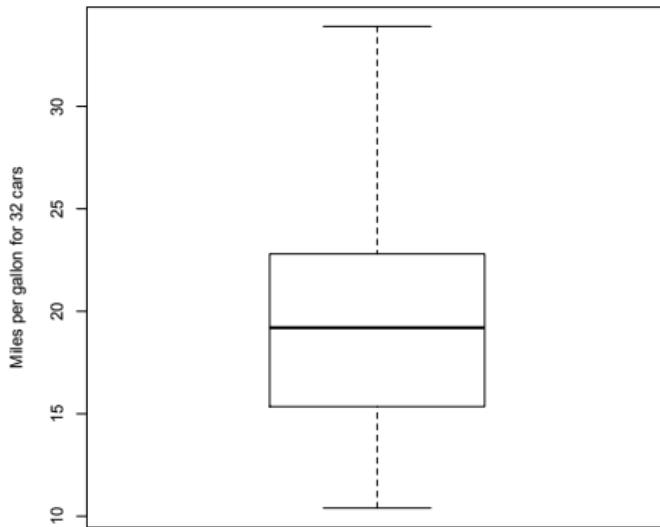
2. Percentages (relative frequencies): Those are given by

$$\text{area} = \text{height} \times \text{width}.$$

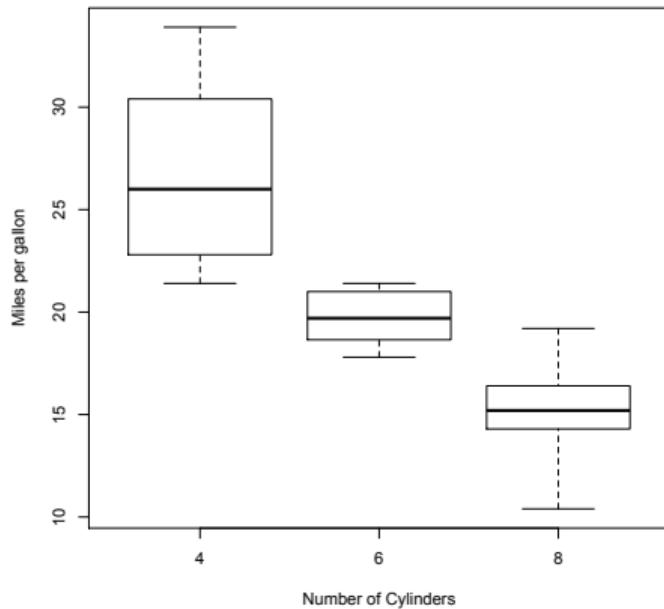
For example, about 14% of all subjects fall into the age range 60–80, because the corresponding area is  $(20 \text{ years}) \times (0.7 \% \text{ per year}) = 14\%$ . Alternatively, you can find this answer by eyeballing that this area makes up roughly  $1/7$  of the total area of the histogram, so roughly  $1/7 = 14\%$  of all subjects fall in that range.

## The boxplot (box-and-whisker plot)

The **boxplot** depicts five key numbers of the data:

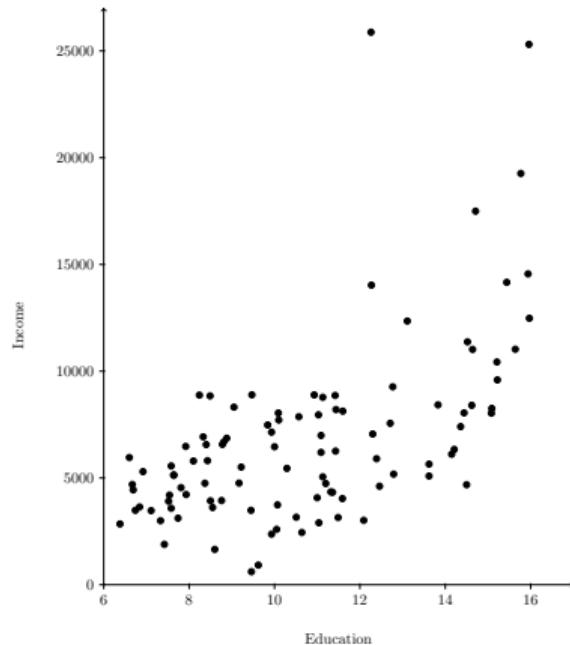


The **boxplot** conveys less information than a histogram, but it takes up less space and so is well suited to compare several datasets:



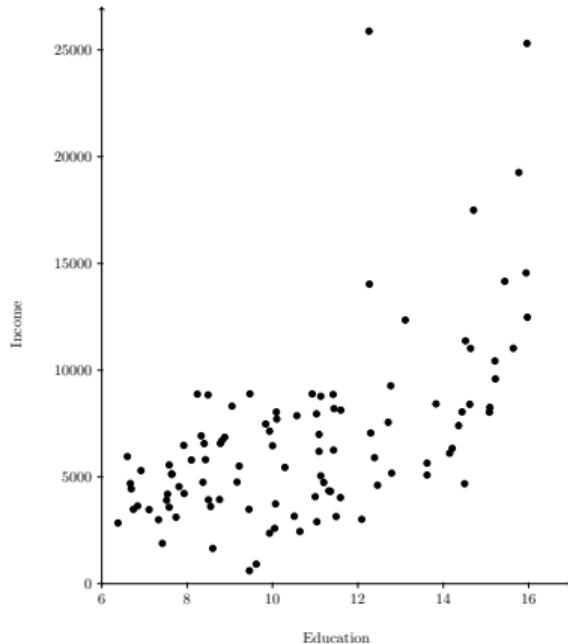
# The scatterplot

The **scatterplot** is used to depict data that come as *pairs*.



# The scatterplot

The **scatterplot** is used to depict data that come as *pairs*.



The scatterplot visualizes the relationship between the two variables.

## Providing context is important

Statistical analyses typically compare the observed data to a reference. Therefore context is essential for graphical integrity.

## Providing context is important

Statistical analyses typically compare the observed data to a reference. Therefore context is essential for graphical integrity.

- ▶ ‘The Visual Display of Quantitative Information’ by Edward Tufte (p.74)

## Providing context is important

Statistical analyses typically compare the observed data to a reference. Therefore context is essential for graphical integrity.

- ▶ ‘The Visual Display of Quantitative Information’ by Edward Tufte (p.74)

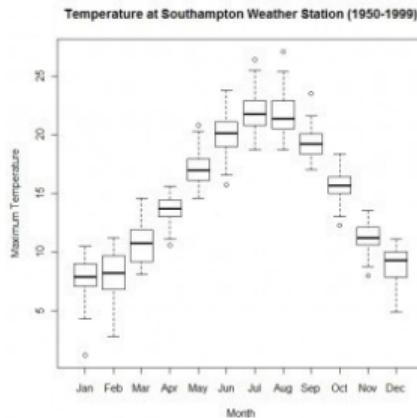
One way to provide context is by using *small multiples*.

## Providing context is important

Statistical analyses typically compare the observed data to a reference. Therefore context is essential for graphical integrity.

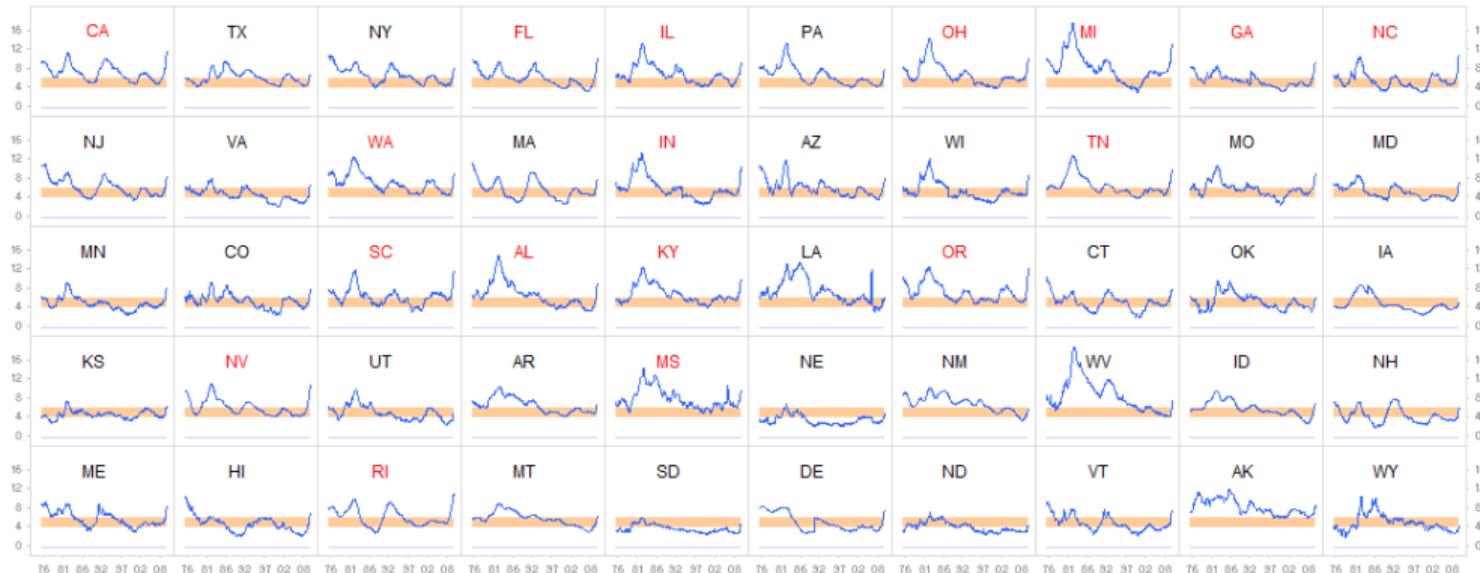
- ▶ ‘The Visual Display of Quantitative Information’ by Edward Tufte (p.74)

One way to provide context is by using *small multiples*. The compact design of the boxplot makes it well suited for this task:



# Providing context with small multiples

Monthly Unemployment Rates by State, Jan 1976 - Apr 2009



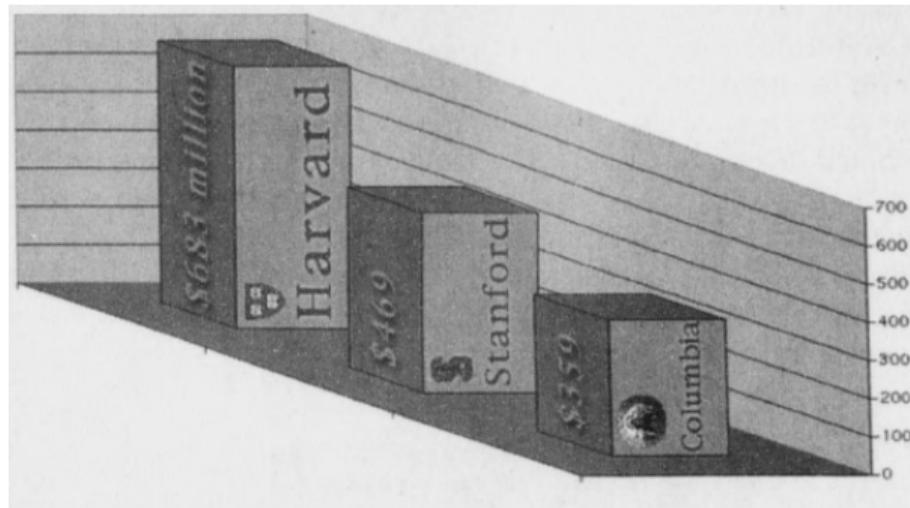
Source: Bureau of Labor Statistics

Notes: The orange band denotes a "normal" unemployment rate (4%-6%);

State code in red: unemployment rate in April 2009 is higher than the US average

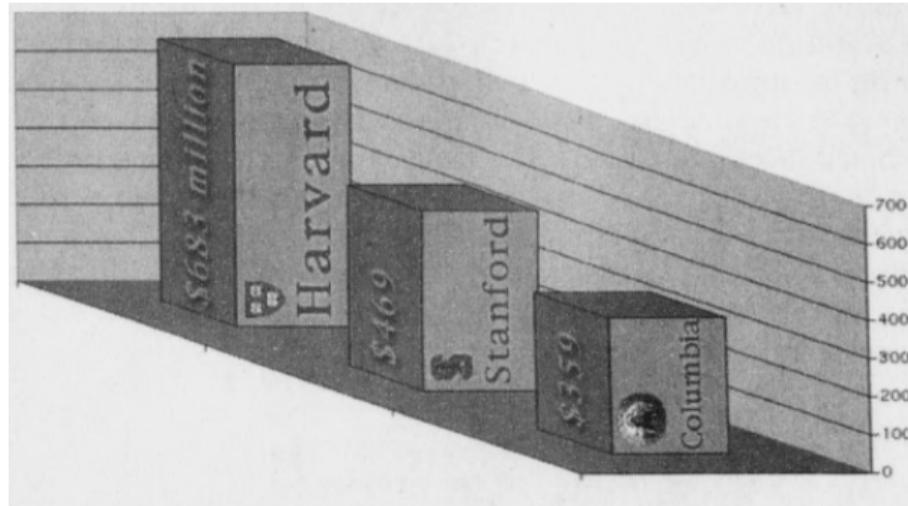
## Pitfalls when visualizing data

Sophisticated software makes it tempting to produce showy but poor visualizations:



## Pitfalls when visualizing data

Sophisticated software makes it tempting to produce showy but poor visualizations:



- ▶ The 'Ghettysburg Powerpoint Presentation' by Peter Norvig

## Numerical summary measures

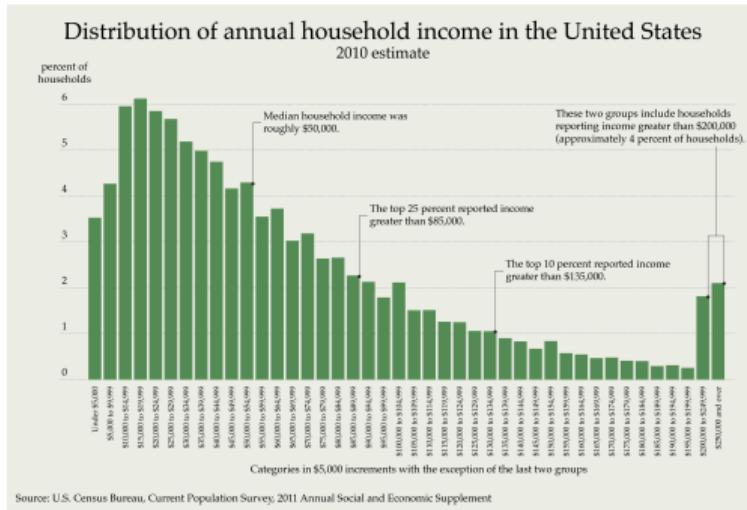
For summarizing data with one number, use the **mean** (=average) or the **median**.

## Numerical summary measures

For summarizing data with one number, use the **mean** (=average) or the **median**.  
The median is the number that is larger than half the data and smaller than the other half.

# Numerical summary measures

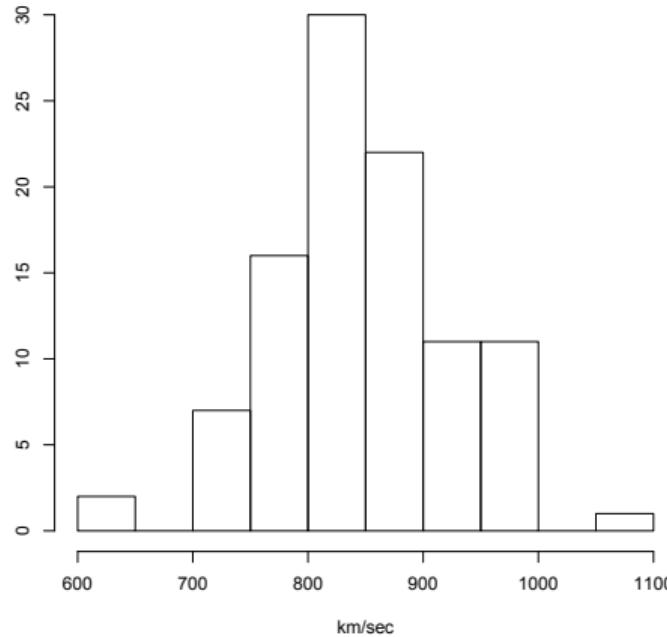
For summarizing data with one number, use the **mean** (=average) or the **median**.  
The median is the number that is larger than half the data and smaller than the other half.



## Mean vs. median

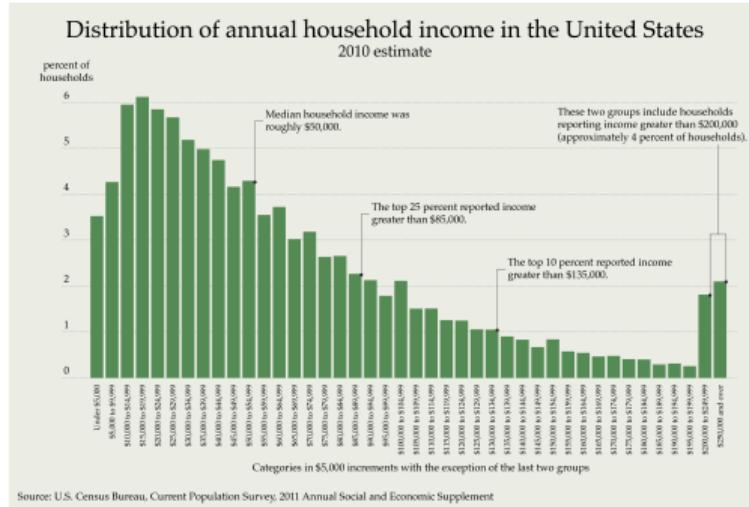
Mean and median are the same when the histogram is symmetric.

100 measurements of the speed of light



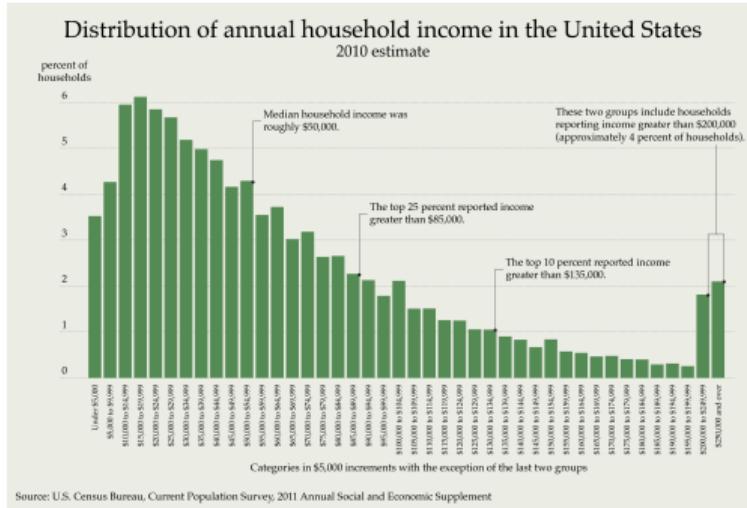
# Mean vs. median

When the histogram is *skewed to the right*, then the mean can be much larger than the median.



## Mean vs. median

When the histogram is *skewed to the right*, then the mean can be much larger than the median.



So if the histogram is very skewed, then use the median.

## Mean vs. median

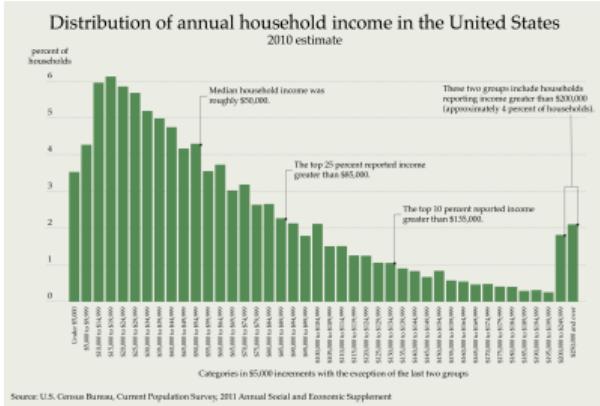
If the median sales price of 10 homes is \$ 1 million, then we know that 5 homes sold for \$ 1 million or more.

## Mean vs. median

If the median sales price of 10 homes is \$ 1 million, then we know that 5 homes sold for \$ 1 million or more.

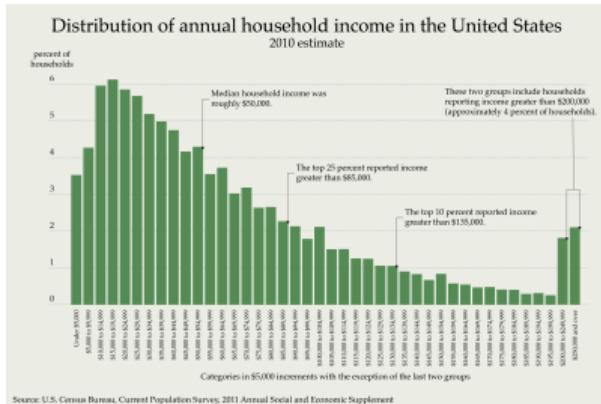
If we are told that the average sale price is \$ 1 million, then we can't draw such a conclusion:

# Percentiles



The 90th percentile of incomes is \$ 135,000: 90% of households report an income of \$ 135,000 or less, 10% report more.

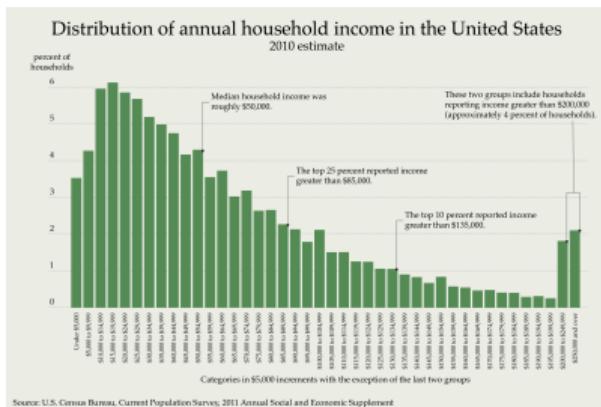
# Percentiles



The 90th percentile of incomes is \$ 135,000: 90% of households report an income of \$ 135,000 or less, 10% report more.

The 75th percentile is called **3rd quartile**: \$ 85,000

# Percentiles

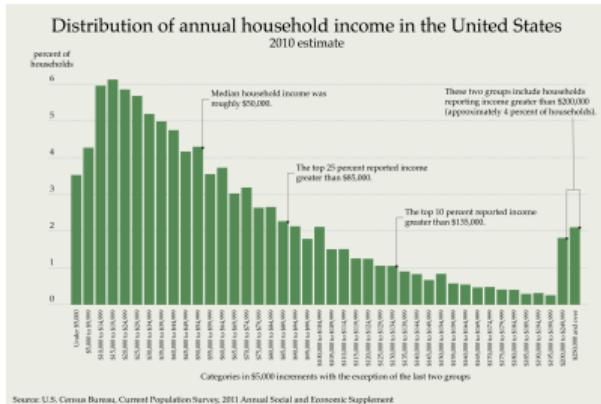


The 90th percentile of incomes is \$ 135,000: 90% of households report an income of \$ 135,000 or less, 10% report more.

The 75th percentile is called **3rd quartile**: \$ 85,000

The 50th percentile is the **median**: \$ 50,000

# Percentiles



The 90th percentile of incomes is \$ 135,000: 90% of households report an income of \$ 135,000 or less, 10% report more.

The 75th percentile is called **3rd quartile**: \$ 85,000

The 50th percentile is the **median**: \$ 50,000

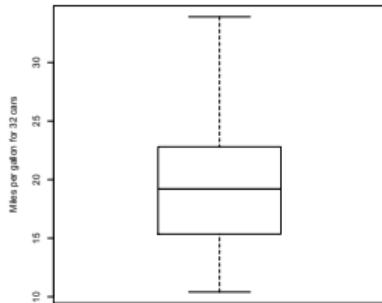
The 25th percentile is called **1st quartile**.

## Five-number summary

Recall that the boxplot gives a **five-number summary** of the data:  
the smallest number, 1st quartile, median, 3rd quartile, largest number.

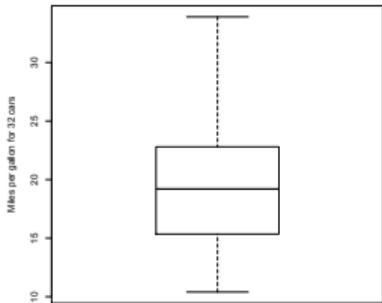
## Five-number summary

Recall that the boxplot gives a **five-number summary** of the data:  
the smallest number, 1st quartile, median, 3rd quartile, largest number.



## Five-number summary

Recall that the boxplot gives a **five-number summary** of the data:  
the smallest number, 1st quartile, median, 3rd quartile, largest number.



The **interquartile range** = 3rd quartile – 1st quartile.  
It measures how spread out the data are.

## The standard deviation

A more commonly used measure of spread is the **standard deviation**.

## The standard deviation

A more commonly used measure of spread is the **standard deviation**.

$\bar{x}$  stands for the average of the numbers  $x_1, \dots, x_n$ .

## The standard deviation

A more commonly used measure of spread is the **standard deviation**.

$\bar{x}$  stands for the average of the numbers  $x_1, \dots, x_n$ .

The standard deviation of these numbers is

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

## The standard deviation

A more commonly used measure of spread is the **standard deviation**.

$\bar{x}$  stands for the average of the numbers  $x_1, \dots, x_n$ .

The standard deviation of these numbers is

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

The two numbers  $\bar{x}$  and  $s$  are often used to summarize data. Both are sensitive to a few large or small data.

## The standard deviation

A more commonly used measure of spread is the **standard deviation**.

$\bar{x}$  stands for the average of the numbers  $x_1, \dots, x_n$ .

The standard deviation of these numbers is

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

The two numbers  $\bar{x}$  and  $s$  are often used to summarize data. Both are sensitive to a few large or small data.

If that is a concern, use the median and the interquartile range.

## Mini quiz

- ▶ For each of the following two data sets state an appropriate way for visualizing the data:
  - a) A list of the eye colors of 120 people.
  - b) A list that gives both the size (measured in square feet) and the number of bedrooms for 1513 houses.
- ▶ The average sales price for houses in a certain county during the last year was \$ 342,000. Do the houses that sold for more than \$ 342,000 constitute more/equal/less than 50% of all sales?

## Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

## Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

This is difficult to determine exactly as there are more than 250 million people of voting age in the U.S.

## Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

This is difficult to determine exactly as there are more than 250 million people of voting age in the U.S.

But it's not difficult to *estimate* this percentage quite well:

## Producing data, sampling

What percentage of voters approve of the way the U.S. President is handling his job?

This is difficult to determine exactly as there are more than 250 million people of voting age in the U.S.

But it's not difficult to *estimate* this percentage quite well:

Sample 1,000 (say) voters at random. Then use the approval percentage among those voters as an estimate for the approval percentage of all voters.

## What is statistical inference?

**Population:** the entire group of subjects about which we want information

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in  
approval percentage among all U.S. voters

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in  
approval percentage among all U.S. voters

**Sample:** the part of the population from which we collect information

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in  
approval percentage among all U.S. voters

**Sample:** the part of the population from which we collect information  
the 1,000 voters selected at random

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in  
approval percentage among all U.S. voters

**Sample:** the part of the population from which we collect information  
the 1,000 voters selected at random

**Statistic (estimate):** the quantity we are interested in as measured in the sample

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in  
approval percentage among all U.S. voters

**Sample:** the part of the population from which we collect information  
the 1,000 voters selected at random

**Statistic (estimate):** the quantity we are interested in as measured in the sample  
approval percentage among the sampled voters

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in  
approval percentage among all U.S. voters

**Sample:** the part of the population from which we collect information  
the 1,000 voters selected at random

**Statistic (estimate):** the quantity we are interested in as measured in the sample  
approval percentage among the sampled voters

Key point: even a relatively small sample (100 or 1,000) will produce an estimate that  
is close to the parameter of a very large population of 250 million subjects.

## What is statistical inference?

**Population:** the entire group of subjects about which we want information  
all U.S. voters

**Parameter:** the quantity about the population we are interested in  
approval percentage among all U.S. voters

**Sample:** the part of the population from which we collect information  
the 1,000 voters selected at random

**Statistic (estimate):** the quantity we are interested in as measured in the sample  
approval percentage among the sampled voters

Key point: even a relatively small sample (100 or 1,000) will produce an estimate that is close to the parameter of a very large population of 250 million subjects. This is the reason why statistics is so powerful.

Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

## Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

## Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

## Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

**selection bias:** a sample of convenience makes it more likely to sample certain subjects than others

## Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

**selection bias:** a sample of convenience makes it more likely to sample certain subjects than others

**non-response bias:** parents are less likely to answer a survey request at 6 pm because they are busy with children and dinner

## Sampling correctly is very important

It's tempting to sample 1,000 voters in your hometown.

This is a **sample of convenience**. This is not a good way to sample because the voters will be different from the population of U.S. voters.

This will introduce **bias**, i.e. this sampling will favor a certain outcome.

**selection bias:** a sample of convenience makes it more likely to sample certain subjects than others

**non-response bias:** parents are less likely to answer a survey request at 6 pm because they are busy with children and dinner

**voluntary response bias:** websites that post reviews of businesses are more likely to get responses from customers who had very bad or very good experiences

## Sampling designs

The best methods for sampling use chance in a planned way:

## Sampling designs

The best methods for sampling use chance in a planned way:

a **simple random sample** selects subjects at random without replacement

## Sampling designs

The best methods for sampling use chance in a planned way:

a **simple random sample** selects subjects at random without replacement

a **stratified random sample** divides the population into groups of similar subjects called *strata* (e.g. urban, suburban, and rural voters).

## Sampling designs

The best methods for sampling use chance in a planned way:

a **simple random sample** selects subjects at random without replacement

a **stratified random sample** divides the population into groups of similar subjects called *strata* (e.g. urban, suburban, and rural voters). Then one chooses a simple random sample in each stratum and combines these.

## Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**.

## Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

## Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}$$

## Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}$$

The chance error (sampling error) will get smaller as the sample size gets bigger.

## Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}$$

The chance error (sampling error) will get smaller as the sample size gets bigger. Moreover, we can compute how large the chance error will be.

## Bias and chance error

Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error**. Drawing another sample will result in a different chance error.

$$\text{estimate} = \text{parameter} + \text{bias} + \text{chance error}$$

The chance error (sampling error) will get smaller as the sample size gets bigger. Moreover, we can compute how large the chance error will be.

This is not the case for the bias (systematic error):

Increasing the sample size just repeats the error on a larger scale, and typically we don't know how large the bias is.

## Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

## Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- ▶ This means that there is an **association** between red meat consumption and cancer: there is a link between these two.

## Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- ▶ This means that there is an **association** between red meat consumption and cancer: there is a link between these two.
- ▶ But this does **not** mean that eating red meat *causes* cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.

## Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- ▶ This means that there is an **association** between red meat consumption and cancer: there is a link between these two.
- ▶ But this does **not** mean that eating red meat *causes* cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.

This is an **observational study**: It measures outcomes of interest and this can be used to establish association.

## Observational Studies

People who eat red meat have higher rates of certain cancers than people who don't eat red meat.

- ▶ This means that there is an **association** between red meat consumption and cancer: there is a link between these two.
- ▶ But this does **not** mean that eating red meat *causes* cancer: people who don't eat red meat are known to exercise more and drink less alcohol, and it could be the latter two issues that cause the difference in cancer rates.

This is an **observational study**: It measures outcomes of interest and this can be used to establish association.

But **association is not causation**, because there may be **confounding factors** such as exercise that are associated both with red meat consumption and cancer.

## Randomized controlled experiments

To establish causation, an **experiment** is required:

## Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

## Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

## Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- ▶ The **subjects** are assigned into treatment and control groups at **random**.

## Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- ▶ The **subjects** are assigned into treatment and control groups at **random**.
- ▶ When possible, subjects in the control group get a **placebo**: it resembles the treatment but is neutral.

## Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- ▶ The **subjects** are assigned into treatment and control groups at **random**.
- ▶ When possible, subjects in the control group get a **placebo**: it resembles the treatment but is neutral. Assigning a placebo makes sure that both groups are equally affected by the **placebo effect**: the idea of being treated may have an effect by itself.

## Randomized controlled experiments

To establish causation, an **experiment** is required:

A **treatment** (e.g. eating red meat) is *assigned* to people in the **treatment group** but not to people in the **control group**.

Then the outcomes in the two groups are compared. To rule out confounders, both groups should be similar, apart from the treatment. To this end:

- ▶ The **subjects** are assigned into treatment and control groups at **random**.
- ▶ When possible, subjects in the control group get a **placebo**: it resembles the treatment but is neutral. Assigning a placebo makes sure that both groups are equally affected by the **placebo effect**: the idea of being treated may have an effect by itself.
- ▶ The experiment is **double-blind**: neither the subjects nor the evaluators know the assignments to treatment and control.

## The placebo effect

The placebo effect is still not fully understood and is one of the most interesting phenomena in science.

## The placebo effect

The placebo effect is still not fully understood and is one of the most interesting phenomena in science.

'The weird power of the placebo effect, explained' by Brian Resnick (7/7/2017)

## The logic of randomized controlled experiments

Randomization serves two purposes:

## The logic of randomized controlled experiments

Randomization serves two purposes:

- ▶ It makes the treatment group similar to the control group. Therefore influences other than the treatment operate equally on both groups, apart from differences due to chance.

## The logic of randomized controlled experiments

Randomization serves two purposes:

- ▶ It makes the treatment group similar to the control group. Therefore influences other than the treatment operate equally on both groups, apart from differences due to chance.
- ▶ It allows to assess how relevant the treatment effect is, by calculating the size of chance effects when comparing the outcomes in the two groups (see later).

## Mini quiz

For each of the following three sampling plans, say whether it represents simple random sampling or whether it leads to selection bias or to non-response bias, or to voluntary response bias:

- ▶ A news company located next to Times Square in New York wants to get a sense how people feel about a proposed law on immigration. A reporter steps out of the building and randomly selects 100 people walking there and asks them about the proposed law.
- ▶ A car company wants to get a sense how satisfied the owners of its new car model are with the quality of that car. It randomly selects 250 numbers from the all the vehicle registration numbers that have been issued for this model and contacts the owners of that model.
- ▶ An airline wants to do a customer survey in order to improve its service. For one month, it sends an email to a random sample of customers which flew with the airline on the previous day (no customer will be contacted more than once). The email states that the airline would like the customer to fill out a 10 minute survey in order to help the airline improve its service.

## What is probability?

In 2015 there were about 4 million babies born in the US, and 48.8% of the newborns were girls.

## What is probability?

In 2015 there were about 4 million babies born in the US, and 48.8% of the newborns were girls.

We write:  $P(\text{newborn is a girl}) = 48.8\%$ .

## What is probability?

In 2015 there were about 4 million babies born in the US, and 48.8% of the newborns were girls.

We write:  $P(\text{newborn is a girl}) = 48.8\%$ .

The probability of an event is defined as the proportion of times this event occurs in many repetitions.

## What is probability?

In 2015 there were about 4 million babies born in the US, and 48.8% of the newborns were girls.

We write:  $P(\text{newborn is a girl}) = 48.8\%$ .

The probability of an event is defined as the proportion of times this event occurs in many repetitions.

This is the standard definition of probability. It requires that it is possible to repeat this chance experiment many times.

## What is probability?

In 2015 there were about 4 million babies born in the US, and 48.8% of the newborns were girls.

We write:  $P(\text{newborn is a girl}) = 48.8\%$ .

The probability of an event is defined as the proportion of times this event occurs in many repetitions.

This is the standard definition of probability. It requires that it is possible to repeat this chance experiment many times.

While interned during the Second World War, John Kerrich tossed a coin 10,000 times and observed 5,067 tosses resulting in heads.

## What is probability?

The long-run interpretation of probability can make it difficult to interpret it for single event:

'What I was wrong about this year' by David Leonhardt (12/24/2017)

## What is probability?

The long-run interpretation of probability can make it difficult to interpret it for single event:

'What I was wrong about this year' by David Leonhardt (12/24/2017)

Sometimes people use a different interpretation:

'The probability that my best friend calls today is 30%'.

## What is probability?

The long-run interpretation of probability can make it difficult to interpret it for single event:

'What I was wrong about this year' by David Leonhardt (12/24/2017)

Sometimes people use a different interpretation:

'The probability that my best friend calls today is 30%'.  
Such a 'subjective probability' is not based on experiments, and different people may assign different subjective probabilities to the same event.

## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that  $P(\text{newborn is a girl}) = 48.8\%$ .

## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that  $P(\text{newborn is a girl}) = 48.8\%$ .

Therefore  $P(\text{newborn is a boy}) = 51.2\%$ .

## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that  $P(\text{newborn is a girl}) = 48.8\%$ .

Therefore  $P(\text{newborn is a boy}) = 51.2\%$ .

More formally, write A for an event, such as A='newborn is a girl'.

## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that  $P(\text{newborn is a girl}) = 48.8\%$ .

Therefore  $P(\text{newborn is a boy}) = 51.2\%$ .

More formally, write A for an event, such as A='newborn is a girl'.

Complement rule:  $P(A \text{ does not occur}) = 1 - P(A)$

## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that  $P(\text{newborn is a girl}) = 48.8\%$ .

Therefore  $P(\text{newborn is a boy}) = 51.2\%$ .

More formally, write A for an event, such as A='newborn is a girl'.

Complement rule:  $P(A \text{ does not occur}) = 1 - P(A)$

Now let's look at a different chance experiment: Rolling a die.



## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that  $P(\text{newborn is a girl}) = 48.8\%$ .

Therefore  $P(\text{newborn is a boy}) = 51.2\%$ .

More formally, write A for an event, such as A='newborn is a girl'.

Complement rule:  $P(A \text{ does not occur}) = 1 - P(A)$



Now let's look at a different chance experiment: Rolling a die.

Since each of its six faces is equally likely to come up, each has probability  $1/6$ .

## Four basic rules

Probabilities are always between 0 and 1. Computing probabilities comes down to applying a few basic rules repeatedly:

We saw that  $P(\text{newborn is a girl}) = 48.8\%$ .

Therefore  $P(\text{newborn is a boy}) = 51.2\%$ .

More formally, write A for an event, such as A='newborn is a girl'.

Complement rule:  $P(A \text{ does not occur}) = 1 - P(A)$



Now let's look at a different chance experiment: Rolling a die.

Since each of its six faces is equally likely to come up, each has probability  $1/6$ .

Rule for equally likely outcomes: If there are  $n$  possible outcomes and they are equally likely, then  $P(A) = \frac{\text{number of outcomes in } A}{n}$

## Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

## Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.

A=  on first roll, B=  on first roll, C=  on second roll

## Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.

$A = \boxed{\bullet}$  on first roll,  $B = \boxed{\bullet\bullet}$  on first roll,  $C = \boxed{\circ}$  on second roll

A and B are mutually exclusive, but A and C are not.

## Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.

A=  on first roll, B=  on first roll, C=  on second roll

A and B are mutually exclusive, but A and C are not.

Addition rule: If A and B are mutually exclusive, then

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

## Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.

A=  on first roll, B=  on first roll, C=  on second roll

A and B are mutually exclusive, but A and C are not.

Addition rule: If A and B are mutually exclusive, then

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

Two events are *independent* if knowing that one occurs does not change the probability that the other occurs.

## Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.

A=  on first roll, B=  on first roll, C=  on second roll

A and B are mutually exclusive, but A and C are not.

Addition rule: If A and B are mutually exclusive, then

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

Two events are *independent* if knowing that one occurs does not change the probability that the other occurs.

B and C are independent, but A and B are not.

## Four basic rules

The next two rules make it possible to express probabilities of multiple events as those of the individual events.

A and B are *mutually exclusive* if they cannot occur at the same time.

As an example, we roll a die twice.

A=  on first roll, B=  on first roll, C=  on second roll

A and B are mutually exclusive, but A and C are not.

Addition rule: If A and B are mutually exclusive, then

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

Two events are *independent* if knowing that one occurs does not change the probability that the other occurs.

B and C are independent, but A and B are not.

Multiplication rule: If A and B are independent, then

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

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

$P(\text{at least one } \square\square \text{ in three rolls})$

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

$$P(\text{at least one } \square\square \text{ in three rolls})$$

$$= 1 - P(\text{no } \square\square \text{ in three rolls})$$

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

$P(\text{at least one } \square\square \text{ in three rolls})$

$= 1 - P(\text{no } \square\square \text{ in three rolls})$

$= 1 - P((\text{no } \square\square \text{ in first roll}) \text{ and } (\text{no } \square\square \text{ in second roll}) \text{ and } (\text{no } \square\square \text{ in third roll}))$

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

$$P(\text{at least one } \square\square \text{ in three rolls})$$

$$= 1 - P(\text{no } \square\square \text{ in three rolls})$$

$$= 1 - P((\text{no } \square\square \text{ in first roll}) \text{ and } (\text{no } \square\square \text{ in second roll}) \text{ and } (\text{no } \square\square \text{ in third roll}))$$

$$= 1 - P(\text{no } \square\square \text{ in first roll}) \times P(\text{no } \square\square \text{ in second roll}) \times P(\text{no } \square\square \text{ in third roll})$$

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

$$P(\text{at least one } \square\square \text{ in three rolls})$$

$$= 1 - P(\text{no } \square\square \text{ in three rolls})$$

$$= 1 - P((\text{no } \square\square \text{ in first roll}) \text{ and } (\text{no } \square\square \text{ in second roll}) \text{ and } (\text{no } \square\square \text{ in third roll}))$$

$$= 1 - P(\text{no } \square\square \text{ in first roll}) \times P(\text{no } \square\square \text{ in second roll}) \times P(\text{no } \square\square \text{ in third roll})$$

$$= 1 - \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$$

## Four basic rules

Roll a die three times. What is  $P(\text{at least one } \square\square)$  ?

We could write 'at least one  $\square\square$ ' as follows:

$\square\square$  on the first roll    or     $\square\square$  on the second roll    or     $\square\square$  on the third roll.

But these three events are **not** mutually exclusive, so we can't use the addition rule.

Better way: Start with the complement rule.

$$P(\text{at least one } \square\square \text{ in three rolls})$$

$$= 1 - P(\text{no } \square\square \text{ in three rolls})$$

$$= 1 - P((\text{no } \square\square \text{ in first roll}) \text{ and } (\text{no } \square\square \text{ in second roll}) \text{ and } (\text{no } \square\square \text{ in third roll}))$$

$$= 1 - P(\text{no } \square\square \text{ in first roll}) \times P(\text{no } \square\square \text{ in second roll}) \times P(\text{no } \square\square \text{ in third roll})$$

$$= 1 - \frac{5}{6} \times \frac{5}{6} \times \frac{5}{6}$$

$$= 41.1\%$$

## Conditional probability

Spam e-mail has a higher chance to contain the word ‘money’ than ham e-mail:

## Conditional probability

Spam e-mail has a higher chance to contain the word ‘money’ than ham e-mail:

$$P(\text{‘money’ in e-mail} \mid \text{spam}) = 8\%, \quad P(\text{‘money’ in e-mail} \mid \text{ham}) = 1\%.$$

## Conditional probability

Spam e-mail has a higher chance to contain the word ‘money’ than ham e-mail:

$$P(\text{‘money’ in e-mail} \mid \text{spam}) = 8\%, \quad P(\text{‘money’ in e-mail} \mid \text{ham}) = 1\%.$$

The *conditional probability of B given A* is

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

## Conditional probability

Spam e-mail has a higher chance to contain the word ‘money’ than ham e-mail:

$$P(\text{‘money’ in e-mail} \mid \text{spam}) = 8\%, \quad P(\text{‘money’ in e-mail} \mid \text{ham}) = 1\%.$$

The *conditional probability of B given A* is

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

General multiplication rule:  $P(A \text{ and } B) = P(A) P(B|A)$

## Conditional probability

Spam e-mail has a higher chance to contain the word ‘money’ than ham e-mail:

$$P(\text{‘money’ in e-mail} \mid \text{spam}) = 8\%, \quad P(\text{‘money’ in e-mail} \mid \text{ham}) = 1\%.$$

The *conditional probability of B given A* is

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

General multiplication rule:  $P(A \text{ and } B) = P(A)P(B|A)$

In the special case where A and B are independent:  $P(A \text{ and } B) = P(A)P(B)$

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

Idea is to artificially introduce the event 'spam/ham'.

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam      or      money appears and e-mail is ham

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam      or      money appears and e-mail is ham

$P(\text{money appears})$

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam      or      money appears and e-mail is ham

$P(\text{money appears})$

$= P(\text{money and spam}) + P(\text{money and ham})$

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam      or      money appears and e-mail is ham

$P(\text{money appears})$

$= P(\text{money and spam}) + P(\text{money and ham})$

$= P(\text{money} \mid \text{spam}) P(\text{spam}) + P(\text{money} \mid \text{ham}) P(\text{ham})$

## Conditional probability

Computing probabilities by total enumeration:

$P(\text{spam}) = 20\%$ . What is the probability that 'money' appears in an e-mail?

From data we know  $P(\text{money} \mid \text{spam}) = 8\%$ ,  $P(\text{money} \mid \text{ham}) = 1\%$ .

Idea is to artificially introduce the event 'spam/ham'.

The event 'money appears in the e-mail' can be written as:

money appears and e-mail is spam      or      money appears and e-mail is ham

$$P(\text{money appears})$$

$$= P(\text{money and spam}) + P(\text{money and ham})$$

$$= P(\text{money} \mid \text{spam}) P(\text{spam}) + P(\text{money} \mid \text{ham}) P(\text{ham})$$

$$= 0.08 \times 0.2 + 0.01 \times 0.8$$

$$= 2.4\%$$

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

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

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

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

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

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

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

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

$$P(\text{spam} \mid \text{money}) = \frac{P(\text{money} \mid \text{spam}) P(\text{spam})}{P(\text{money})}$$

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

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

$$P(\text{spam} \mid \text{money}) = \frac{P(\text{money} \mid \text{spam}) P(\text{spam})}{P(\text{money})} = \frac{0.08 \times 0.2}{0.024} = 67\%$$

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

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

$$P(\text{spam} \mid \text{money}) = \frac{P(\text{money} \mid \text{spam}) P(\text{spam})}{P(\text{money})} = \frac{0.08 \times 0.2}{0.024} = 67\%$$

Bayes' rule:

$$P(B|A) = \frac{P(A|B) P(B)}{P(A)}$$

## Bayes' rule

From data we know  $P(\text{money appears in e-mail} \mid \text{e-mail is spam}) = 8\%$ , but what we need to build a spam filter is  $P(\text{e-mail is spam} \mid \text{money appears in e-mail})$ .

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

$$P(\text{spam} \mid \text{money}) = \frac{P(\text{money} \mid \text{spam}) P(\text{spam})}{P(\text{money})} = \frac{0.08 \times 0.2}{0.024} = 67\%$$

Bayes' rule:

$$\begin{aligned} P(B|A) &= \frac{P(A|B) P(B)}{P(A)} \\ &= \frac{P(A|B) P(B)}{P(A|B)P(B) + P(A|\text{not } B) P(\text{not } B)} \end{aligned}$$

## Bayesian analysis

The spam filter classifies e-mail as spam via a *Bayesian analysis*:

## Bayesian analysis

The spam filter classifies e-mail as spam via a *Bayesian analysis*:

- ▶ Before examining the e-mail, there is a *prior probability* of 20% that it is spam.

## Bayesian analysis

The spam filter classifies e-mail as spam via a *Bayesian analysis*:

- ▶ Before examining the e-mail, there is a *prior probability* of 20% that it is spam.
- ▶ After examining the e-mail for certain keywords such as 'money', the filter updates this prior probability using Bayes' rule to arrive at the *posterior probability* that the e-mail is spam.

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive.

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know  $P(D) = 1\%$

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know  $P(D) = 1\%$  ,  $P(+|D) = 95\%$

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know  $P(D) = 1\%$  ,  $P(+|D) = 95\%$  ,  $P(+|\text{no } D) = 2\%$ .

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know  $P(D) = 1\%$  ,  $P(+|D) = 95\%$  ,  $P(+|\text{no } D) = 2\%$ .

Want  $P(D|+)$

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know  $P(D) = 1\%$  ,  $P(+|D) = 95\%$  ,  $P(+|\text{no } D) = 2\%$ .

$$\text{Want } P(D|+) = \frac{P(+|D) P(D)}{P(+)}$$

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know  $P(D) = 1\%$  ,  $P(+|D) = 95\%$  ,  $P(+|\text{no } D) = 2\%$ .

$$\begin{aligned}\text{Want } P(D|+) &= \frac{P(+|D) P(D)}{P(+)} \\ &= \frac{P(+|D) P(D)}{P(+|D) P(D) + P(+|\text{no } D) P(\text{no } D)}\end{aligned}$$

## Examples and case studies: False positives

1% of the population has a certain disease. If an infected person is tested, then there is a 95% chance that the test is positive. If the person is not infected, then there is a 2% chance that the test gives an erroneous positive result ('false positive').

Given that a person tests positive, what are the chances that he has the disease?

Know  $P(D) = 1\%$  ,  $P(+|D) = 95\%$  ,  $P(+|\text{no } D) = 2\%$ .

$$\begin{aligned}\text{Want } P(D|+) &= \frac{P(+|D) P(D)}{P(+)} \\ &= \frac{P(+|D) P(D)}{P(+|D) P(D) + P(+|\text{no } D) P(\text{no } D)} \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.02 \times 0.99} = 32.4\%\end{aligned}$$

## Case study: Warner's randomized response model

What percentage of students have cheated during an exam in college?

## Case study: Warner's randomized response model

What percentage of students have cheated during an exam in college?

Problem: Students may be too embarrassed to answer truthfully.

## Case study: Warner's randomized response model

What percentage of students have cheated during an exam in college?

Problem: Students may be too embarrassed to answer truthfully.

Randomization comes to the rescue:

We do a survey that first instructs students to toss a coin twice. If the student gets 'tails' on the first toss, then the student has to answer question 1, otherwise the student answers question 2.

## Case study: Warner's randomized response model

What percentage of students have cheated during an exam in college?

Problem: Students may be too embarrassed to answer truthfully.

Randomization comes to the rescue:

We do a survey that first instructs students to toss a coin twice. If the student gets 'tails' on the first toss, then the student has to answer question 1, otherwise the student answers question 2.

Q1: Have you ever cheated on an exam in college?

Q2: Did you get 'tails' on the second toss?

## Case study: Warner's randomized response model

What percentage of students have cheated during an exam in college?

Problem: Students may be too embarrassed to answer truthfully.

Randomization comes to the rescue:

We do a survey that first instructs students to toss a coin twice. If the student gets 'tails' on the first toss, then the student has to answer question 1, otherwise the student answers question 2.

Q1: Have you ever cheated on an exam in college?

Q2: Did you get 'tails' on the second toss?

So the answer will be partly random: We don't know whether a 'yes' answer is due to the student cheating or to getting tails on the second toss. This should put the student at ease to answer truthfully.

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$P(\text{yes}) = P(\text{yes and Q1}) + P(\text{yes and Q2})$$

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$\begin{aligned} P(\text{yes}) &= P(\text{yes and Q1}) + P(\text{yes and Q2}) \\ &= P(\text{yes} \mid \text{Q1}) P(\text{Q1}) + P(\text{yes} \mid \text{Q2}) P(\text{Q2}) \end{aligned}$$

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$\begin{aligned} P(\text{yes}) &= P(\text{yes and Q1}) + P(\text{yes and Q2}) \\ &= P(\text{yes} \mid \text{Q1}) P(\text{Q1}) + P(\text{yes} \mid \text{Q2}) P(\text{Q2}) \end{aligned}$$

$$\text{Solve for } P(\text{yes} \mid \text{Q1}) = \frac{P(\text{yes}) - P(\text{yes} \mid \text{Q2}) P(\text{Q2})}{P(\text{Q1})}$$

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$\begin{aligned} P(\text{yes}) &= P(\text{yes and Q1}) + P(\text{yes and Q2}) \\ &= P(\text{yes} \mid \text{Q1}) P(\text{Q1}) + P(\text{yes} \mid \text{Q2}) P(\text{Q2}) \end{aligned}$$

$$\text{Solve for } P(\text{yes} \mid \text{Q1}) = \frac{P(\text{yes}) - P(\text{yes} \mid \text{Q2}) P(\text{Q2})}{P(\text{Q1})}$$

In one survey, 27 students answered 'yes' and 30 answered 'no'.

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$\begin{aligned} P(\text{yes}) &= P(\text{yes and Q1}) + P(\text{yes and Q2}) \\ &= P(\text{yes} \mid \text{Q1}) P(\text{Q1}) + P(\text{yes} \mid \text{Q2}) P(\text{Q2}) \end{aligned}$$

$$\text{Solve for } P(\text{yes} \mid \text{Q1}) = \frac{P(\text{yes}) - P(\text{yes} \mid \text{Q2}) P(\text{Q2})}{P(\text{Q1})}$$

In one survey, 27 students answered 'yes' and 30 answered 'no'.

$$\text{So we estimate } P(\text{yes}) = \frac{27}{27+30} = 47\%$$

Key point: While we don't know what an *individual* 'yes' means, we can estimate the proportion of cheaters using all the answers *collectively*:

$$\begin{aligned} P(\text{yes}) &= P(\text{yes and Q1}) + P(\text{yes and Q2}) \\ &= P(\text{yes} \mid \text{Q1}) P(\text{Q1}) + P(\text{yes} \mid \text{Q2}) P(\text{Q2}) \end{aligned}$$

$$\text{Solve for } P(\text{yes} \mid \text{Q1}) = \frac{P(\text{yes}) - P(\text{yes} \mid \text{Q2}) P(\text{Q2})}{P(\text{Q1})}$$

In one survey, 27 students answered 'yes' and 30 answered 'no'.

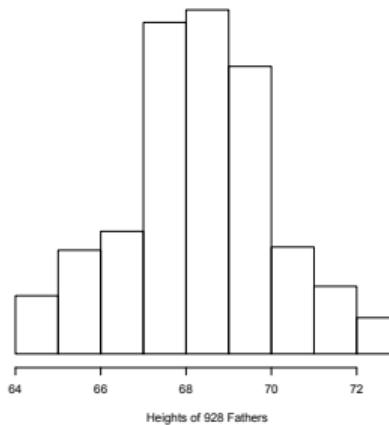
$$\text{So we estimate } P(\text{yes}) = \frac{27}{27+30} = 47\% \text{ and get } P(\text{yes} \mid \text{Q1}) = \frac{0.47 - 0.5 \times 0.5}{0.5} = 44\%$$

## Mini quiz

- ▶ A fair coin is tossed 5 times. Find the probability of getting at most 4 tails.
- ▶ 3% of all applicants to the Stanford Medical School are admitted. 70% of all applicants have a GPA of 3.6 or above. Of those who are admitted, 95% have a GPA of 3.6 or above.  
What are the chances of being admitted if the GPA is 3.6 or above?

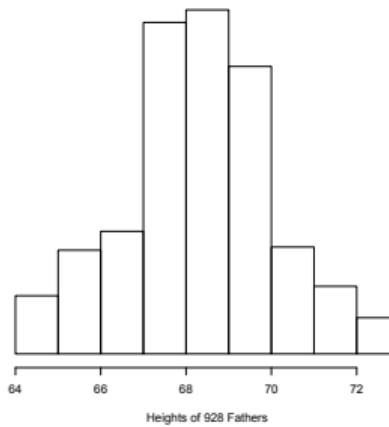
## The normal curve

Many data have histograms that look bell-shaped, e.g. heights, weights, IQ scores:



## The normal curve

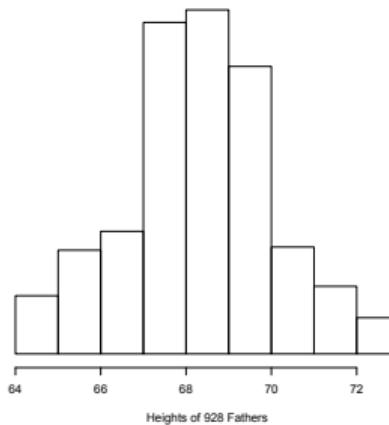
Many data have histograms that look bell-shaped, e.g. heights, weights, IQ scores:



'The data follow the normal curve.'

## The normal curve

Many data have histograms that look bell-shaped, e.g. heights, weights, IQ scores:



'The data follow the normal curve.'

But remember that some data have histograms that look quite different, e.g. incomes, house prices.

## The empirical rule

If the data follow the normal curve, then

## The empirical rule

If the data follow the normal curve, then

- ▶ about 2/3 (68%) of the data fall within one standard deviation of the mean

## The empirical rule

If the data follow the normal curve, then

- ▶ about 2/3 (68%) of the data fall within one standard deviation of the mean
- ▶ about 95% fall within 2 standard deviations of the mean

## The empirical rule

If the data follow the normal curve, then

- ▶ about 2/3 (68%) of the data fall within one standard deviation of the mean
- ▶ about 95% fall within 2 standard deviations of the mean
- ▶ about 99.7% fall within 3 standard deviations of the mean

## The empirical rule

If the data follow the normal curve, then

- ▶ about 2/3 (68%) of the data fall within one standard deviation of the mean
- ▶ about 95% fall within 2 standard deviations of the mean
- ▶ about 99.7% fall within 3 standard deviations of the mean

Galton's measurements of heights of fathers have  $\bar{x} = 68.3$  in and  $s = 1.8$  in.

Therefore about 95% of all heights are between  $68.3$  in  $- 2 \times 1.8$  in =  $64.7$  in and  $68.3$  in  $+ 2 \times 1.8$  in =  $71.9$  in.

## The empirical rule

Recall that in a histogram, percentages are given by areas:

## Standardizing data

A normal curve is determined by  $\bar{x}$  and  $s$ : If the data follow the normal curve, then knowing  $\bar{x}$  and  $s$  means knowing the whole histogram.

## Standardizing data

A normal curve is determined by  $\bar{x}$  and  $s$ : If the data follow the normal curve, then knowing  $\bar{x}$  and  $s$  means knowing the whole histogram.

To compute areas under the normal curve, we first **standardize** the data by subtracting off  $\bar{x}$  and then dividing by  $s$ :

$$z = \frac{\text{height} - \bar{x}}{s}$$

## Standardizing data

A normal curve is determined by  $\bar{x}$  and  $s$ : If the data follow the normal curve, then knowing  $\bar{x}$  and  $s$  means knowing the whole histogram.

To compute areas under the normal curve, we first **standardize** the data by subtracting off  $\bar{x}$  and then dividing by  $s$ :

$$z = \frac{\text{height} - \bar{x}}{s}$$

$z$  is called the **standardized value** or **z-score**.

## Standardizing data

A normal curve is determined by  $\bar{x}$  and  $s$ : If the data follow the normal curve, then knowing  $\bar{x}$  and  $s$  means knowing the whole histogram.

To compute areas under the normal curve, we first **standardize** the data by subtracting off  $\bar{x}$  and then dividing by  $s$ :

$$z = \frac{\text{height} - \bar{x}}{s}$$

$z$  is called the **standardized value** or **z-score**.

$z$  has no unit (height,  $\bar{x}$  and  $s$  all have the unit 'inches')

## Standardizing data

A normal curve is determined by  $\bar{x}$  and  $s$ : If the data follow the normal curve, then knowing  $\bar{x}$  and  $s$  means knowing the whole histogram.

To compute areas under the normal curve, we first **standardize** the data by subtracting off  $\bar{x}$  and then dividing by  $s$ :

$$z = \frac{\text{height} - \bar{x}}{s}$$

$z$  is called the **standardized value** or **z-score**.

$z$  has no unit (height,  $\bar{x}$  and  $s$  all have the unit 'inches')

For example,  $z = 2$  means the height is 2 standard deviations above average.

## Standardizing data

A normal curve is determined by  $\bar{x}$  and  $s$ : If the data follow the normal curve, then knowing  $\bar{x}$  and  $s$  means knowing the whole histogram.

To compute areas under the normal curve, we first **standardize** the data by subtracting off  $\bar{x}$  and then dividing by  $s$ :

$$z = \frac{\text{height} - \bar{x}}{s}$$

$z$  is called the **standardized value** or **z-score**.

$z$  has no unit (height,  $\bar{x}$  and  $s$  all have the unit 'inches')

For example,  $z = 2$  means the height is 2 standard deviations above average.

$z = -1.5$  means the height is 1.5 standard deviations *below* average.

## Random variables

You play an online game 10 times. Each time there is a 20% chance to win. The outcomes of the 10 games are due to chance, so the number of wins is random:

## Random variables

You play an online game 10 times. Each time there is a 20% chance to win. The outcomes of the 10 games are due to chance, so the number of wins is random: One set of 10 games might result in 4 wins, another set might result in 7 wins.

## Random variables

You play an online game 10 times. Each time there is a 20% chance to win. The outcomes of the 10 games are due to chance, so the number of wins is random: One set of 10 games might result in 4 wins, another set might result in 7 wins.

$X$  = 'number of successes' is called a **random variable**.

## Random variables

You play an online game 10 times. Each time there is a 20% chance to win. The outcomes of the 10 games are due to chance, so the number of wins is random: One set of 10 games might result in 4 wins, another set might result in 7 wins.

$X$  = 'number of successes' is called a **random variable**.

One can calculate  $P(X = 2) = 30.2\%$ .

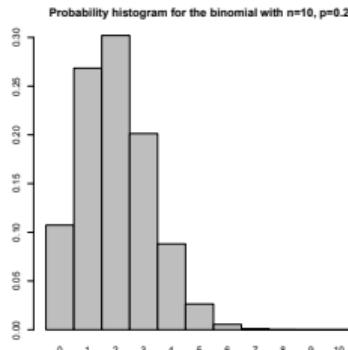
## Random variables

You play an online game 10 times. Each time there is a 20% chance to win. The outcomes of the 10 games are due to chance, so the number of wins is random: One set of 10 games might result in 4 wins, another set might result in 7 wins.

$X$  = 'number of successes' is called a **random variable**.

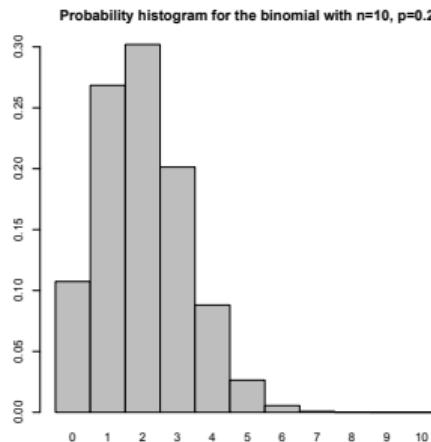
One can calculate  $P(X = 2) = 30.2\%$ .

We can visualize the probabilities of the various outcomes of  $X$  with a **probability histogram**:



## The probability histogram

We can visualize the probabilities of the various outcomes of  $X$  with a **probability histogram**:



A histogram of data gives percentages for observed data. In contrast, a probability histogram is a theoretical construct: it visualizes probabilities rather than data that have been empirically observed.

## Parameter and statistic

What is the average height of adult men in the US?

## Parameter and statistic

What is the average height of adult men in the US?

This is difficult to evaluate as there are 120 million adult men, but it can be *estimated* quite well with a relatively small sample.

## Parameter and statistic

What is the average height of adult men in the US?

This is difficult to evaluate as there are 120 million adult men, but it can be *estimated* quite well with a relatively small sample.

The **population** consists of all adult men in the US (about 120 million).

## Parameter and statistic

What is the average height of adult men in the US?

This is difficult to evaluate as there are 120 million adult men, but it can be *estimated* quite well with a relatively small sample.

The **population** consists of all adult men in the US (about 120 million).

A **parameter** is a quantity of interest about the population: the population average  $\mu$ , or the population standard deviation  $\sigma$ .

## Parameter and statistic

What is the average height of adult men in the US?

This is difficult to evaluate as there are 120 million adult men, but it can be *estimated* quite well with a relatively small sample.

The **population** consists of all adult men in the US (about 120 million).

A **parameter** is a quantity of interest about the population: the population average  $\mu$ , or the population standard deviation  $\sigma$ .

A **statistic (estimate)** is the quantity of interest as measured in the sample: the sample average  $\bar{x}$ , or the sample standard deviation  $s$ .

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

How about  $\bar{x}_n$ , the average of  $n$  draws?

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

How about  $\bar{x}_n$ , the average of  $n$  draws?

The **expected value of the sample average**,  $E(\bar{x}_n)$ , is the population average  $\mu$ .

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

How about  $\bar{x}_n$ , the average of  $n$  draws?

The **expected value of the sample average**,  $E(\bar{x}_n)$ , is the population average  $\mu$ .

But remember that  $\bar{x}_n$  is a random variable because sampling is a random process.

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

How about  $\bar{x}_n$ , the average of  $n$  draws?

The **expected value of the sample average**,  $E(\bar{x}_n)$ , is the population average  $\mu$ .

But remember that  $\bar{x}_n$  is a random variable because sampling is a random process.

So  $\bar{x}_n$  won't be exactly equal to  $\mu = 69.3$  in: We might get, say,  $\bar{x}_n = 70.1$  in. Taking another sample of size  $n$  might result in  $\bar{x}_n = 69.1$  in.

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

How about  $\bar{x}_n$ , the average of  $n$  draws?

The **expected value of the sample average**,  $E(\bar{x}_n)$ , is the population average  $\mu$ .

But remember that  $\bar{x}_n$  is a random variable because sampling is a random process.

So  $\bar{x}_n$  won't be exactly equal to  $\mu = 69.3$  in: We might get, say,  $\bar{x}_n = 70.1$  in. Taking another sample of size  $n$  might result in  $\bar{x}_n = 69.1$  in.

How far off from  $\mu$  will  $\bar{x}_n$  be?

## The expected value

If we sample an adult male at random, then we expect his height to be around the population average  $\mu$ , give or take about one standard deviation  $\sigma$ .

The **expected value** of one random draw is the population average  $\mu$ .

How about  $\bar{x}_n$ , the average of  $n$  draws?

The **expected value of the sample average**,  $E(\bar{x}_n)$ , is the population average  $\mu$ .

But remember that  $\bar{x}_n$  is a random variable because sampling is a random process.

So  $\bar{x}_n$  won't be exactly equal to  $\mu = 69.3$  in: We might get, say,  $\bar{x}_n = 70.1$  in. Taking another sample of size  $n$  might result in  $\bar{x}_n = 69.1$  in.

How far off from  $\mu$  will  $\bar{x}_n$  be?

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

## The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

## The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation  $\sigma$  plays for one observation drawn at random.

## The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation  $\sigma$  plays for one observation drawn at random.

The **square root law** is key for statistical inference:

$$\text{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$

## The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation  $\sigma$  plays for one observation drawn at random.

The **square root law** is key for statistical inference:

$$\text{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$

The importance of the square root law is twofold:

- ▶ It shows that the SE becomes smaller if we use a larger sample size  $n$ .

## The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation  $\sigma$  plays for one observation drawn at random.

The **square root law** is key for statistical inference:

$$\text{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$

The importance of the square root law is twofold:

- ▶ It shows that the SE becomes smaller if we use a larger sample size  $n$ . We can use the formula to determine what sample size is required for a desired accuracy.

## The standard error for the sample average

The **standard error (SE)** of a statistic tells roughly how far off the statistic will be from its expected value.

So the SE for a statistic plays the same role that the standard deviation  $\sigma$  plays for one observation drawn at random.

The **square root law** is key for statistical inference:

$$\text{SE}(\bar{x}_n) = \frac{\sigma}{\sqrt{n}}$$

The importance of the square root law is twofold:

- ▶ It shows that the SE becomes smaller if we use a larger sample size  $n$ . We can use the formula to determine what sample size is required for a desired accuracy.
- ▶ The formula for the standard error **does not depend on the size of the population**, only on the size of the sample.

## Supplement: Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

## Supplement: Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the **framework for counting and classifying**:

## Supplement: Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the **framework for counting and classifying**:

- ▶ The population consists of all likely voters (about 140 million).

## Supplement: Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the **framework for counting and classifying**:

- ▶ The population consists of all likely voters (about 140 million).
- ▶ Each likely voter falls into one of two categories: approve or not approve.

## Supplement: Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the **framework for counting and classifying**:

- ▶ The population consists of all likely voters (about 140 million).
- ▶ Each likely voter falls into one of two categories: approve or not approve.
- ▶ Put the label '1' on each likely voter who approves, and '0' on each who doesn't.

## Supplement: Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the **framework for counting and classifying**:

- ▶ The population consists of all likely voters (about 140 million).
- ▶ Each likely voter falls into one of two categories: approve or not approve.
- ▶ Put the label '1' on each likely voter who approves, and '0' on each who doesn't.
- ▶ Then the number of likely voters who approve equals the sum of all 140 million labels.

## Supplement: Expected value and standard error for percentages

What percentage of likely voters approve of the way the US President is handling his job?

The 'percentage of likely voters' is an average. This becomes clear by using the **framework for counting and classifying**:

- ▶ The population consists of all likely voters (about 140 million).
- ▶ Each likely voter falls into one of two categories: approve or not approve.
- ▶ Put the label '1' on each likely voter who approves, and '0' on each who doesn't.
- ▶ Then the number of likely voters who approve equals the sum of all 140 million labels.
- ▶ The percentage of likely voters who approve is the percentage of 1s among the labels, which is the average of the labels.

## Supplement: Expected value and standard error for percentages

In a sample of  $n$  likely voters

- ▶ the number of voters in the sample who are approving is the sum of the draws

## Supplement: Expected value and standard error for percentages

In a sample of  $n$  likely voters

- ▶ the number of voters in the sample who are approving is the sum of the draws
- ▶ the percentage of voters approving is the percentage of 1s, which is
$$\frac{\text{sum}}{n} \times 100\% = \bar{x}_n \times 100\%$$

## Supplement: Expected value and standard error for percentages

In a sample of  $n$  likely voters

- ▶ the number of voters in the sample who are approving is the sum of the draws
- ▶ the percentage of voters approving is the percentage of 1s, which is  
$$\frac{\text{sum}}{n} \times 100\% = \bar{x}_n \times 100\%$$

Therefore

$$E(\text{percentage of 1s}) = \mu \times 100\% \quad SE(\text{percentage of 1s}) = \frac{\sigma}{\sqrt{n}} \times 100\%$$

where  $\mu$  is the population average (=proportion of 1s) and  $\sigma$  is the standard deviation of the population of 0s and 1s.

## Supplement: Expected value and standard error for percentages

In a sample of  $n$  likely voters

- ▶ the number of voters in the sample who are approving is the sum of the draws
- ▶ the percentage of voters approving is the percentage of 1s, which is  
$$\frac{\text{sum}}{n} \times 100\% = \bar{x}_n \times 100\%$$

Therefore

$$E(\text{percentage of 1s}) = \mu \times 100\% \quad SE(\text{percentage of 1s}) = \frac{\sigma}{\sqrt{n}} \times 100\%$$

where  $\mu$  is the population average (=proportion of 1s) and  $\sigma$  is the standard deviation of the population of 0s and 1s.

All of the above formulas are for sampling with replacement. They are still approximately true when sampling without replacement if the sample size is much smaller than the size of the population.

## The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes:  
0, 1, 2, ..., 100.

## The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes:  
0, 1, 2, ..., 100.

How likely is each outcome?

## The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes:  
0, 1, 2, ..., 100.

How likely is each outcome?

The number of tails has the binomial distribution with  $n = 100$  and  $p = 0.5$ .

## The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes:  
0, 1, 2, ..., 100.

How likely is each outcome?

The number of tails has the binomial distribution with  $n = 100$  and  $p = 0.5$ .

So if the statistic of interest is  $S_n$ ='number of tails', then  $S_n$  is a random variable whose probability histogram is given by the binomial distribution.

## The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes:  
0, 1, 2, ..., 100.

How likely is each outcome?

The number of tails has the binomial distribution with  $n = 100$  and  $p = 0.5$ .

So if the statistic of interest is  $S_n$ ='number of tails', then  $S_n$  is a random variable whose probability histogram is given by the binomial distribution. This is called the **sampling distribution** of the statistic  $S_n$ .

## The sampling distribution

Toss a coin 100 times. The number of tails has the following possible outcomes:  
0, 1, 2, ..., 100.

How likely is each outcome?

The number of tails has the binomial distribution with  $n = 100$  and  $p = 0.5$ .

So if the statistic of interest is  $S_n$  = 'number of tails', then  $S_n$  is a random variable whose probability histogram is given by the binomial distribution. This is called the **sampling distribution** of the statistic  $S_n$ .

The sampling distribution of  $S_n$  provides more detailed information about the chance properties of  $S_n$  than the summary numbers given by the expected value and the standard error.

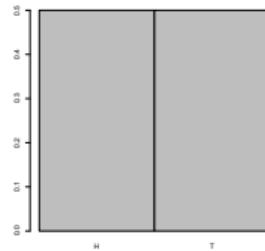
There are three histograms

The chance process of tossing a coin 100 times comes with three different histograms:

## There are three histograms

The chance process of tossing a coin 100 times comes with three different histograms:

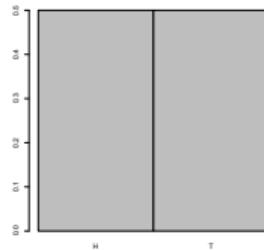
1. The probability histogram for producing the data:



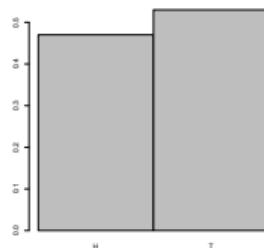
## There are three histograms

The chance process of tossing a coin 100 times comes with three different histograms:

1. The probability histogram for producing the data:

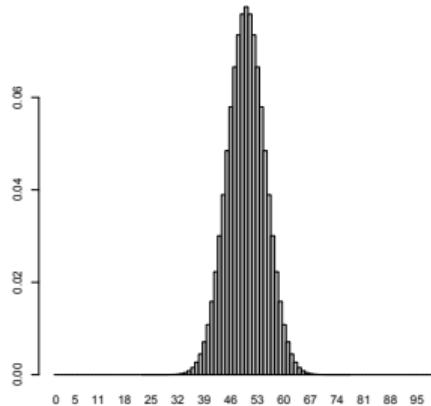


2. The histogram of the 100 observed tosses. This is an empirical histogram of real data:



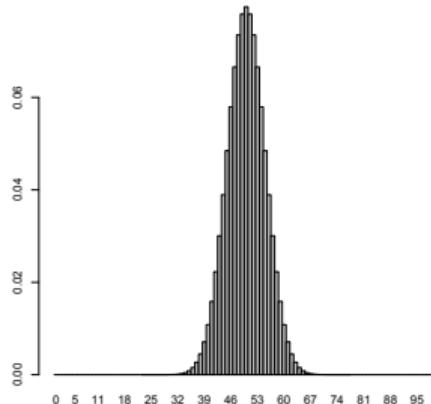
There are three histograms

3. The probability histogram of the statistic  $S_{100}$  = ‘number of tails’, which shows the sampling distribution of  $S_{100}$ :



There are three histograms

3. The probability histogram of the statistic  $S_{100}$  = ‘number of tails’, which shows the sampling distribution of  $S_{100}$ :



When doing statistical inference it is important to carefully distinguish these three histograms.

## The law of large numbers

The square root law says that  $SE(\bar{x}_n)$ , the standard error of the sample mean, goes to zero as the sample size increases.

## The law of large numbers

The square root law says that  $\text{SE}(\bar{x}_n)$ , the standard error of the sample mean, goes to zero as the sample size increases.

Therefore the sample mean  $\bar{x}_n$  will likely be close to its expected value  $\mu$  if the sample size is large. This is the **law of large numbers**.

## The law of large numbers

The square root law says that  $SE(\bar{x}_n)$ , the standard error of the sample mean, goes to zero as the sample size increases.

Therefore the sample mean  $\bar{x}_n$  will likely be close to its expected value  $\mu$  if the sample size is large. This is the **law of large numbers**.

Keep in mind that the law of large numbers applies

- ▶ for averages and therefore also for percentages, but not for sums as their SE *increases*

## The law of large numbers

The square root law says that  $\text{SE}(\bar{x}_n)$ , the standard error of the sample mean, goes to zero as the sample size increases.

Therefore the sample mean  $\bar{x}_n$  will likely be close to its expected value  $\mu$  if the sample size is large. This is the **law of large numbers**.

Keep in mind that the law of large numbers applies

- ▶ for averages and therefore also for percentages, but not for sums as their SE *increases*
- ▶ for sampling with replacement from a population, or for simulating data from a probability histogram

## The law of large numbers

The square root law says that  $\text{SE}(\bar{x}_n)$ , the standard error of the sample mean, goes to zero as the sample size increases.

Therefore the sample mean  $\bar{x}_n$  will likely be close to its expected value  $\mu$  if the sample size is large. This is the **law of large numbers**.

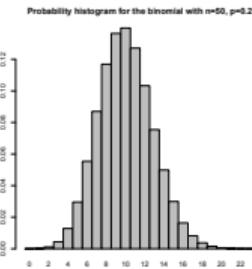
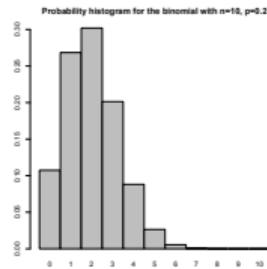
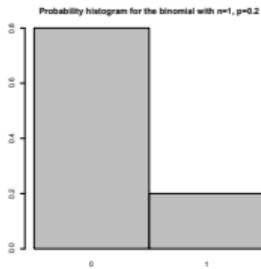
Keep in mind that the law of large numbers applies

- ▶ for averages and therefore also for percentages, but not for sums as their SE *increases*
- ▶ for sampling with replacement from a population, or for simulating data from a probability histogram

More advanced versions of the law of large numbers state that the empirical histogram of the data (the histogram in 2. in the previous section) will be close to the probability histogram in 1. if the sample size is large.

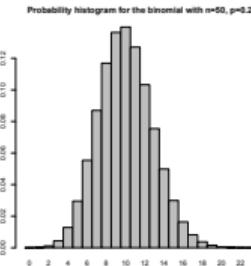
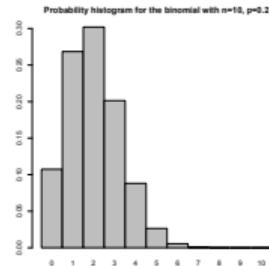
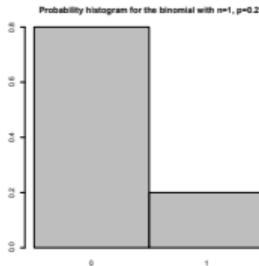
## The central limit theorem

Recall the online game where you win with probability 0.2. We looked at the random variable  $X$  = 'number of wins' in  $n$  gambles and found that  $X$  has the binomial distribution with that  $n$  and  $p = 0.2$ .



## The central limit theorem

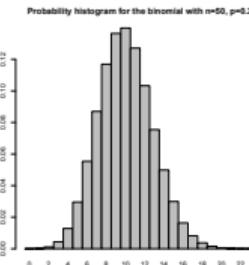
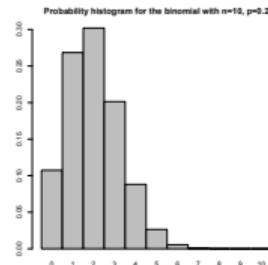
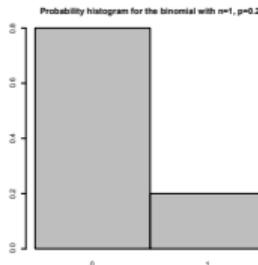
Recall the online game where you win with probability 0.2. We looked at the random variable  $X$  = 'number of wins' in  $n$  gambles and found that  $X$  has the binomial distribution with that  $n$  and  $p = 0.2$ .



As  $n$  gets large, the probability histogram looks more and more similar to the normal curve. This is an example of the **central limit theorem**:

## The central limit theorem

Recall the online game where you win with probability 0.2. We looked at the random variable  $X$  = 'number of wins' in  $n$  gambles and found that  $X$  has the binomial distribution with that  $n$  and  $p = 0.2$ .

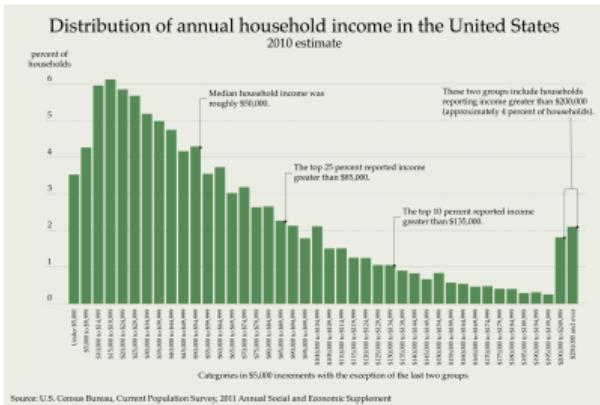


As  $n$  gets large, the probability histogram looks more and more similar to the normal curve. This is an example of the **central limit theorem**:

When sampling with replacement and  $n$  is large, then the sampling distribution of the sample average (or sum or percentage) approximately follows the normal curve. To standardize, subtract off the expected value of the statistic, then divide by its SE.

# The central limit theorem

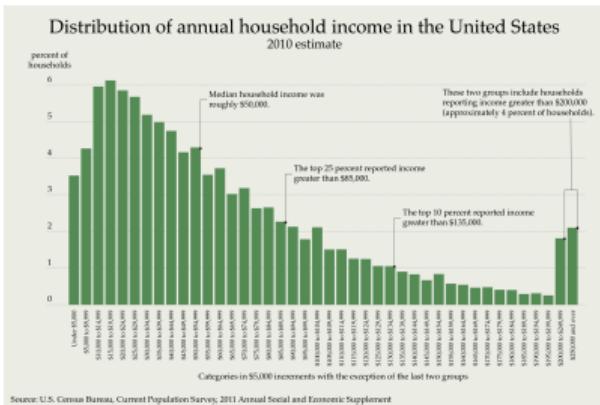
The key point of the theorem is that we know that the sampling distribution of the statistic is normal *no matter what the population histogram is:*



$$\mu = \$67,000$$
$$\sigma = \$38,000$$

## The central limit theorem

The key point of the theorem is that we know that the sampling distribution of the statistic is normal *no matter what the population histogram is*:



$$\mu = \$67,000$$
$$\sigma = \$38,000$$

If we sample  $n$  incomes at random, then the sample average  $\bar{x}_n$  follows the normal curve centered at  $E(\bar{x}_n) = \mu = \$67,000$  and with its spread given by

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38,000}{\sqrt{n}}$$

## The central limit theorem

If we sample  $n$  incomes at random, then the sample average  $\bar{x}_n$  follows the normal curve centered at  $E(\bar{x}_n) = \mu = \$67,000$  and with its spread given by

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38,000}{\sqrt{n}}.$$

## The central limit theorem

If we sample  $n$  incomes at random, then the sample average  $\bar{x}_n$  follows the normal curve centered at  $E(\bar{x}_n) = \mu = \$67,000$  and with its spread given by

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38,000}{\sqrt{n}}.$$

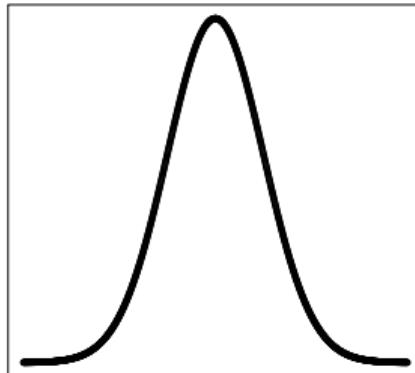
For example, if we sample 100 incomes, then by the empirical rule there is about a 16% chance that  $\bar{x}_n$  is larger than \$ 70,800:

## The central limit theorem

If we sample  $n$  incomes at random, then the sample average  $\bar{x}_n$  follows the normal curve centered at  $E(\bar{x}_n) = \mu = \$67,000$  and with its spread given by

$$SE(\bar{x}_n) = \frac{\sigma}{\sqrt{n}} = \frac{\$38,000}{\sqrt{n}}.$$

For example, if we sample 100 incomes, then by the empirical rule there is about a 16% chance that  $\bar{x}_n$  is larger than \$ 70,800:



## When does the central limit theorem apply?

For the normal approximation to work, the key requirements are:

- ▶ We sample with replacement, or we simulate independent random variables from the same distribution.

## When does the central limit theorem apply?

For the normal approximation to work, the key requirements are:

- ▶ We sample with replacement, or we simulate independent random variables from the same distribution.
- ▶ The statistic of interest is a sum (averages and percentages are sums in disguise).

## When does the central limit theorem apply?

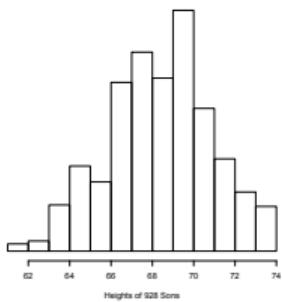
For the normal approximation to work, the key requirements are:

- ▶ We sample with replacement, or we simulate independent random variables from the same distribution.
- ▶ The statistic of interest is a sum (averages and percentages are sums in disguise).
- ▶ The sample size is large enough: the more skewed the population histogram is, the larger the required sample size  $n$ .  
(if there is no strong skewness then  $n \geq 15$  is sufficient)

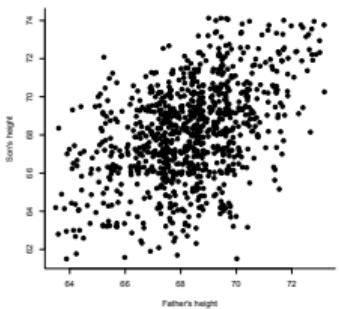
## Mini quiz

1. There are two candidates running for governor in CA and they are said to have roughly equal support from the voters. To get a better idea who is ahead, a company polls 400 of the 20 million registered voters in California. Likewise, there are two candidates running for mayor in Palo Alto who are said to have roughly equal support, and the company polls 400 out of the 20,000 registered voters in Palo Alto. Will the first poll be more/equal/less accurate than the second?
2. The average taxable income reported on tax returns for the year 2016 is \$ 45,000, and the standard deviation of the taxable incomes is \$ 23,000. For each of the following two statements, state whether it is true or false:
  - a. The percentage of taxable incomes that fall below \$ 30,000 can be computed from the above information using normal approximation.
  - b. The chances that the sum of 100 randomly selected taxable incomes exceeds \$ 4 million can be computed from the above information using normal approximation.

## Regression: Prediction is a key task of statistics

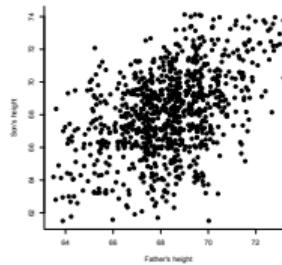
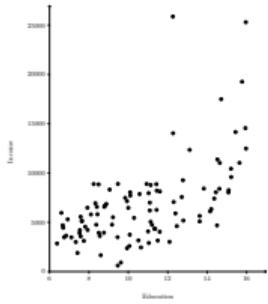


Predict the height of a son who is chosen at random from 928 sons. The average height of sons, 68.1 in, is the 'best' predictor.



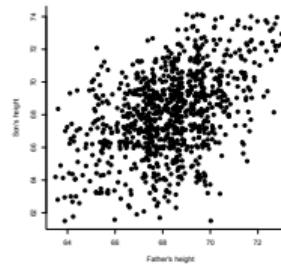
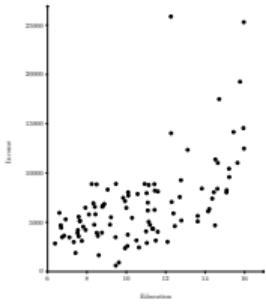
Predict the height of a son whose father is 72 in tall. This additional information about the father should allow us to make a better prediction. Regression does just that.

# The correlation coefficient



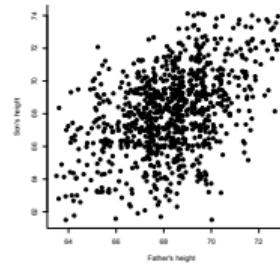
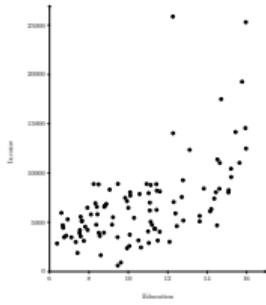
The scatterplot visualizes the relationship between two quantitative variables.

# The correlation coefficient



The scatterplot visualizes the relationship between two quantitative variables. It may have a direction (sloping up or down), form (a scatter that clusters around a line is called *linear*) and strength (how closely do the points follow the form?).

# The correlation coefficient



The scatterplot visualizes the relationship between two quantitative variables. It may have a direction (sloping up or down), form (a scatter that clusters around a line is called *linear*) and strength (how closely do the points follow the form?).

If the form is linear, then a good measure of strength is the **correlation coefficient  $r$** : Our data are  $(x_i, y_i)$ ,  $i = 1, \dots, n$ .

$$r = \frac{1}{n} \sum_{i=1}^n \frac{x_i - \bar{x}}{s_x} \times \frac{y_i - \bar{y}}{s_y}$$

(divide by  $n - 1$  instead of  $n$  if this is also done for the standard deviations  $s_x, s_y$ ).

## Correlation measures linear association

A numerical summary of these pairs of data is given by:  $\bar{x}, s_x, \bar{y}, s_y, r$ .

## Correlation measures linear association

A numerical summary of these pairs of data is given by:  $\bar{x}, s_x, \bar{y}, s_y, r$ .

As a convention the variable on the horizontal axis is called **explanatory variable** or **predictor**, the one on the vertical axis is called **response variable**.

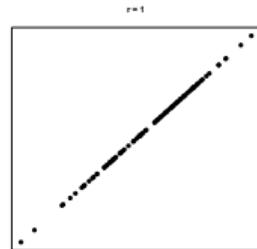
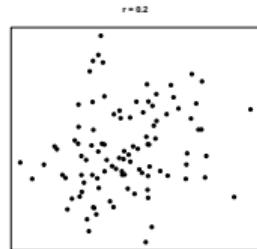
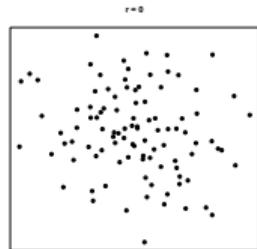
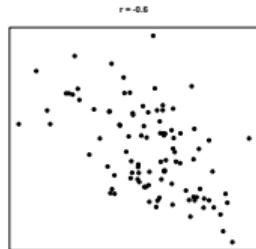
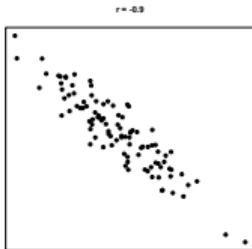
$r$  is always between  $-1$  and  $1$ . The sign of  $r$  gives the direction of the association and its absolute value gives the strength:

## Correlation measures linear association

A numerical summary of these pairs of data is given by:  $\bar{x}, s_x, \bar{y}, s_y, r$ .

As a convention the variable on the horizontal axis is called **explanatory variable** or **predictor**, the one on the vertical axis is called **response variable**.

$r$  is always between  $-1$  and  $1$ . The sign of  $r$  gives the direction of the association and its absolute value gives the strength:

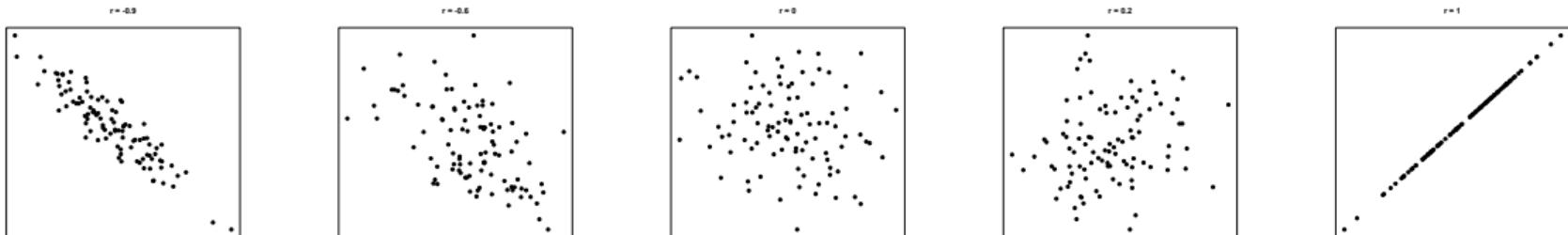


## Correlation measures linear association

A numerical summary of these pairs of data is given by:  $\bar{x}, s_x, \bar{y}, s_y, r$ .

As a convention the variable on the horizontal axis is called **explanatory variable** or **predictor**, the one on the vertical axis is called **response variable**.

$r$  is always between  $-1$  and  $1$ . The sign of  $r$  gives the direction of the association and its absolute value gives the strength:



Since both  $x$  and  $y$  were standardized when computing  $r$ ,  $r$  has no units and is not affected by changing the center or the scale of either variable.

## Correlation measures linear association

Keep in mind that  $r$  is only useful for measuring *linear* association:



$$r = 0$$

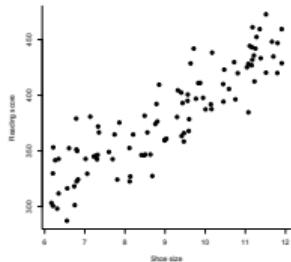
## Correlation measures linear association

Keep in mind that  $r$  is only useful for measuring *linear* association:



$$r = 0$$

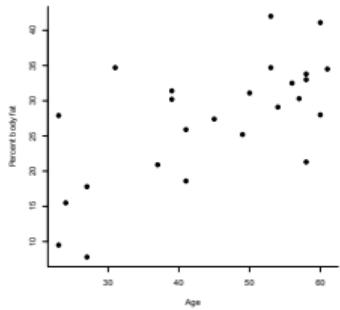
Also remember that correlation does not mean causation:



Among school children there is a high correlation between shoe size and reading ability. Both are driven by the *lurking variable* 'age'.

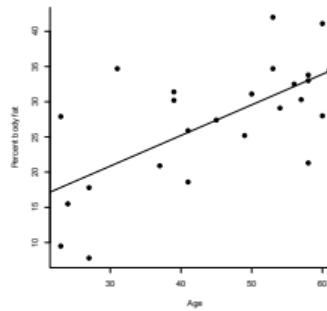
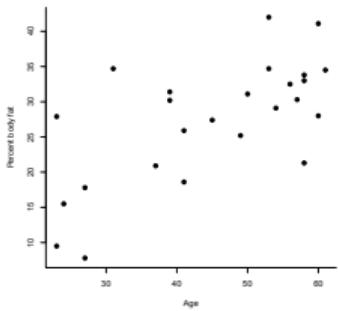
## The regression line

If the scatterplot shows a linear association, then this relationship can be summarized by a line.



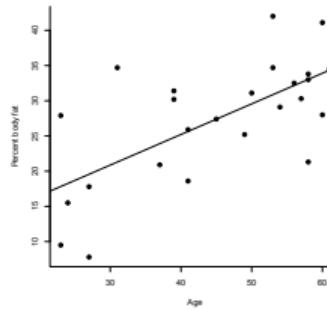
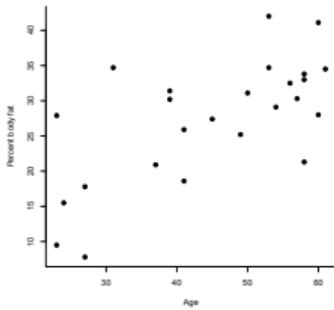
## The regression line

If the scatterplot shows a linear association, then this relationship can be summarized by a line.



## The regression line

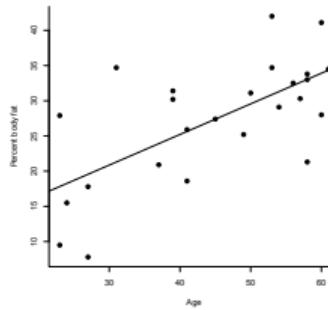
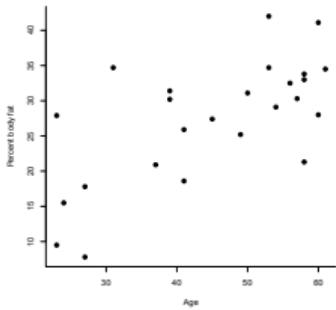
If the scatterplot shows a linear association, then this relationship can be summarized by a line.



To find this line for  $n$  pairs of data  $(x_1, y_1), \dots, (x_n, y_n)$ , recall that the equation of a line produces the y-value  $\hat{y}_i = a + bx_i$ .

## The regression line

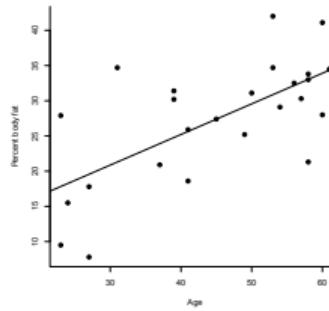
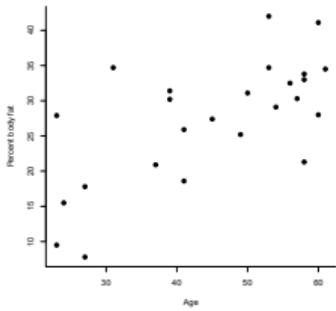
If the scatterplot shows a linear association, then this relationship can be summarized by a line.



To find this line for  $n$  pairs of data  $(x_1, y_1), \dots, (x_n, y_n)$ , recall that the equation of a line produces the y-value  $\hat{y}_i = a + bx_i$ . The idea is to choose the line that minimizes the sum of the squared distances between the observed  $y_i$  and the  $\hat{y}_i$ .

## The regression line

If the scatterplot shows a linear association, then this relationship can be summarized by a line.



To find this line for  $n$  pairs of data  $(x_1, y_1), \dots, (x_n, y_n)$ , recall that the equation of a line produces the y-value  $\hat{y}_i = a + bx_i$ . The idea is to choose the line that minimizes the sum of the squared distances between the observed  $y_i$  and the  $\hat{y}_i$ . In other words, find  $a$  and  $b$  that minimize

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

## The method of least squares

For  $n$  pairs of data  $(x_1, y_1), \dots, (x_n, y_n)$ , find  $a$  and  $b$  that minimize

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

## The method of least squares

For  $n$  pairs of data  $(x_1, y_1), \dots, (x_n, y_n)$ , find  $a$  and  $b$  that minimize

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

This is the **method of least squares**. It turns out that  $b = r \frac{s_y}{s_x}$  and  $a = \bar{y} - b\bar{x}$ . This line  $\hat{y} = a + bx$  is called the **regression line**.

## The method of least squares

For  $n$  pairs of data  $(x_1, y_1), \dots, (x_n, y_n)$ , find  $a$  and  $b$  that minimize

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2$$

This is the **method of least squares**. It turns out that  $b = r \frac{s_y}{s_x}$  and  $a = \bar{y} - b\bar{x}$ . This line  $\hat{y} = a + bx$  is called the **regression line**.

There is another interpretation of the regression line:  
it computes the average value of  $y$  when the first coordinate is near  $x$ .

Remember that often times an average is the 'best' predictor. This shows how the regression line incorporates the information given by  $x$  to produce a good predictor of  $y$ .

## Regression to the mean

The main use of regression is to predict  $y$  from  $x$ :

Given  $x$ , predict  $y$  to be  $\hat{y} = a + bx$ .

## Regression to the mean

The main use of regression is to predict  $y$  from  $x$ :

Given  $x$ , predict  $y$  to be  $\hat{y} = a + bx$ .

The prediction for  $y$  at  $x = \bar{x}$  is simply  $\hat{y} = \bar{y}$ .

## Regression to the mean

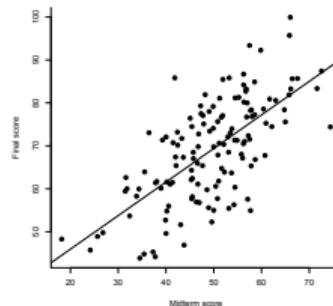
The main use of regression is to predict  $y$  from  $x$ :

Given  $x$ , predict  $y$  to be  $\hat{y} = a + bx$ .

The prediction for  $y$  at  $x = \bar{x}$  is simply  $\hat{y} = \bar{y}$ .

But  $b = r \frac{s_y}{s_x}$  means that if  $x$  is one standard deviation  $s_x$  above  $\bar{x}$ , then the predicted  $\hat{y}$  is only  $r s_y$  above  $\bar{y}$ .

Since  $r$  is between  $-1$  and  $1$ , the prediction is ‘towards the mean’:  $\hat{y}$  is fewer standard deviations away from  $\bar{y}$  than  $x$  is from  $\bar{x}$ .



## Regression to the mean

This is called **regression to the mean** (or: the **regression effect**). It can be observed in data whose scatter is football-shaped such as the exam scores: In such a test-retest situation, the top group on the test will drop down somewhat on the retest, while the bottom group moves up.

## Regression to the mean

This is called **regression to the mean** (or: the **regression effect**). It can be observed in data whose scatter is football-shaped such as the exam scores: In such a test-retest situation, the top group on the test will drop down somewhat on the retest, while the bottom group moves up.

A heuristic explanation is this: To score among the very top on the midterm requires excellent preparation as well as some luck. This luck may not be there any more on the final exam, and so we expect this group to fall back a bit.

## Regression to the mean

This is called **regression to the mean** (or: the **regression effect**). It can be observed in data whose scatter is football-shaped such as the exam scores: In such a test-retest situation, the top group on the test will drop down somewhat on the retest, while the bottom group moves up.

A heuristic explanation is this: To score among the very top on the midterm requires excellent preparation as well as some luck. This luck may not be there any more on the final exam, and so we expect this group to fall back a bit.

This effect is simply a consequence of there being a scatter around the line. Erroneously assuming that this occurs due to some action (e.g. ‘the top scorers on the midterm slackened off’) is the **regression fallacy**.

## Predicting $y$ from $x$ and $x$ from $y$

If we are given  $x$ , then we use the regression line  $\hat{y} = a + bx$  to predict  $y$ .  
To find this regression line we need only  $\bar{x}, \bar{y}, s_x, s_y$  and  $r$ .

## Predicting $y$ from $x$ and $x$ from $y$

If we are given  $x$ , then we use the regression line  $\hat{y} = a + bx$  to predict  $y$ .

To find this regression line we need only  $\bar{x}, \bar{y}, s_x, s_y$  and  $r$ .

The line can be computed with software, e.g. 'lm' in R, but it can also be done easily by hand.

## Predicting $y$ from $x$ and $x$ from $y$

If we are given  $x$ , then we use the regression line  $\hat{y} = a + bx$  to predict  $y$ .

To find this regression line we need only  $\bar{x}, \bar{y}, s_x, s_y$  and  $r$ .

The line can be computed with software, e.g. 'lm' in R, but it can also be done easily by hand.

When predicting  $x$  from  $y$  **it is a mistake** to use the regression line  $\hat{y} = a + bx$ , derived for regressing  $y$  on  $x$ , and solve for  $x$ . This is because regressing  $x$  on  $y$  will result in a **different regression line**.

## Predicting $y$ from $x$ and $x$ from $y$

If we are given  $x$ , then we use the regression line  $\hat{y} = a + bx$  to predict  $y$ .

To find this regression line we need only  $\bar{x}, \bar{y}, s_x, s_y$  and  $r$ .

The line can be computed with software, e.g. 'lm' in R, but it can also be done easily by hand.

When predicting  $x$  from  $y$  **it is a mistake** to use the regression line  $\hat{y} = a + bx$ , derived for regressing  $y$  on  $x$ , and solve for  $x$ . This is because regressing  $x$  on  $y$  will result in a **different regression line**.

To avoid confusing these, always put the predictor on the x-axis and proceed as on the previous slide.

## Normal approximation in regression

Regression requires that the scatter is football-shaped. Then one may use normal approximation for the  $y$ -values conditional on  $x$ . That is, the observations whose first coordinate is near that  $x$  have  $y$ -values that approximately follow the normal curve.

## Normal approximation in regression

Regression requires that the scatter is football-shaped. Then one may use normal approximation for the  $y$ -values conditional on  $x$ . That is, the observations whose first coordinate is near that  $x$  have  $y$ -values that approximately follow the normal curve.

To standardize, subtract off the predicted value  $\hat{y}$ , then divide by  $\sqrt{1 - r^2} \times s_y$ .

## Residuals

The differences between observed and predicted  $y$ -values are called **residuals**:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

## Residuals

The differences between observed and predicted y-values are called **residuals**:

$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

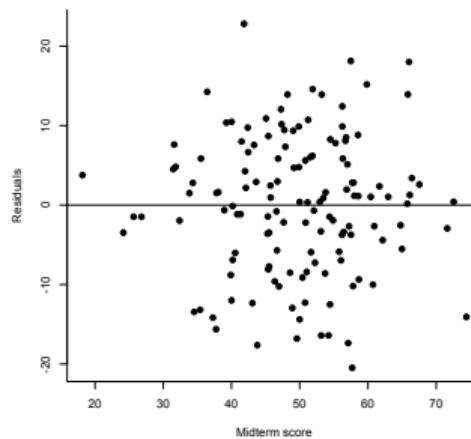
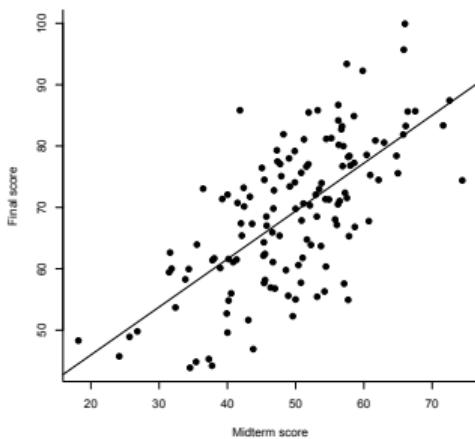
Residuals are used to check whether the use of regression is appropriate. The **residual plot** is a scatterplot of the residuals against the x-values. It should show an unstructured horizontal band.

## Residuals

The differences between observed and predicted  $y$ -values are called **residuals**:

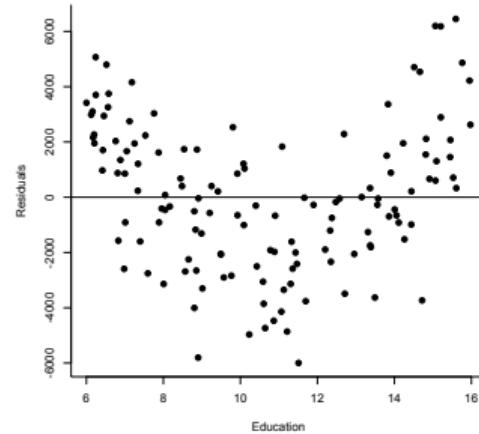
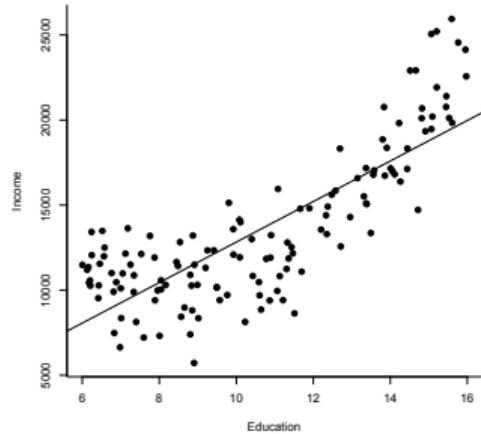
$$e_i = y_i - \hat{y}_i, \quad i = 1, \dots, n$$

Residuals are used to check whether the use of regression is appropriate. The **residual plot** is a scatterplot of the residuals against the  $x$ -values. It should show an unstructured horizontal band.



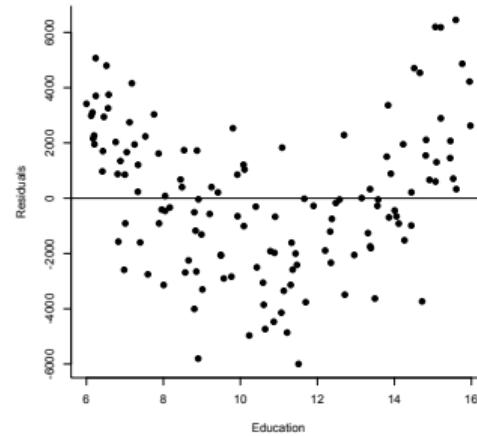
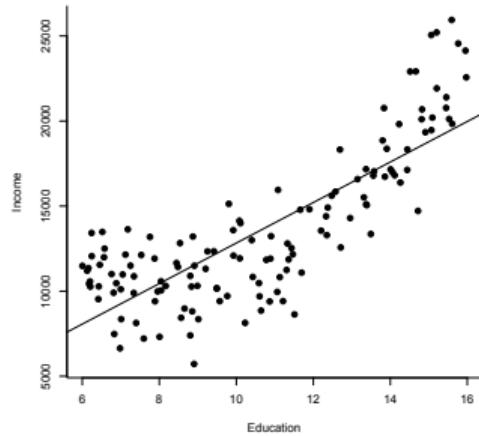
## Residual plots

A curved pattern suggests that the scatter is not linear:



## Residual plots

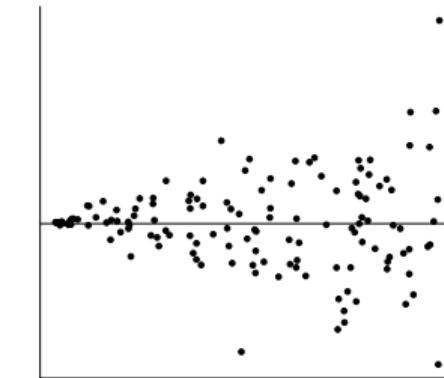
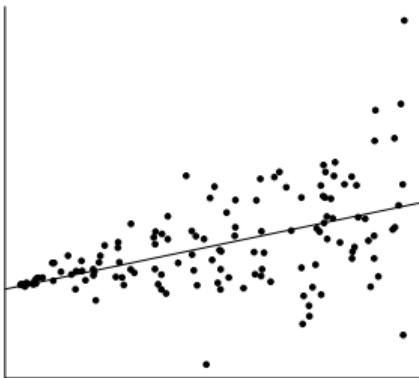
A curved pattern suggests that the scatter is not linear:



But it may still be possible to analyze these data with regression! Regression may be applicable after **transforming** the data, e.g. regress  $\sqrt{\text{income}}$  or  $\log(\text{income})$  on Education.

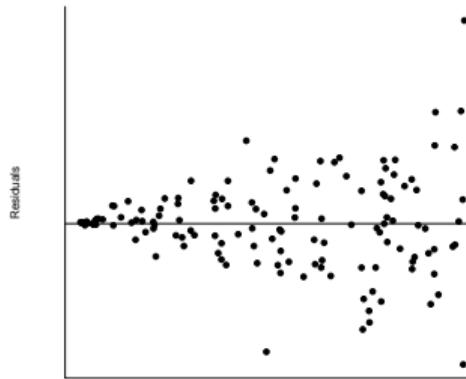
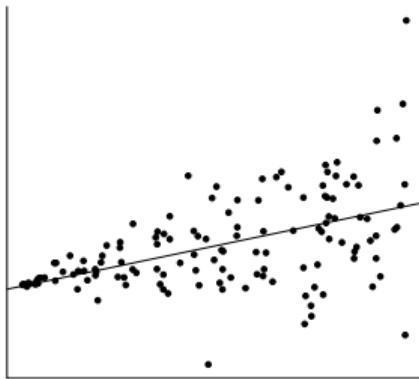
## Transformations of the variables

Another violation of the football-shaped assumption about the scatter arises if the scatter is **heteroscedastic**:



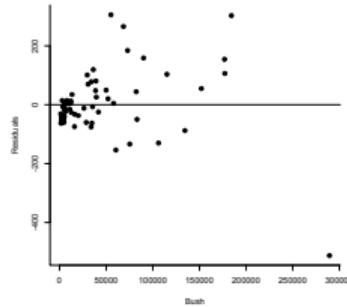
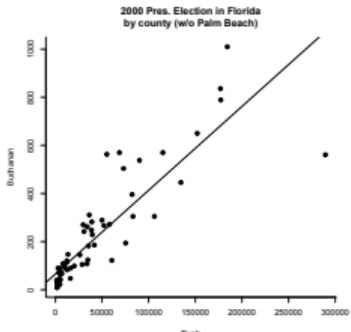
## Transformations of the variables

Another violation of the football-shaped assumption about the scatter arises if the scatter is **heteroscedastic**:



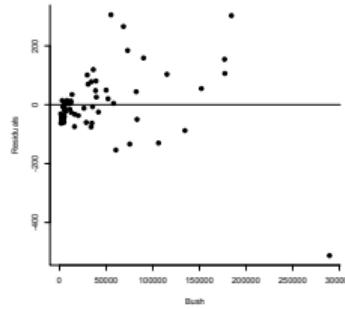
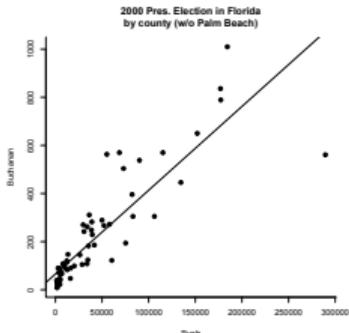
A transformation of the y-variables may produce a **homoscedastic** scatter, i.e. result in equal spread of the residuals across  $x$ . (However, it may also result in a non-linear scatter, which may require a second transformation of the x-values to fix!)

## Transformation of the variables

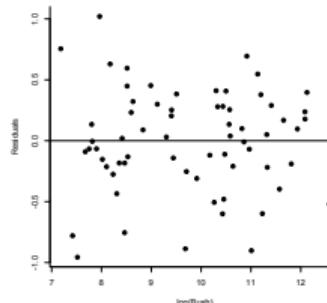
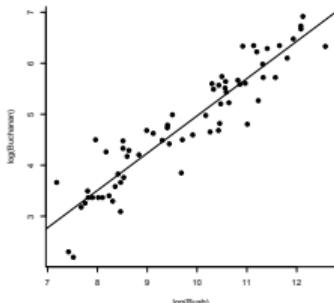


The residual plot looks heteroscedastic. Taking log of both variables produces a residual plot that is very satisfactory:

## Transformation of the variables

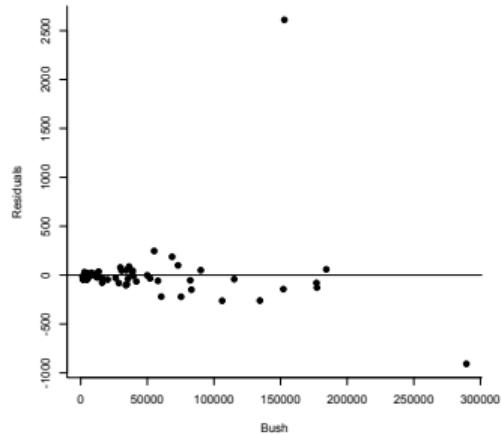
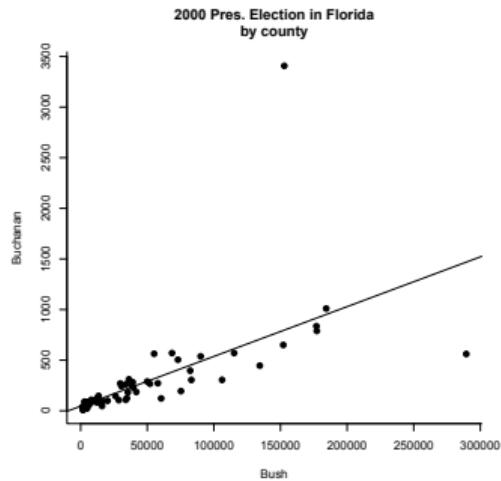


The residual plot looks heteroscedastic. Taking log of both variables produces a residual plot that is very satisfactory:



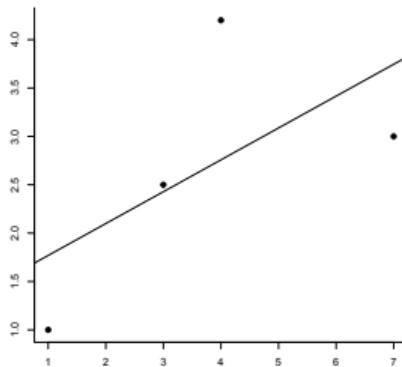
# Outliers

Points with very large residuals (**outliers**) should be examined: they may represent typos or interesting phenomena.



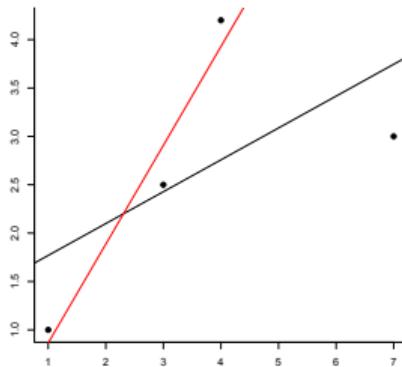
## Leverage and influential points

A point whose x-value is far from the mean of the x-values has high **leverage**: it has the potential to cause a big change the regression line.



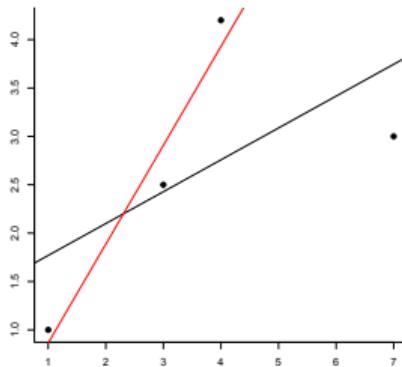
## Leverage and influential points

A point whose x-value is far from the mean of the x-values has high **leverage**: it has the potential to cause a big change the regression line.



## Leverage and influential points

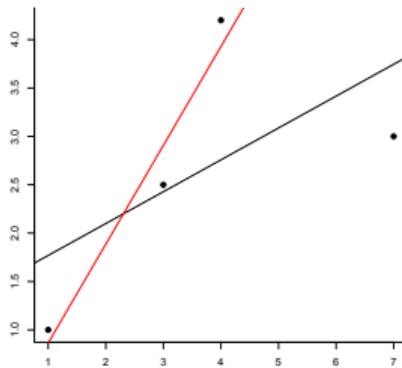
A point whose x-value is far from the mean of the x-values has high **leverage**: it has the potential to cause a big change the regression line.



Whether it does change the line a lot ( $\rightarrow$  **influential point**) or not can only be determined by refitting the regression without the point.

## Leverage and influential points

A point whose x-value is far from the mean of the x-values has high **leverage**: it has the potential to cause a big change the regression line.



Whether it does change the line a lot ( $\rightarrow$  **influential point**) or not can only be determined by refitting the regression without the point. An influential point may have a small residual (because it is influential!), so a residual plot is not helpful for this analysis.

## Some other issues

- ▶ Avoid predicting  $y$  by **extrapolation**, i.e. at  $x$ -values that are outside the range of the  $x$ -values that were used for the regression: The linear relationship often breaks down outside a certain range.

## Some other issues

- ▶ Avoid predicting  $y$  by **extrapolation**, i.e. at  $x$ -values that are outside the range of the  $x$ -values that were used for the regression: The linear relationship often breaks down outside a certain range.
- ▶ Beware of data that are summaries (e.g. averages of some data). Those are less variable than individual observations and correlations between averages tend to overstate the strength of the relationship.

## Some other issues

- ▶ Avoid predicting  $y$  by **extrapolation**, i.e. at  $x$ -values that are outside the range of the  $x$ -values that were used for the regression: The linear relationship often breaks down outside a certain range.
- ▶ Beware of data that are summaries (e.g. averages of some data). Those are less variable than individual observations and correlations between averages tend to overstate the strength of the relationship.
- ▶ Regression analyses often report ‘R-squared’:  $R^2 = r^2$ . It gives the fraction of the variation in the  $y$ -values that is explained by the regression line. (So  $1 - r^2$  is the fraction of the variation in the  $y$ -values that is left in the residuals.)

## Mini quiz

1. Some people believe that musical activity (e.g. playing an instrument) enhances mathematical ability. 100 high school students were selected at random. For each student, musical activity was recorded in hours per week and mathematical ability was assessed by a test. The correlation coefficient was found to be 0.85.
  - a. Does the large correlation coefficient prove that musical activity enhances mathematical ability?
  - b. What would your answer to a) be if you learned that all students in the study came from the same grade?
2. A tutoring center advertises its services by stating that students who sign up improve their GPA on tests by 0.5 points on average. Is this indeed evidence that the tutoring helps or could this be due to the regression effect?
3. True or false: If an observation with large leverage has a small residual, then it is not influential.

## Confidence intervals

Let's look at the Gallup poll on the approval rating of the US President.  
Suppose 60% of the 140 million likely voters approve of the way the president is handling his job.

## Confidence intervals

Let's look at the Gallup poll on the approval rating of the US President.

Suppose 60% of the 140 million likely voters approve of the way the president is handling his job.

Gallup polls 1,000 of them. The resulting approval percentage in the sample will be off the population percentage of 60% due to chance error. How much?

## Confidence intervals

Let's look at the Gallup poll on the approval rating of the US President. Suppose 60% of the 140 million likely voters approve of the way the president is handling his job. Gallup polls 1,000 of them. The resulting approval percentage in the sample will be off the population percentage of 60% due to chance error. How much? The SE tells us the likely size of the chance error.

## Confidence intervals

Let's look at the Gallup poll on the approval rating of the US President. Suppose 60% of the 140 million likely voters approve of the way the president is handling his job.

Gallup polls 1,000 of them. The resulting approval percentage in the sample will be off the population percentage of 60% due to chance error. How much? The SE tells us the likely size of the chance error.

Confidence intervals give a more precise statement.

## Confidence intervals

According to the central limit theorem, the sample percentage follows the normal curve with expected value  $\mu = 60\%$  and SE equal to  $\frac{\sigma}{\sqrt{1000}} = \frac{0.49}{\sqrt{1000}} = 1.6\%$ .

(Because we sample from a population of 140 million labels, of which 60% are 1s and 40% are 0s.)

## Confidence intervals

According to the central limit theorem, the sample percentage follows the normal curve with expected value  $\mu = 60\%$  and SE equal to  $\frac{\sigma}{\sqrt{1000}} = \frac{0.49}{\sqrt{1000}} = 1.6\%$ .

(Because we sample from a population of 140 million labels, of which 60% are 1s and 40% are 0s.)

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage  $\mu$ .

## Confidence intervals

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage  $\mu$ .

## Confidence intervals

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage  $\mu$ .

But that is the same as saying that the population percentage is no more than 2 SEs away from the sample percentage.

## Confidence intervals

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage  $\mu$ .

But that is the same as saying that the population percentage is no more than 2 SEs away from the sample percentage.

So once we compute the sample percentage from our sample, say 58%, then we can give a range of plausible values for the unknown population percentage by going 2 SEs in each direction:

## Confidence intervals

By the empirical rule there is a 95% chance that the sample percentage is no more than 2 SEs away from the population percentage  $\mu$ .

But that is the same as saying that the population percentage is no more than 2 SEs away from the sample percentage.

So once we compute the sample percentage from our sample, say 58%, then we can give a range of plausible values for the unknown population percentage by going 2 SEs in each direction:

[54.8%, 61.2%] is a 95% **confidence interval** for the population percentage.

## Interpretation of a confidence interval

[54.8%, 61.2%] is a 95% confidence interval for the population percentage.  
Why ‘confidence’ instead of ‘probability’ ?

## Interpretation of a confidence interval

[54.8%, 61.2%] is a 95% confidence interval for the population percentage.

Why ‘confidence’ instead of ‘probability’?

The population percentage  $\mu$  is a *fixed* number, which either falls into [54.8%, 61.2%] or not. There are no chances involved.

## Interpretation of a confidence interval

[54.8%, 61.2%] is a 95% confidence interval for the population percentage.

Why ‘confidence’ instead of ‘probability’?

The population percentage  $\mu$  is a *fixed* number, which either falls into [54.8%, 61.2%] or not. There are no chances involved.

Rather, the chances are in the sampling procedure:

a different sample of 1,000 voters will give a slightly different interval.

## Interpretation of a confidence interval

[54.8%, 61.2%] is a 95% confidence interval for the population percentage.

Why ‘confidence’ instead of ‘probability’?

The population percentage  $\mu$  is a *fixed* number, which either falls into [54.8%, 61.2%] or not. There are no chances involved.

Rather, the chances are in the sampling procedure:

a different sample of 1,000 voters will give a slightly different interval.

If one does many polls (the Gallup poll is done frequently), then 95% of these intervals trap the population percentage, and 5% will miss it.

## Interpretation of a confidence interval

[54.8%, 61.2%] is a 95% confidence interval for the population percentage.

Why ‘confidence’ instead of ‘probability’?

The population percentage  $\mu$  is a *fixed* number, which either falls into [54.8%, 61.2%] or not. There are no chances involved.

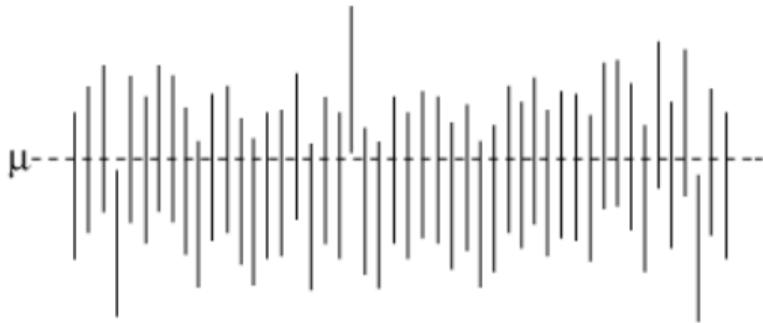
Rather, the chances are in the sampling procedure:

a different sample of 1,000 voters will give a slightly different interval.

If one does many polls (the Gallup poll is done frequently), then 95% of these intervals trap the population percentage, and 5% will miss it.

95% is called the **confidence level**.

## Interpretation of a confidence interval



"I am 95% confident that the President's approval rating is between 54.8% and 61.2%" means that 95% of the time I am correct when making such a statement based on a poll.

## Interpretation of a confidence interval



"I am 95% confident that the President's approval rating is between 54.8% and 61.2%" means that 95% of the time I am correct when making such a statement based on a poll.

Keep in mind that the interval varies from sample to sample, while the population percentage is a fixed number.

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ .

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ . Usually the confidence interval is centered at an estimate for  $\mu$  which is an average.

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ . Usually the confidence interval is centered at an estimate for  $\mu$  which is an average.

Examples:

1.  $\mu$  = approval percentage among all 140 million likely voters.  
estimate = approval percentage among voters in the sample.

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ . Usually the confidence interval is centered at an estimate for  $\mu$  which is an average.

Examples:

1.  $\mu$  = approval percentage among all 140 million likely voters.  
estimate = approval percentage among voters in the sample.
2.  $\mu$  = speed of light.  
estimate = average of 30 measurements.

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ . Usually the confidence interval is centered at an estimate for  $\mu$  which is an average.

Examples:

1.  $\mu$  = approval percentage among all 140 million likely voters.  
estimate = approval percentage among voters in the sample.
2.  $\mu$  = speed of light.  
estimate = average of 30 measurements.

Since the central limit theorem applies for averages, the confidence interval has a simple form:

$$\text{estimate} \pm z \text{SE}$$

where  $z$  is the z-score corresponding to the desired confidence level:

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ . Usually the confidence interval is centered at an estimate for  $\mu$  which is an average.

Examples:

1.  $\mu$  = approval percentage among all 140 million likely voters.  
estimate = approval percentage among voters in the sample.
2.  $\mu$  = speed of light.  
estimate = average of 30 measurements.

Since the central limit theorem applies for averages, the confidence interval has a simple form:

$$\text{estimate} \pm z \text{SE}$$

where  $z$  is the z-score corresponding to the desired confidence level:

95% confidence level  $\rightarrow z = 1.96$

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ . Usually the confidence interval is centered at an estimate for  $\mu$  which is an average.

Examples:

1.  $\mu$  = approval percentage among all 140 million likely voters.  
estimate = approval percentage among voters in the sample.
2.  $\mu$  = speed of light.  
estimate = average of 30 measurements.

Since the central limit theorem applies for averages, the confidence interval has a simple form:

$$\text{estimate} \pm z \text{SE}$$

where  $z$  is the z-score corresponding to the desired confidence level:

95% confidence level  $\rightarrow z = 1.96$

90% confidence level  $\rightarrow z = 1.65$

## Confidence intervals via the Central Limit Theorem

A confidence interval gives a range of plausible values for a *population parameter*  $\mu$ . Usually the confidence interval is centered at an estimate for  $\mu$  which is an average.

Examples:

1.  $\mu$  = approval percentage among all 140 million likely voters.  
estimate = approval percentage among voters in the sample.
2.  $\mu$  = speed of light.  
estimate = average of 30 measurements.

Since the central limit theorem applies for averages, the confidence interval has a simple form:

$$\text{estimate} \pm z \text{SE}$$

where  $z$  is the z-score corresponding to the desired confidence level:

95% confidence level  $\rightarrow z = 1.96$

90% confidence level  $\rightarrow z = 1.65$

99% confidence level  $\rightarrow z = 2.58$

## Estimating the SE with the bootstrap principle

SE is the standard error of the estimate. If the estimate is an average (e.g. a percentage), then we know that

$$SE = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  is the standard deviation of the population.

## Estimating the SE with the bootstrap principle

SE is the standard error of the estimate. If the estimate is an average (e.g. a percentage), then we know that

$$SE = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  is the standard deviation of the population.

Now we have a problem: we don't know  $\sigma$  because we don't know all the data in the population (that's the reason why we sample in the first place!)

## Estimating the SE with the bootstrap principle

SE is the standard error of the estimate. If the estimate is an average (e.g. a percentage), then we know that

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

where  $\sigma$  is the standard deviation of the population.

Now we have a problem: we don't know  $\sigma$  because we don't know all the data in the population (that's the reason why we sample in the first place!)

The **bootstrap principle** states that we can estimate  $\sigma$  by its sample version  $s$  and still get an approximately correct confidence interval.

## Estimating the SE with the bootstrap principle

The **bootstrap principle** states that we can estimate  $\sigma$  by its sample version  $s$  and still get an approximately correct confidence interval.

Examples:

1. We poll 1,000 likely voters and find that 58% approve of the way the president handles his job.

## Estimating the SE with the bootstrap principle

The **bootstrap principle** states that we can estimate  $\sigma$  by its sample version  $s$  and still get an approximately correct confidence interval.

Examples:

1. We poll 1,000 likely voters and find that 58% approve of the way the president handles his job.

$$\text{SE} = \frac{\sigma}{\sqrt{n}} \times 100\%, \text{ where } \sigma = \sqrt{p(1-p)}, p = \text{proportion of all voters who approve.}$$

## Estimating the SE with the bootstrap principle

The **bootstrap principle** states that we can estimate  $\sigma$  by its sample version  $s$  and still get an approximately correct confidence interval.

Examples:

1. We poll 1,000 likely voters and find that 58% approve of the way the president handles his job.

$$\text{SE} = \frac{\sigma}{\sqrt{n}} \times 100\%, \text{ where } \sigma = \sqrt{p(1-p)}, p = \text{proportion of all voters who approve.}$$

The bootstrap principle replaces  $\sigma$  by  $s$  = standard deviation of the 0/1 labels in the sample =  $\sqrt{0.58(1 - 0.58)} = 0.49$ .

## Estimating the SE with the bootstrap principle

The **bootstrap principle** states that we can estimate  $\sigma$  by its sample version  $s$  and still get an approximately correct confidence interval.

Examples:

1. We poll 1,000 likely voters and find that 58% approve of the way the president handles his job.

$$\text{SE} = \frac{\sigma}{\sqrt{n}} \times 100\%, \text{ where } \sigma = \sqrt{p(1-p)}, p = \text{proportion of all voters who approve.}$$

The bootstrap principle replaces  $\sigma$  by  $s$  = standard deviation of the 0/1 labels in the sample =  $\sqrt{0.58(1 - 0.58)} = 0.49$ .

So a 95% confidence interval for  $p$  is

$$58\% \pm 2 \frac{0.49}{\sqrt{1000}}, \quad \text{which is } [54.9\%, 61.1\%]$$

## Estimating the SE with the bootstrap principle

2. estimate = average of 30 measurements of the speed of light.

## Estimating the SE with the bootstrap principle

2. estimate = average of 30 measurements of the speed of light.

What is the population or probability histogram from which we sample?

## Estimating the SE with the bootstrap principle

2. estimate = average of 30 measurements of the speed of light.

What is the population or probability histogram from which we sample?

The reason we get 30 different measurements is because each is off by a chance error:

## Estimating the SE with the bootstrap principle

2. estimate = average of 30 measurements of the speed of light.

What is the population or probability histogram from which we sample?

The reason we get 30 different measurements is because each is off by a chance error:

$$\text{measurement} = \text{speed of light} + \text{measurement error}$$

## Estimating the SE with the bootstrap principle

2. estimate = average of 30 measurements of the speed of light.

What is the population or probability histogram from which we sample?

The reason we get 30 different measurements is because each is off by a chance error:

$$\text{measurement} = \text{speed of light} + \text{measurement error}$$

The measurement error follows a probability histogram that is unknown to us. We estimate the standard deviation  $\sigma$  of this probability histogram by the standard deviation  $s$  of the sample of 30 measurements.

## More about confidence intervals

- ▶ The width of the confidence interval is determined by  $z \text{SE}$ , which is called the **margin of error**.

## More about confidence intervals

- ▶ The width of the confidence interval is determined by  $z \text{SE}$ , which is called the **margin of error**.  
A larger sample size  $n$  will result in a smaller margin of error since  $\text{SE} = \frac{\sigma}{\sqrt{n}}$ , but we need *four times the sample size to cut the width in half*.

## More about confidence intervals

- ▶ The width of the confidence interval is determined by  $z \text{SE}$ , which is called the **margin of error**.  
A larger sample size  $n$  will result in a smaller margin of error since  $\text{SE} = \frac{\sigma}{\sqrt{n}}$ , but we need *four times the sample size to cut the width in half*.  
We can also make the width smaller by making  $z$  smaller, e.g. use a 80% confidence level instead of 95%.

## More about confidence intervals

- ▶ The width of the confidence interval is determined by  $z \text{SE}$ , which is called the **margin of error**.  
A larger sample size  $n$  will result in a smaller margin of error since  $\text{SE} = \frac{\sigma}{\sqrt{n}}$ , but we need *four times the sample size to cut the width in half*.  
We can also make the width smaller by making  $z$  smaller, e.g. use a 80% confidence level instead of 95%. Then the price for more precision (i.e. shorter interval) is less confidence that it covers the parameter  $\mu$ .

## More about confidence intervals

- ▶ The width of the confidence interval is determined by  $z \text{SE}$ , which is called the **margin of error**.  
A larger sample size  $n$  will result in a smaller margin of error since  $\text{SE} = \frac{\sigma}{\sqrt{n}}$ , but we need *four times the sample size to cut the width in half*.  
We can also make the width smaller by making  $z$  smaller, e.g. use a 80% confidence level instead of 95%. Then the price for more precision (i.e. shorter interval) is less confidence that it covers the parameter  $\mu$ .
- ▶ There is an easy to remember formula for a 95% confidence interval for a percentage:

$$\text{estimated percentage} \pm \frac{1}{\sqrt{n}}$$

That's because  $\sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$  no matter what  $p$  is.

## More about confidence intervals

- ▶ The width of the confidence interval is determined by  $z \text{SE}$ , which is called the **margin of error**.  
A larger sample size  $n$  will result in a smaller margin of error since  $\text{SE} = \frac{\sigma}{\sqrt{n}}$ , but we need *four times the sample size to cut the width in half*.  
We can also make the width smaller by making  $z$  smaller, e.g. use a 80% confidence level instead of 95%. Then the price for more precision (i.e. shorter interval) is less confidence that it covers the parameter  $\mu$ .
- ▶ There is an easy to remember formula for a 95% confidence interval for a percentage:  
$$\text{estimated percentage} \pm \frac{1}{\sqrt{n}}$$
That's because  $\sigma = \sqrt{p(1-p)} \leq \frac{1}{2}$  no matter what  $p$  is.
- ▶ It is journalistic convention to use a 95% confidence level unless stated otherwise.

## Mini quiz

1. A random sample of 500 sales prices of recently purchased homes in a county is taken. From that sample a 90% confidence interval for the average sales price of all homes in the county is computed to be \$ 215,000 +/- \$ 35,000. True or false:
  - a. About 90% of all home sales in the county have a sales price in the range \$ 215,000 +/- \$ 35,000.
  - b. There is a 90% chance that the average sales price of all homes in the county is in the range \$ 215,000 +/- \$ 35,000.
2. Based on a sample of 500 salaries in a large city we want to find a confidence interval for the average salary in that city.
  - a. Is it possible to do this using the formula 'average +/- z SE'? (Keep in mind that the histogram of salaries is not normal but quite skewed.)
  - b. The margin of error for this confidence interval turns out to be \$ 5,400. How many salaries do we need to sample in order to shrink the margin of error to about \$ 2,000?

3. You are interested what the current starting salary for jobs in data science is. You solicit feedback on an online forum about data science and you get 230 replies with salary numbers. Can you use the formula 'average +/- z SE' to find a confidence interval for the average starting salary?

## The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

## The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis**,  $H_0$ , states that "nothing extraordinary is going on". So in this case

$$H_0: P(T) = \frac{1}{2}$$

## The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis**,  $H_0$ , states that "nothing extraordinary is going on". So in this case

$$H_0: P(T) = \frac{1}{2}$$

The **alternative hypothesis**,  $H_A$ , states that there is a different chance process that generates the data. Here we can take

$$H_A: P(T) \neq \frac{1}{2}$$

## The logic behind testing hypotheses

We toss a coin 10 times and get 7 tails. Is this sufficient evidence to conclude that the coin is biased?

The **null hypothesis**,  $H_0$ , states that "nothing extraordinary is going on". So in this case

$$H_0: P(T) = \frac{1}{2}$$

The **alternative hypothesis**,  $H_A$ , states that there is a different chance process that generates the data. Here we can take

$$H_A: P(T) \neq \frac{1}{2}$$

Hypothesis testing proceeds by collecting data and evaluating whether the data are compatible with  $H_0$  or not (in which case one **rejects**  $H_0$ ).

## The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

## The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

In this case "nothing extraordinary going on" means that the drug has no effect.

## The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

In this case "nothing extraordinary going on" means that the drug has no effect. So

$H_0$ : no change in blood pressure       $H_A$ : blood pressure drops

## The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

In this case "nothing extraordinary going on" means that the drug has no effect. So

$H_0$ : no change in blood pressure       $H_A$ : blood pressure drops

Note that in this case the company would like to reject  $H_0$ !

## The logic behind testing hypotheses

A different example: A company develops a new drug to lower blood pressure. It tests it with an experiment involving 1,000 patients.

In this case "nothing extraordinary going on" means that the drug has no effect. So

$H_0$ : no change in blood pressure       $H_A$ : blood pressure drops

Note that in this case the company would like to reject  $H_0$ !

So the logic of testing is typically indirect: One assumes that nothing extraordinary is happening and then hopes to reject this assumption  $H_0$ .

## Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if  $H_0$  were true.

## Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if  $H_0$  were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

## Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if  $H_0$  were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

'Observed' is a statistic that is appropriate for assessing  $H_0$ . In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

## Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if  $H_0$  were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

'Observed' is a statistic that is appropriate for assessing  $H_0$ . In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

'Expected' and SE are the expected value and the SE of this statistic, *computed under the assumption that  $H_0$  is true*.

## Setting up a test statistic

A **test statistic** measures how far away the data are from what we would expect if  $H_0$  were true.

The most common test statistic is the **z-statistic**:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

'Observed' is a statistic that is appropriate for assessing  $H_0$ . In the example of the 10 coin tosses, appropriate statistics would be the number of tails or the percent of tails.

'Expected' and SE are the expected value and the SE of this statistic, *computed under the assumption that  $H_0$  is true*.

In the example: Using the formulas for the sum of 0/1 labels we get

'expected' =  $10 \times \frac{1}{2} = 5$  and  $\text{SE} = \sqrt{10} \sqrt{\frac{1}{2} \times \frac{1}{2}} = 1.58$ . So

$$z = \frac{7 - 5}{1.58} = 1.27$$

## p-values measure the evidence against $H_0$

Large values of  $|z|$  are evidence against  $H_0$ : The larger  $|z|$  is, the stronger the evidence.  
The strength of the evidence is measured by the  
**p-value** (or: **observed significance level**):

## p-values measure the evidence against $H_0$

Large values of  $|z|$  are evidence against  $H_0$ : The larger  $|z|$  is, the stronger the evidence.  
The strength of the evidence is measured by the  
**p-value** (or: **observed significance level**):

The p-value is the probability of getting a value of  $z$  as extreme or more extreme than the observed  $z$ , assuming  $H_0$  is true.

## p-values measure the evidence against $H_0$

Large values of  $|z|$  are evidence against  $H_0$ : The larger  $|z|$  is, the stronger the evidence.  
The strength of the evidence is measured by the  
**p-value** (or: **observed significance level**):

The p-value is the probability of getting a value of  $z$  as extreme or more extreme than the observed  $z$ , assuming  $H_0$  is true.

But if  $H_0$  is true, then  $z$  follows that standard normal curve, according to the central limit theorem, so the p-value can be computed with normal approximation:

## p-values measure the evidence against $H_0$

Large values of  $|z|$  are evidence against  $H_0$ : The larger  $|z|$  is, the stronger the evidence.  
The strength of the evidence is measured by the  
**p-value** (or: **observed significance level**):

The p-value is the probability of getting a value of  $z$  as extreme or more extreme than the observed  $z$ , assuming  $H_0$  is true.

But if  $H_0$  is true, then  $z$  follows that standard normal curve, according to the central limit theorem, so the p-value can be computed with normal approximation:

The smaller the p-value, the stronger the evidence against  $H_0$ . Often the criterion for rejecting  $H_0$  is a p-value smaller than 5%. Then the result is called **statistically significant**.

p-values measure the evidence against  $H_0$

In the example:

p-values measure the evidence against  $H_0$

In the example:

Note that the p-value **does not** give the probability that  $H_0$  is true, as  $H_0$  is either true or not - there are no chances involved. Rather, it gives the probability of seeing a statistic as extreme, or more extreme, than the observed one, assuming  $H_0$  is true.

## Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

## Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

## Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

## Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers:    1 = correct answer,  0 = wrong answer

## Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers:    1 = correct answer,  0 = wrong answer

$$H_0: P(0) = P(1) = \frac{1}{2} \quad H_A: P(1) > \frac{1}{2}$$

## Distinguishing Coke and Pepsi by taste

It has been said that it is difficult to distinguish Coke and Pepsi by taste alone, without the visual cue of the bottle or can.

In an experiment that I did in a class at Stanford, 10 cups were filled at random with either Coke or Pepsi. A student volunteer tasted each of the 10 cups and correctly named the contents of seven. Is this sufficient evidence to conclude that the student can tell apart Coke and Pepsi?

"Nothing extraordinary is going on" means that the student does not have any special ability to tell them apart and is just guessing.

To write this down formally we introduce 0/1 labels since we are counting correct answers:    1 = correct answer,  0 = wrong answer

$$H_0: P(0) = P(1) = \frac{1}{2} \quad H_A: P(1) > \frac{1}{2}$$

This is a **one-sided test**: the alternative hypothesis for  $P(1)$  we are interested in is on one side of  $\frac{1}{2}$ .

## Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7 - 5}{1.58} = 1.27$$

## Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7 - 5}{1.58} = 1.27$$

But since we do a one-sided test instead of a two-sided test, the p-value is only half as large:

## Distinguishing Coke and Pepsi by taste

Since we are looking at the sum of ten 0/1 labels, the z-statistic is the same that we had for coin-tossing:

$$z = \frac{\text{observed sum} - \text{expected sum}}{\text{SE of sum}} = \frac{7 - 5}{1.58} = 1.27$$

But since we do a one-sided test instead of a two-sided test, the p-value is only half as large:

Since 10.2% is not smaller than 5%, we don't reject  $H_0$ : We are not convinced that the student can distinguish Coke and Pepsi.

## Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

$$H_A: P(1) \neq \frac{1}{2}$$

$H_A$  corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

## Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

$$H_A: P(1) \neq \frac{1}{2}$$

$H_A$  corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

One has to carefully consider whether the alternative should be one-sided or two-sided, as the p-value gets doubled in the latter case.

## Distinguishing Coke and Pepsi

A two-sided alternative might also be appropriate:

$$H_A: P(1) \neq \frac{1}{2}$$

$H_A$  corresponds to a student who is more likely than not to distinguish Coke and Pepsi, but who may confuse them. Such a student might get one correct answer (say).

One has to carefully consider whether the alternative should be one-sided or two-sided, as the p-value gets doubled in the latter case.

It is not ok to change the alternative afterwards in order to get the p-value below 5%.

## The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

## The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration  $\mu$  in the reservoir is above the standard of 15 ppb?

## The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration  $\mu$  in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

## The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration  $\mu$  in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration  $\mu$  is below 15 ppb, but measurement error results in an average of 15.6 ppb.

## The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration  $\mu$  in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration  $\mu$  is below 15 ppb, but measurement error results in an average of 15.6 ppb.

$$H_0: \mu = 15 \text{ ppb} \quad H_A: \mu > 15 \text{ ppb}$$

## The t-test

The health guideline for lead in drinking water is a concentration of not more than 15 parts per billion (ppb).

Five independent samples from a reservoir average 15.6 ppb. Is this sufficient evidence to conclude that the concentration  $\mu$  in the reservoir is above the standard of 15 ppb?

Recall our model for measurements:

$$\text{measurement} = \mu + \text{measurement error}$$

So it may be that the concentration  $\mu$  is below 15 ppb, but measurement error results in an average of 15.6 ppb.

$$H_0: \mu = 15 \text{ ppb} \quad H_A: \mu > 15 \text{ ppb}$$

We can try a z-test for the average of the measurements:

$$z = \frac{\text{observed average} - \text{expected average}}{\text{SE of average}} = \frac{15.6 \text{ ppb} - 15 \text{ ppb}}{\text{SE of average}}$$

since the measurement error has expected value zero.

## The t-test

SE of average =  $\frac{\sigma}{\sqrt{n}}$ , but the standard deviation  $\sigma$  of the measurement error is unknown.

## The t-test

SE of average =  $\frac{\sigma}{\sqrt{n}}$ , but the standard deviation  $\sigma$  of the measurement error is unknown.

We can estimate  $\sigma$  by  $s$ , the sample standard deviation of the measurements. However:

## The t-test

SE of average =  $\frac{\sigma}{\sqrt{n}}$ , but the standard deviation  $\sigma$  of the measurement error is unknown.

We can estimate  $\sigma$  by  $s$ , the sample standard deviation of the measurements. However:

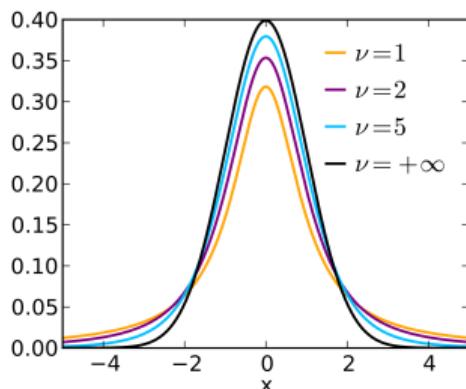
If we estimate  $\sigma$  and  $n$  is small ( $n \leq 20$ ), then the normal curve is not a good enough approximation to the distribution of the z-statistic. Rather, an appropriate approximation is **Student's t-distribution with  $n - 1$  degrees of freedom**:

## The t-test

SE of average =  $\frac{\sigma}{\sqrt{n}}$ , but the standard deviation  $\sigma$  of the measurement error is unknown.

We can estimate  $\sigma$  by  $s$ , the sample standard deviation of the measurements. However:

If we estimate  $\sigma$  and  $n$  is small ( $n \leq 20$ ), then the normal curve is not a good enough approximation to the distribution of the z-statistic. Rather, an appropriate approximation is **Student's t-distribution with  $n - 1$  degrees of freedom**:



## The t-test

The fatter tails account for the additional uncertainty introduced by estimating  $\sigma$  by  
 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ .

## The t-test

The fatter tails account for the additional uncertainty introduced by estimating  $\sigma$  by  
 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ .

Using the **t-test** in place of the z-test is only necessary for small samples:  $n \leq 20$  (say).

## The t-test

The fatter tails account for the additional uncertainty introduced by estimating  $\sigma$  by  
 $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$ .

Using the **t-test** in place of the z-test is only necessary for small samples:  $n \leq 20$  (say).  
In that case it is also better to replace the confidence interval  $\bar{x} \pm z \text{ SE}$  by

$$\bar{x} \pm t_{n-1} \text{SE}$$

## More on testing

- ▶ *Statistically significant does not mean that the effect size is important:*

## More on testing

- ▶ *Statistically significant does not mean that the effect size is important:*  
Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.

## More on testing

- ▶ *Statistically significant does not mean that the effect size is important:*  
Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.  
That may not be of practical concern, even though the test may be highly significant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.

## More on testing

- ▶ *Statistically significant does not mean that the effect size is important:*  
Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.  
That may not be of practical concern, even though the test may be highly significant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.  
Reason: A large sample size  $n$  makes  $SE = \frac{\sigma}{\sqrt{n}}$  small, so even a small exceedance over the limit by (say) 0.05 ppb may give a statistically significant result.

## More on testing

- ▶ *Statistically significant does not mean that the effect size is important:*  
Suppose the sample average shows a lead concentration that is only slightly above the health standard of 15 ppb: say the sample average is 15.05 ppb.  
That may not be of practical concern, even though the test may be highly significant: Statistical significance convinces us that there is an effect, but it doesn't say how big the effect is.  
Reason: A large sample size  $n$  makes  $SE = \frac{\sigma}{\sqrt{n}}$  small, so even a small exceedance over the limit by (say) 0.05 ppb may give a statistically significant result.  
Therefore it is helpful to complement a test with a confidence interval: In the above case a 95% confidence interval for  $\mu$  might be [15.02 ppb, 15.08 ppb].

## More on testing

- ▶ There is a general connection between confidence intervals and tests:

## More on testing

- ▶ There is a general connection between confidence intervals and tests:  
A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.  
(A **5% significance level** means that the threshold for the p-value is 5%).

## More on testing

- ▶ There is a general connection between confidence intervals and tests:  
A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.  
(A **5% significance level** means that the threshold for the p-value is 5%).
- ▶ There are two ways that a test can result in a wrong decision:

## More on testing

- ▶ There is a general connection between confidence intervals and tests:  
A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.  
(A **5% significance level** means that the threshold for the p-value is 5%).
- ▶ There are two ways that a test can result in a wrong decision:  
 $H_0$  is true, but was erroneously rejected → Type I error ('false positive')

## More on testing

- ▶ There is a general connection between confidence intervals and tests:  
A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.  
(A **5% significance level** means that the threshold for the p-value is 5%).
- ▶ There are two ways that a test can result in a wrong decision:  
 $H_0$  is true, but was erroneously rejected → Type I error ('false positive')  
 $H_0$  is false, but we fail to reject it → Type II error

## More on testing

- ▶ There is a general connection between confidence intervals and tests:  
A 95% confidence interval contains all values for the null hypothesis that will not be rejected by a two-sided test at a 5% significance level.  
(A **5% significance level** means that the threshold for the p-value is 5%).
- ▶ There are two ways that a test can result in a wrong decision:  
 $H_0$  is true, but was erroneously rejected → Type I error ('false positive')  
 $H_0$  is false, but we fail to reject it → Type II error  
Rejecting  $H_0$  if the p-value is smaller than 5% means  $P(\text{type I error}) \leq 5\%$

## The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

## The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1$  = proportion of all likely voters approving last month

is equal to

$p_2$  = proportion of all likely voters approving this month

## The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1$  = proportion of all likely voters approving last month

is equal to

$p_2$  = proportion of all likely voters approving this month

"nothing unusual is going on" means  $p_1 = p_2$ .

## The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1$  = proportion of all likely voters approving last month

is equal to

$p_2$  = proportion of all likely voters approving this month

"nothing unusual is going on" means  $p_1 = p_2$ . It's common to look at the difference  $p_2 - p_1$  instead:

$$H_0 : p_2 - p_1 = 0$$

$$H_1 : p_2 - p_1 \neq 0$$

## The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1$  = proportion of all likely voters approving last month

is equal to

$p_2$  = proportion of all likely voters approving this month

"nothing unusual is going on" means  $p_1 = p_2$ . It's common to look at the difference  $p_2 - p_1$  instead:

$$H_0 : p_2 - p_1 = 0$$

$$H_1 : p_2 - p_1 \neq 0$$

$p_1$  is estimated by  $\hat{p}_1 = 55\%$ ,  $p_2$  by  $\hat{p}_2 = 58\%$ .

## The two-sample z-test

Last month, the President's approval rating in a sample of 1,000 likely voters was 55%. This month, a poll of 1,500 likely voters resulted in a rating of 58%. Is this sufficient evidence to conclude that the rating has changed?

We want to assess whether

$p_1$  = proportion of all likely voters approving last month

is equal to

$p_2$  = proportion of all likely voters approving this month

"nothing unusual is going on" means  $p_1 = p_2$ . It's common to look at the difference  $p_2 - p_1$  instead:

$$H_0 : p_2 - p_1 = 0$$

$$H_1 : p_2 - p_1 \neq 0$$

$p_1$  is estimated by  $\hat{p}_1 = 55\%$ ,  $p_2$  by  $\hat{p}_2 = 58\%$ . The central limit theorem applies to the difference  $\hat{p}_2 - \hat{p}_1$  just as it does to  $\hat{p}_1$  and  $\hat{p}_2$ . So we can use a z-test:

## The two-sample z-test

We can use a z-test for the difference  $\hat{p}_2 - \hat{p}_1$ :

## The two-sample z-test

We can use a z-test for the difference  $\hat{p}_2 - \hat{p}_1$ :

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

## The two-sample z-test

We can use a z-test for the difference  $\hat{p}_2 - \hat{p}_1$ :

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if  $\hat{p}_1$  and  $\hat{p}_2$  are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}.$$

## The two-sample z-test

We can use a z-test for the difference  $\hat{p}_2 - \hat{p}_1$ :

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if  $\hat{p}_1$  and  $\hat{p}_2$  are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}. \quad \text{So}$$

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - 0}{\sqrt{\sqrt{\frac{p_1(1-p_1)}{1000}}^2 + \sqrt{\frac{p_2(1-p_2)}{1500}}^2}} =$$

## The two-sample z-test

We can use a z-test for the difference  $\hat{p}_2 - \hat{p}_1$ :

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if  $\hat{p}_1$  and  $\hat{p}_2$  are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}. \quad \text{So}$$

$$z = \frac{(\hat{p}_2 - \hat{p}_1) - 0}{\sqrt{\sqrt{\frac{p_1(1-p_1)}{1000}}^2 + \sqrt{\frac{p_2(1-p_2)}{1500}}^2}} = \frac{0.03}{0.0202} = 1.48$$

## The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

## The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$SE(\bar{x}_2 - \bar{x}_1) = \sqrt{(SE(\bar{x}_1))^2 + (SE(\bar{x}_2))^2}$$

## The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$SE(\bar{x}_2 - \bar{x}_1) = \sqrt{(SE(\bar{x}_1))^2 + (SE(\bar{x}_2))^2}$$

and  $SE(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$  is estimated by  $\frac{s_1}{\sqrt{n_1}}$ .

## The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$\text{SE}(\bar{x}_2 - \bar{x}_1) = \sqrt{(\text{SE}(\bar{x}_1))^2 + (\text{SE}(\bar{x}_2))^2}$$

and  $\text{SE}(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$  is estimated by  $\frac{s_1}{\sqrt{n_1}}$ .

All of the above two-sample tests require that the two samples are independent. They are also applicable in special situations where the samples are dependent, e.g. to compare the treatment effect when subjects are randomized into treatment and control groups.

## The paired-difference test

Do husbands tend to be older than their wives?

## The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

## The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The two-sample t-test is not applicable since the two samples are not independent.

## The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The two-sample t-test is not applicable since the two samples are not independent. Even if they were independent, the small differences in ages would not be significant since the standard deviations are large for husbands and also for the wives.

## The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

## The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

$H_0$ : population difference has mean zero

## The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

$H_0$ : population difference has mean zero

$$t = \frac{\bar{d} - 0}{\text{SE}(\bar{d})}, \quad \text{where } d_i \text{ is the age difference of the } i\text{th couple.}$$

## The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

$H_0$ : population difference has mean zero

$$t = \frac{\bar{d} - 0}{\text{SE}(\bar{d})}, \text{ where } d_i \text{ is the age difference of the } i\text{th couple.}$$

$$\text{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}. \text{ Estimate } \sigma_d \text{ by } s_d = 0.55.$$

## The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

$H_0$ : population difference has mean zero

$t = \frac{\bar{d} - 0}{\text{SE}(\bar{d})}$ , where  $d_i$  is the age difference of the  $i$ th couple.

$\text{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$ . Estimate  $\sigma_d$  by  $s_d = 0.55$ . Then  $t = \frac{1.4 - 0}{0.55/\sqrt{5}} = 5.69$

## The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular t-test, which in this context of **matched pairs** is called **paired t-test**:

$H_0$ : population difference has mean zero

$$t = \frac{\bar{d} - 0}{\text{SE}(\bar{d})}, \text{ where } d_i \text{ is the age difference of the } i\text{th couple.}$$

$$\text{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}. \text{ Estimate } \sigma_d \text{ by } s_d = 0.55. \text{ Then } t = \frac{1.4 - 0}{0.55/\sqrt{5}} = 5.69$$

The independence assumption is in the sampling of the couples.

## The sign test

What if didn't know the age difference  $d_i$  but only if the husband was older or not?

## The sign test

What if didn't know the age difference  $d_i$  but only if the husband was older or not?

We can test

$H_0$ : half the husbands in the population are older than their wives

## The sign test

What if didn't know the age difference  $d_i$  but only if the husband was older or not?

We can test

$H_0$ : half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as we tested whether a coin is fair:

## The sign test

What if didn't know the age difference  $d_i$  but only if the husband was older or not?

We can test

$H_0$ : half the husbands in the population are older than their wives  
using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of } 1s - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5} \frac{1}{2}} = 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on } H_0.$$

## The sign test

What if didn't know the age difference  $d_i$  but only if the husband was older or not?

We can test

$H_0$ : half the husbands in the population are older than their wives  
using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of } 1s - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5} \frac{1}{2}} = 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on } H_0.$$

The p-value of this **sign-test** is less significant than that of the paired t-test. This is because the latter uses more information, namely the size of the differences.

## The sign test

What if didn't know the age difference  $d_i$  but only if the husband was older or not?

We can test

$H_0$ : half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as we tested whether a coin is fair:

$$z = \frac{\text{sum of } 1s - \frac{n}{2}}{\text{SE of sum}} = \frac{5 - \frac{5}{2}}{\sqrt{5} \frac{1}{2}} = 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on } H_0.$$

The p-value of this **sign-test** is less significant than that of the paired t-test. This is because the latter uses more information, namely the size of the differences. On the other hand, the sign test has the virtue of easy interpretation due to the analogy to coin tossing.

## Mini quiz

1. True or false:
  - a. The p-value depends on the data.
  - b. If the p-value is smaller than 5%, then there is less than a 5% chance that the null hypothesis is true.
  - c. If the null hypothesis is true, then there is less than a 5% chance to get a p-value that is smaller than 5%.
  - d. If a data scientist does many tests, then even if all the null hypotheses are true, a certain proportion will be rejected in error.

2. For each of the following situations, indicate which test is appropriate to address the respective question: z-test, t-test, two-sample z-test, sign test, or paired-difference test.
- a. You want to test whether plain M&Ms really contain 24% blue M&Ms as claimed on the manufacturer's web site. You sample 500 plain M&Ms at random and count the fraction of blue M&Ms.
  - b. A high school principal wants to find out whether the average SAT score of this year's graduating class is higher than last year's. She samples 13 students from this year's graduating class at random and wants to compare their average SAT score to the average SAT score from last year's graduating class.
  - c. To investigate whether there are difference in scholastic abilities between first-borns and second-born siblings, 600 families that have at least two children were randomly selected. The scholastic abilities of the first-born and the second-born siblings were assessed with a test and are to be compared.

## Data snooping and the multiple testing fallacy

In 1992 a Swedish study examined whether living near a power line causes adverse health effects. It reported a statistically highly significant increase in childhood leukemia.

## Data snooping and the multiple testing fallacy

In 1992 a Swedish study examined whether living near a power line causes adverse health effects. It reported a statistically highly significant increase in childhood leukemia.

A careful follow-up analysis failed to confirm this result.

Why did the study find a statistically significant result?

## Data snooping and the multiple testing fallacy

In 1992 a Swedish study examined whether living near a power line causes adverse health effects. It reported a statistically highly significant increase in childhood leukemia.

A careful follow-up analysis failed to confirm this result.

Why did the study find a statistically significant result?

The study looked at 800 different health effects:

There were 800 statistical tests involved.

## Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:  
A smaller p-value means stronger evidence.

## Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:  
A smaller p-value means stronger evidence.

p-value < 1% → test is ‘highly significant’

## Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:  
A smaller p-value means stronger evidence.

p-value < 1% → test is ‘highly significant’

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

## Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:  
A smaller p-value means stronger evidence.

p-value < 1% → test is ‘highly significant’

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

So if we do 800 tests, then even if there is no effect at all we expect to see  $800 \times 1\% = 8$  highly significant results just by chance!

## Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:  
A smaller p-value means stronger evidence.

p-value < 1% → test is ‘highly significant’

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

So if we do 800 tests, then even if there is no effect at all we expect to see  $800 \times 1\% = 8$  highly significant results just by chance!

This is called the **multiple testing fallacy** or **look-elsewhere effect**.

## Multiple comparisons

A statistical test summarizes the evidence for an effect by reporting a p-value:  
A smaller p-value means stronger evidence.

p-value < 1% → test is ‘highly significant’

Interpretation: If there is no effect, then there is only a 1% chance to get such a highly significant result.

So if we do 800 tests, then even if there is no effect at all we expect to see  $800 \times 1\% = 8$  highly significant results just by chance!

This is called the **multiple testing fallacy** or **look-elsewhere effect**.

When analyzing large amounts of data it is easy to fall into this trap because there are so many potential relationships to explore, which leads to **data snooping (=data dredging)**.

## Reproducibility and Replicability

Data snooping and other problems have lead to a crisis with regard to **replicability** (getting similar conclusions with different samples, procedures and data analysis methods) and **reproducibility** (getting the same results when using the same data and methods of analysis.)

## Reproducibility and Replicability

Data snooping and other problems have lead to a crisis with regard to **replicability** (getting similar conclusions with different samples, procedures and data analysis methods) and **reproducibility** (getting the same results when using the same data and methods of analysis.)

- ▶ 'How science goes wrong' in The Economist (10/13/2013)
- ▶ 'Why most published research findings are false' by J. Ioannidis (2005)

How can one account for multiple testing?

How can one account for multiple testing?

**Bonferroni correction:** If there are  $m$  tests, multiply the p-values by  $m$ .

## How can one account for multiple testing?

**Bonferroni correction:** If there are  $m$  tests, multiply the p-values by  $m$ .

The Bonferroni correction makes sure that  $P(\text{any of the } m \text{ tests rejects in error}) \leq 5\%$ .

## How can one account for multiple testing?

**Bonferroni correction:** If there are  $m$  tests, multiply the p-values by  $m$ .

The Bonferroni correction makes sure that  $P(\text{any of the } m \text{ tests rejects in error}) \leq 5\%$ .

The Bonferroni correction is often very restrictive: It guards against having even one false positive among the  $m$  tests.

## How can one account for multiple testing?

**Bonferroni correction:** If there are  $m$  tests, multiply the p-values by  $m$ .

The Bonferroni correction makes sure that  $P(\text{any of the } m \text{ tests rejects in error}) \leq 5\%$ .

The Bonferroni correction is often very restrictive: It guards against having even one false positive among the  $m$  tests.

As a consequence the adjusted p-values may not be significant any more even if a noticeable effect is present.

## Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

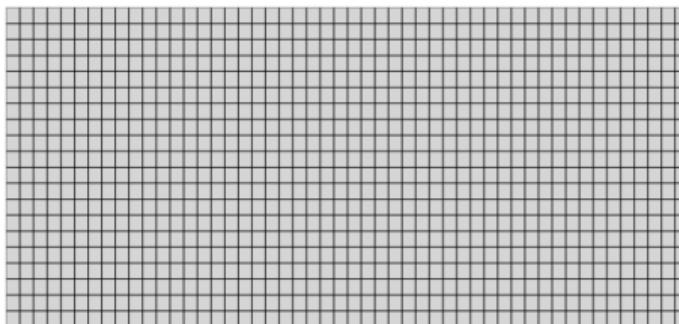
## Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



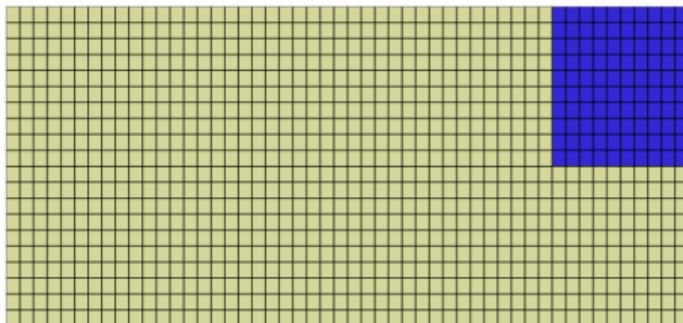
## Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a 'discovery' occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



In 900 cases the null hypothesis is true ("Nothing is going on"), and in 100 cases an alternative hypothesis is true ("There is an effect: something is going on").

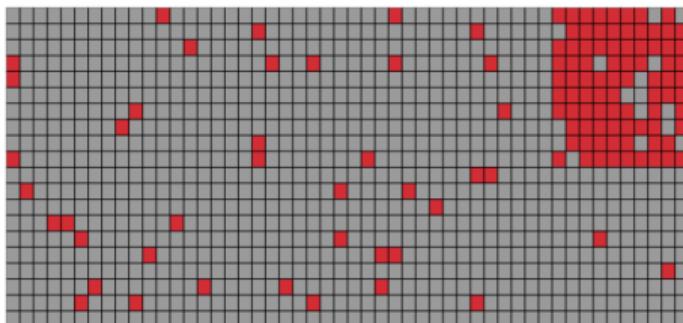
## Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a ‘discovery’ occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



Doing 1,000 tests results in  
**Discoveries** and Non-discoveries.

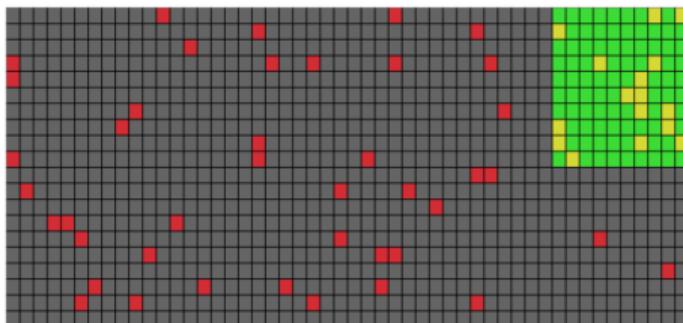
## Accounting for multiple testing

Alternatively, we can try to control the **False Discovery Proportion (FDP)**:

$$\text{FDP} = \frac{\text{number of false discoveries}}{\text{total number of discoveries}}$$

where a ‘discovery’ occurs when a test rejects the null hypothesis.

As an example, we test 1,000 hypotheses.



We made **80 true discoveries** and **41 false discoveries**. The false discovery proportion is  $41/121=0.34$ .

## Accounting for multiple testing with FDR

**False discovery rate (FDR)**: Controls the expected proportion of discoveries that are false.

## Accounting for multiple testing with FDR

**False discovery rate (FDR)**: Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level  $\alpha = 5\%$  (say):

## Accounting for multiple testing with FDR

**False discovery rate (FDR):** Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level  $\alpha = 5\%$  (say):

1. Sort the p-values:  $p_{(1)} \leq \dots \leq p_{(m)}$

## Accounting for multiple testing with FDR

**False discovery rate (FDR):** Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level  $\alpha = 5\%$  (say):

1. Sort the p-values:  $p_{(1)} \leq \dots \leq p_{(m)}$
2. Find the largest  $k$  such that  $p_{(k)} \leq \frac{k}{m}\alpha$

## Accounting for multiple testing with FDR

**False discovery rate (FDR):** Controls the expected proportion of discoveries that are false.

Benjamini-Hochberg procedure to control the FDR at level  $\alpha = 5\%$  (say):

1. Sort the p-values:  $p_{(1)} \leq \dots \leq p_{(m)}$
2. Find the largest  $k$  such that  $p_{(k)} \leq \frac{k}{m}\alpha$
3. Declare discoveries for all tests  $i$  from 1 to  $k$ .

## Accounting for multiple testing with validation set

**Using a validation set:** Split the data into a *model-building set* and a *validation set* before the analysis.

## Accounting for multiple testing with validation set

**Using a validation set:** Split the data into a *model-building set* and a *validation set* before the analysis.

You may use data snooping on the model-building set to find something interesting.

## Accounting for multiple testing with validation set

**Using a validation set:** Split the data into a *model-building set* and a *validation set* before the analysis.

You may use data snooping on the model-building set to find something interesting.

Then test this hypothesis on the validation set.

## Accounting for multiple testing with validation set

**Using a validation set:** Split the data into a *model-building set* and a *validation set* before the analysis.

You may use data snooping on the model-building set to find something interesting.

Then test this hypothesis on the validation set.

This approach requires strict discipline: You are not allowed to look at the validation set during the exploratory step!

## Mini quiz

1. A medical study examines whether there is a significant correlation between any of 12 lifestyle choices and high blood pressure. It doesn't find any significant correlation, but upon further examination the researchers find a highly significant ( $p\text{-value} < 0.5\%$ ) correlation between two of the lifestyle choices. This correlation seems not to have been noticed before. Which of the following three statements is an appropriate summary of these findings:
- i) The correlation between these two lifestyle choices is highly significant and should be reported as such.
  - ii) The seemingly significant correlation was found as a consequence of data snooping and therefore the  $p\text{-value}$  is not valid. The researchers shouldn't report anything.
  - iii) The seemingly significant correlation was found as a consequence of data snooping and therefore the  $p\text{-value}$  is not valid. However, this could potentially be a significant new finding. The researchers can report it as such, pointing out that they cannot attach a valid  $p\text{-value}$  to this finding. It can serve as a hypothesis for a future study with new data, which would then allow for statistically valid conclusions.

2. 1,000 tests were evaluated with the Bonferroni correction. 31 tests had corrected p-values smaller than 5%. Which of the following three statements are an appropriate conclusion:

- i) There is a 95% probability that all of these 31 null hypotheses are false.
- ii) This is sufficient evidence to reject all of these 31 null hypotheses, because there is only a 5% chance that any of these 31 p-values would be this small if the null hypothesis were true.
- iii) If we reject these 31 null hypotheses then we can expect that about 5% of them are rejected in error.

3. 1,000 tests were evaluated with the FDR at the 5% level, which resulted in 31 discoveries. Which of the above statements (i)-(iii) are an appropriate conclusion?