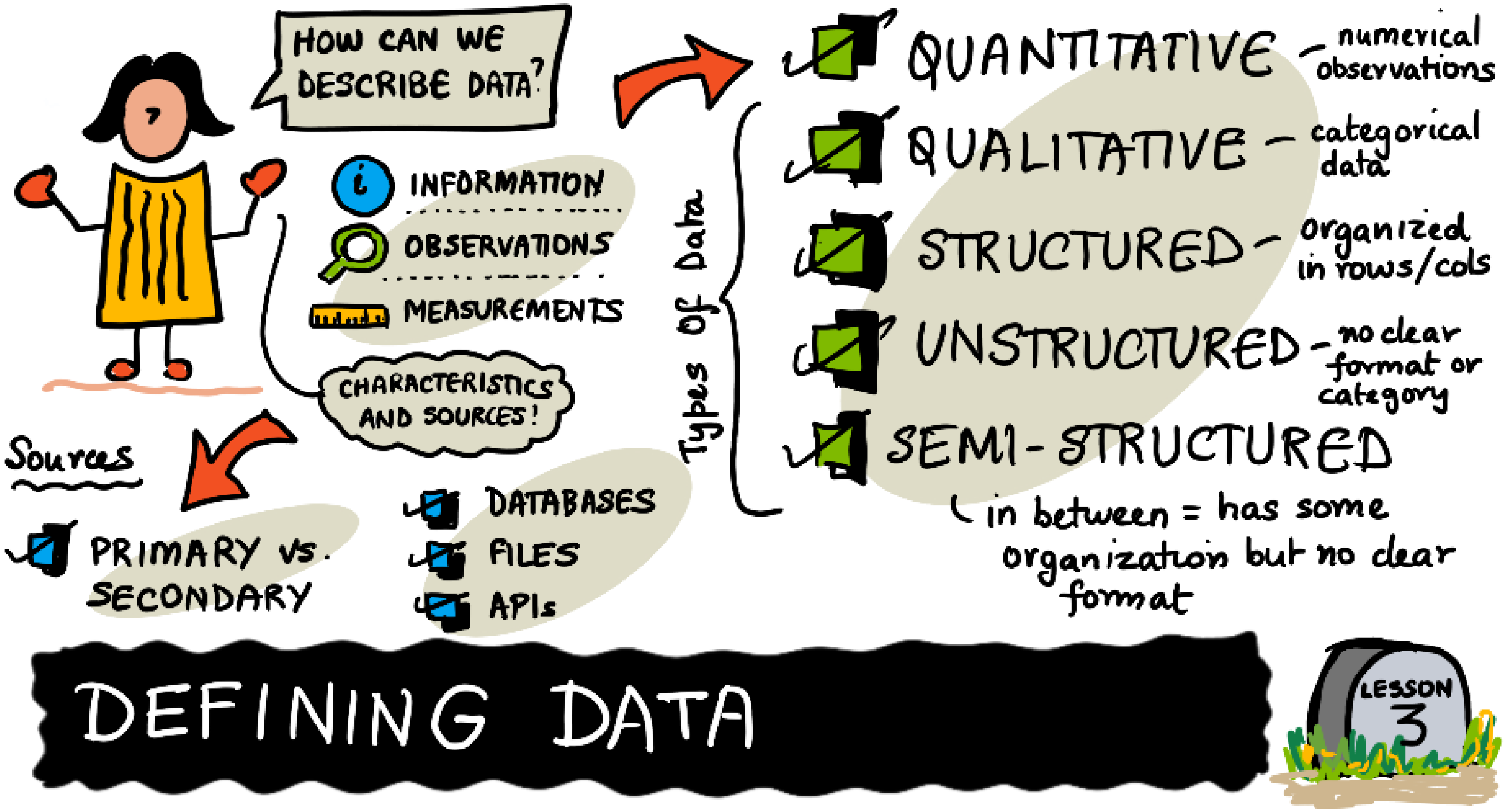
Lecture 2

Defining Data

Statistics and Probability

Defining Data


HOW CAN WE DESCRIBE DATA?
INFORMATION
OBSERVATIONS
MEASUREMENTS
CHARACTERISTICS AND SOURCES!
Sources
PRIMARY vs. SECONDARY
DATABASES
FILES
APIs
Types of Data
QUANTITATIVE - numerical observations
QUALITATIVE - categorical data
STRUCTURED - organized in rows/cols
UNSTRUCTURED - no clear format or category
SEMI-STRUCTURED
in between = has some organization but no clear format
DEFINING DATA
LESSON 3

|Defining Data - *Sketchnote by* @nitya |

Defining Data

- Data is facts, information, observations and measurements that are used to make discoveries and to support informed decisions.
- A data point is a single unit of data with in a dataset, which is collection of data points. Datasets may come in different formats and structures, and will usually be based on its source, or where the data came from.
- For example, a company's monthly earnings might be in a spreadsheet but hourly heart rate data from a smartwatch may be in JSON format.
- It's common for data scientists to work with different types of data within a dataset.

This lecture focuses on identifying and classifying data by its characteristics and its sources.

How Data is Described

- **Raw data** is data that has come from its source in its initial state and has not been analyzed or organized.
- In order to make sense of what is happening with a dataset, it needs to be organized into a format that can be understood by humans as well as the technology they may use to analyze it further.
- The structure of a dataset describes how it's organized and can be classified at structured, unstructured and semi-structured.
- These types of structure will vary, depending on the source but will ultimately fit in these three categories.

Quantitative Data

- `Quantitative data` is numerical observations within a dataset and can typically be analyzed, measured and used mathematically.
- Some examples of quantitative data are: a country's population, a person's height or a company's quarterly earnings.
- With some additional analysis, quantitative data could be used to discover seasonal trends of the Air Quality Index (AQI) or estimate the probability of rush hour traffic on a typical work day.

Qualitative Data

- `Qualitative data`, also known as categorical data is data that cannot be measured objectively like observations of quantitative data.
- It's generally various formats of subjective data that captures the quality of something, such as a product or process.
- Sometimes, qualitative data is numerical and wouldn't be typically used mathematically, like phone numbers or timestamps.
- Some examples of qualitative data are: video comments, the make and model of a car or your closest friends' favorite color.
- Qualitative data could be used to understand which products consumers like best or identifying popular keywords in job application resumes.

Two types of variables

- **categorical**: records a category (e.g., gender, color, T/F, educational level, Likert scales)
- **quantitative variables**: records a numerical measurement

Categorical variables might or might not have an order associated with the categories.

In this lecture we'll focus on **quantitative** variables, which can be either **discrete** or **continuous**:

- **discrete variables**: values are discrete (e.g., year born, counts)
- **continuous variables**: values are real numbers (e.g., length, temperature, time)

(Note categorical variables are always discrete.)

Structured Data

- Structured data is data that is organized into rows and columns, where each row will have the same set of columns.
- Columns represent a value of a particular type and will be identified with a name describing what the value represents, while rows contain the actual values.
- Columns will often have a specific set of rules or restrictions on the values, to ensure that the values accurately represent the column.
- For example imagine a spreadsheet of customers where each row must have a phone number and the phone numbers never contain alphabetical characters.
- There may be rules applied on the phone number column to make sure it's never empty and only contains numbers.

Structured Data

- A benefit of structured data is that it can be organized in such a way that it can be related to other structured data.
- However, because the data is designed to be organized in a specific way, making changes to its overall structure can take a lot of effort to do.
- For example, adding an email column to the customer spreadsheet that cannot be empty means you'll need figure out how you'll add these values to the existing rows of customers in the dataset.
- Examples of structured data: spreadsheets, relational databases, phone numbers, bank statements

Unstructured Data

- Unstructured data typically cannot be categorized into rows or columns and doesn't contain a format or set of rules to follow.
- Because unstructured data has less restrictions on its structure it's easier to add new information in comparison to a structured dataset.
- If a sensor capturing data on barometric pressure every 2 minutes has received an update that now allows it to measure and record temperature, it doesn't require altering the existing data if it's unstructured.
- However, this may make analyzing or investigating this type of data take longer.
- For example, a scientist who wants to find the average temperature of the previous month from the sensors data, but discovers that the sensor recorded an "e" in some of its recorded data to note that it was broken instead of a typical number, which means the data is incomplete.
- Examples of unstructured data: text files, text messages, video files

Semi-structured

- Semi-structured data has features that make it a combination of structured and unstructured data.
- It doesn't typically conform to a format of rows and columns but is organized in a way that is considered structured and may follow a fixed format or set of rules.
- The structure will vary between sources, such as a well defined hierarchy to something more flexible that allows for easy integration of new information.
- Metadata are indicators that help decide how the data is organized and stored and will have various names, based on the type of data.
- Some common names for metadata are tags, elements, entities and attributes.
- For example, a typical email message will have a subject, body and a set of recipients and can be organized by whom or when it was sent.
- Examples of semi-structured data: HTML, CSV files, JSON

Sources of Data

- A data source is the initial location of where the data was generated, or where it "lives" and will vary based on how and when it was collected.
- Data generated by its user(s) are known as primary data while secondary data comes from a source that has collected data for general use.
- For example, a group of scientists collecting observations in a rainforest would be considered primary and if they decide to share it with other scientists it would be considered secondary to those that use it.

Sources of Data

- Databases are a common source and rely on a database management system to host and maintain the data where users use commands called queries to explore the data.
- Files as data sources can be audio, image, and video files as well as spreadsheets like Excel.
- Internet sources are a common location for hosting data, where databases as well as files can be found.
- Application programming interfaces, also known as APIs allow programmers to create ways to share data with external users through the internet, while the process of web scraping extracts data from a web page.

Conclusion

In this lecture we have learned:

- What data is
- How data is described
- How data is classified and categorized
- Where data can be found

🚀 Challenge

Kaggle is an excellent source of open datasets. Use the dataset search tool to find some interesting datasets and classify 3-5 datasets with this criteria:

- Is the data quantitative or qualitative?
- Is the data structured, unstructured, or semi-structured?

Quiz

For each of the following variables, is the variable type categorical, quantitative discrete, or quantitative continuous?

1. Latitude
2. Olympic 50 meter race times
3. Olympic floor gymnastics score
4. College major
5. Number of offspring of a rat

More Quizzes: Classifying Datasets

Instructions

Follow the prompts to identify and classify the data with one of each of the following data types:

Structure Types: Structured, Semi-Structured, or Unstructured

Value Types: Qualitative or Quantitative

Source Types: Primary or Secondary

1. A company has been acquired and now has a parent company. The data scientists have received a spreadsheet of customer phone numbers from the parent company.

- Structure Type:
- Value Type:
- Source Type:

2. A smart watch has been collecting heart rate data from its wearer, and the raw data is in JSON format.

- Structure Type:
- Value Type:
- Source Type:

3. A workplace survey of employee morale that is stored in a CSV file.

- Structure Type:
- Value Type:
- Source Type:

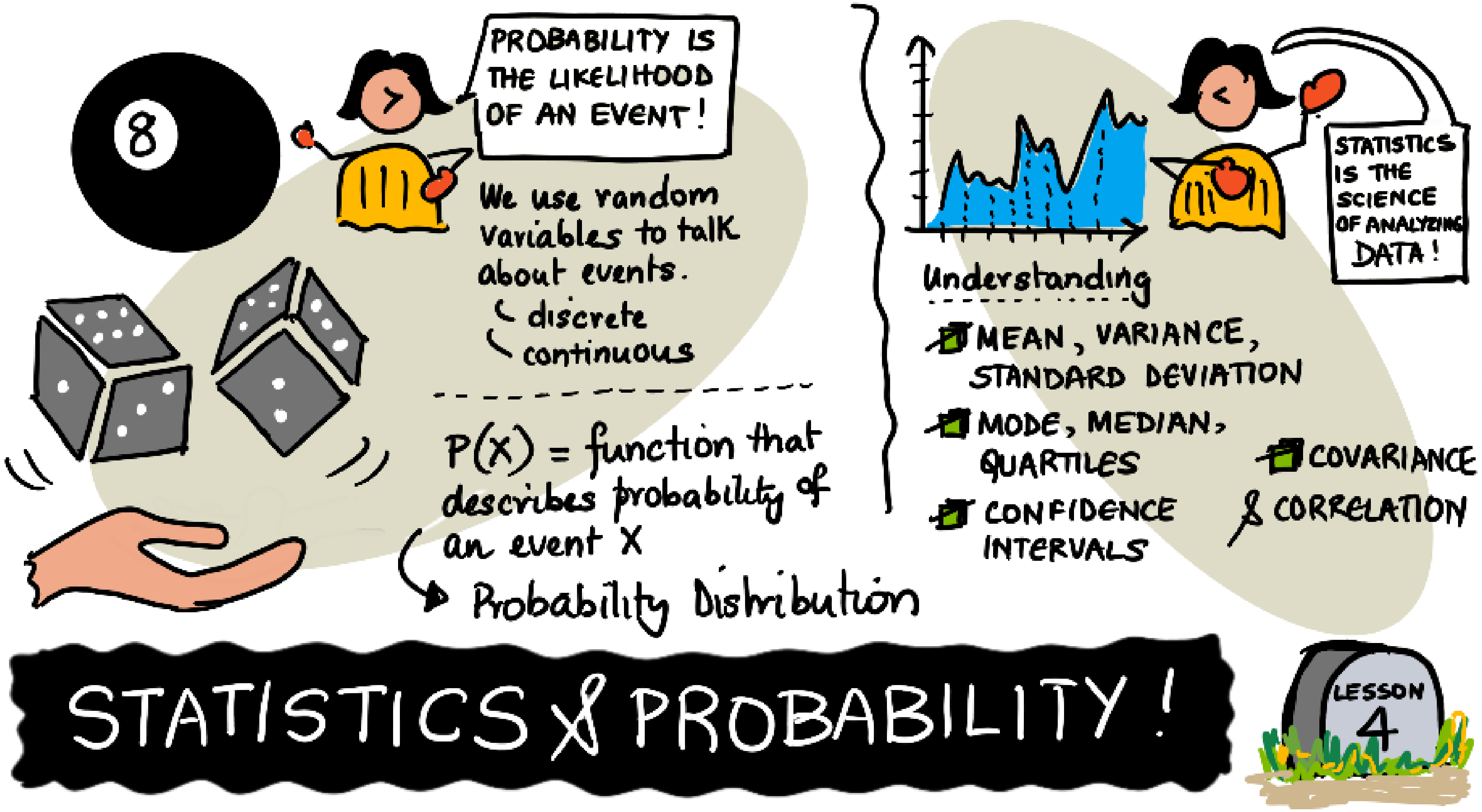
4. Astrophysicists are accessing a database of galaxies that has been collected by a space probe. The data contains the number of planets within in each galaxy.

- Structure Type:
- Value Type:
- Source Type:

5. A personal finance app uses APIs to connect to a user's financial accounts in order to calculate their net worth. They can see all of their transactions in a format of rows and columns and looks similar to a spreadsheet.

- Structure Type:
- Value Type:
- Source Type:

Statistics and Probability


8
PROBABILITY IS THE LIKELIHOOD OF AN EVENT!
We use random Variables to talk about events.
discrete
continuous
P(X) = function that describes probability of an event X
Probability Distribution
STATISTICS IS THE SCIENCE OF ANALYZING DATA!
Understanding
MEAN, VARIANCE, STANDARD DEVIATION
MODE, MEDIAN, QUARTILES
CONFIDENCE INTERVALS
COVARIANCE & CORRELATION
STATISTICS & PROBABILITY!
LESSON 4

| Statistics and Probability - *Sketchnote by @nitya* |

Statistics and Probability

- Statistics and Probability Theory are two highly related areas of Mathematics that are highly relevant to Data Science.
- It is possible to operate with data without deep knowledge of mathematics, but it is still better to know at least some basic concepts.
- This is a short introduction that will help you get started.

Probability and Random Variables

- **Probability** is a number between 0 and 1 that expresses how probable an **event** is.
- It is defined as a number of positive outcomes (that lead to the event), divided by total number of outcomes, given that all outcomes are equally probable.
- For example, when we roll a dice, the probability that we get an even number is $3/6 = 0.5$.

Probability and Random Variables

- When we talk about events, we use **random variables**.
- For example, the random variable that represents a number obtained when rolling a dice would take values from 1 to 6.
- Set of numbers from 1 to 6 is called **sample space**.
- We can talk about the probability of a random variable taking a certain value, for example $P(X = 3) = 1/6$.

Probability and Random Variables

- The random variable in previous example is called **discrete**, because it has a countable sample space, i.e. there are separate values that can be enumerated.
- There are cases when sample space is a range of real numbers, or the whole set of real numbers. Such variables are called **continuous**.
- A good example is the time when the bus arrives.

Probability Distribution

- In the case of discrete random variables, it is easy to describe the probability of each event by a function $P(X)$.
- For each value s from sample space S it will give a number from 0 to 1, such that the sum of all values of $P(X = s)$ for all events would be 1.
- The most well-known discrete distribution is **uniform distribution**, in which there is a sample space of N elements, with equal probability of $1/N$ for each of them.

Probability Distribution

- It is more difficult to describe the probability distribution of a continuous variable, with values drawn from some interval [a,b], or the whole set of real numbers $\mathbb{R}$;.
- Consider the case of bus arrival time. In fact, for each exact arrival time t, the probability of a bus arriving at exactly that time is 0!

> Now you know that events with 0 probability happen, and very often! At least each time when the bus arrives!

Probability Distribution

- We can only talk about the probability of a variable falling in a given interval of values, eg. $P(t_1 \leq X < t_2)$. In this case, probability distribution is described by a **probability density function** p(x), such that

$$P(t * 1 \leq X < t_2) = \int *{t_1}^{t_2} p(x) dx$$

Probability Distribution

- A continuous analog of uniform distribution is called **continuous uniform**, which is defined on a finite interval.
- A probability that the value X falls into an interval of length l is proportional to l, and rises up to 1.
- Another important distribution is **normal distribution**, which we will talk about in more detail below.

Descriptive statistics (quantitative variables)

The goal is to describe a dataset with a small number of statistics or figures

Suppose we are given a sample, $x_1, x_2, \ldots, x_n$, of numerical values

Some *descriptive statistics* for quantitative data are the min, max, median, and mean, $\frac{1}{n}\sum_{i=1}^{n} x_i$

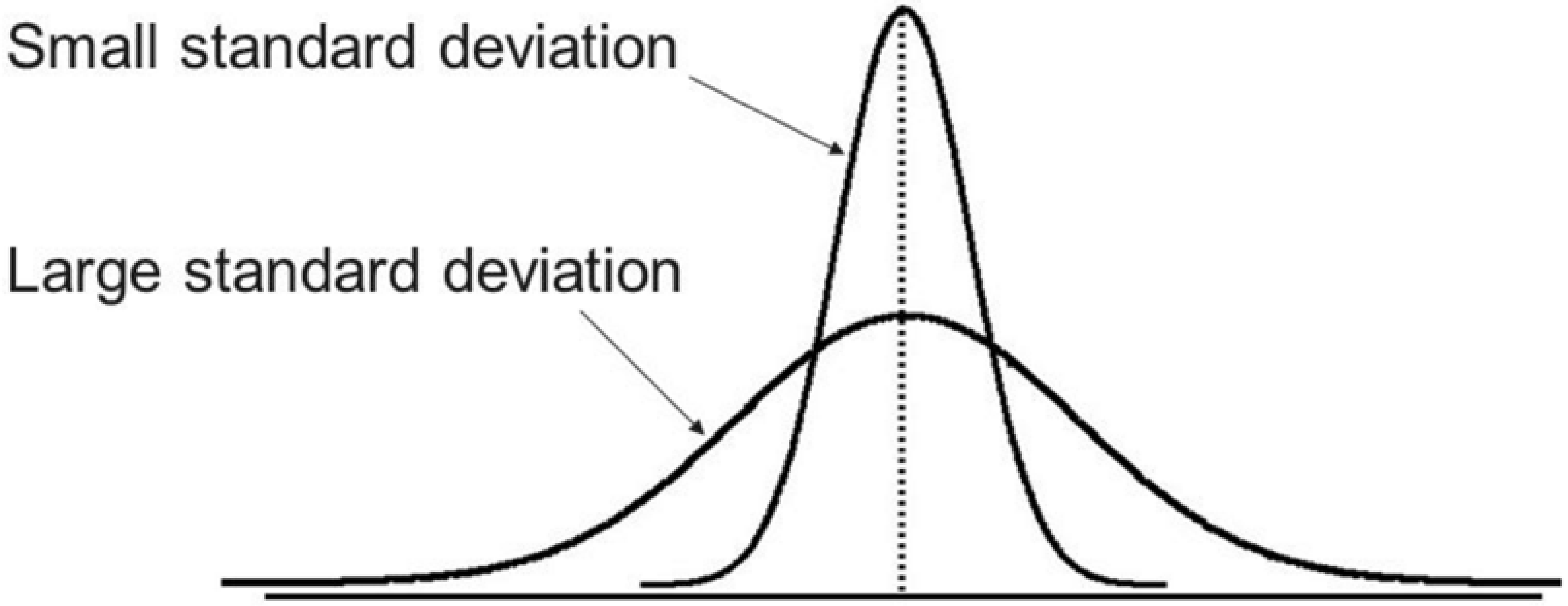
Mean, Variance and Standard Deviation

- Suppose we draw a sequence of n samples of a random variable X: $x_1, x_2, \ldots, x_n$.
- We can define **mean** (or **arithmetic average**) value of the sequence in the traditional way as $x_1 + x_2 + \ldots + x_n/n$.
- As we grow the size of the sample (i.e. take the limit with $n \to \infty$), we will obtain the mean (also called **expectation**) of the distribution.
- We will denote expectation by $E(x)$.

  > It can be demonstrated that for any discrete distribution with values {$x_1, x_2, \ldots, x_N$} and corresponding probabilities $p_1, p_2, \ldots, p_N$, the expectation would equal to $E(X) = x_1p_1 + x_2p_2 + \ldots + x_Np_N$.

- To identify how far the values are spread, we can compute the variance $\sigma^2 = \sum(x_i - \mu)^2/n$, where μ is the mean of the sequence. The value σ is called **standard deviation**, and σ^2 is called a **variance**.

Standard Deviation and Histogram



Mode, Median and Quartiles

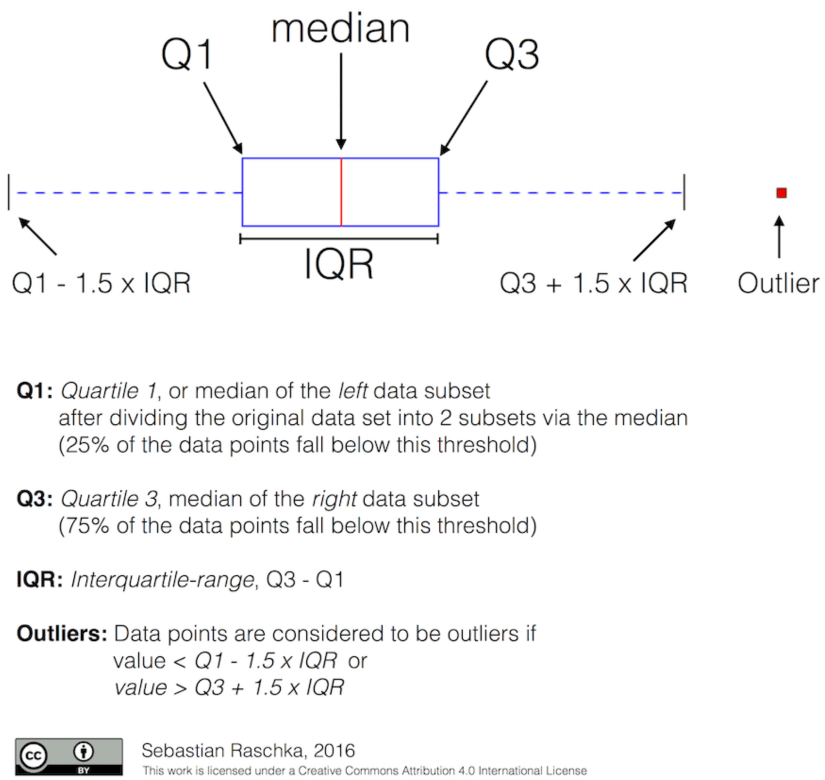
- Sometimes, mean does not adequately represent the "typical" value for data.
- For example, when there are a few extreme values that are completely out of range, they can affect the mean.
- Another good indication is a **median**, a value such that half of data points are lower than it, and another half - higher.

To help us understand the distribution of data, it is helpful to talk about **quartiles**:

- First quartile, or Q1, is a value, such that 25% of the data fall below it
- Third quartile, or Q3, is a value that 75% of the data fall below it

Mode, Median and Quartiles

Graphically we can represent relationship between median and quartiles in a diagram called the **box plot**:



Q1: *Quartile 1*, or median of the *left* data subset
after dividing the original data set into 2 subsets via the median
(25% of the data points fall below this threshold)

Q3: *Quartile 3*, median of the *right* data subset
(75% of the data points fall below this threshold)

IQR: *Interquartile-range*, Q3 - Q1

Outliers: Data points are considered to be outliers if
value < *Q1 - 1.5 x IQR* or
value > Q3 + 1.5 x IQR

Sebastian Raschka, 2016

Here we also compute **inter-quartile range** $IQR = Q3 - Q1$, and so-called **outliers** - values, that lie outside the boundaries $[Q1 - 1.5 * IQR, Q3 + 1.5 * IQR]$.

Mode, Median and Quartiles

- For finite distribution that contains a small number of possible values, a good "typical" value is the one that appears the most frequently, which is called **mode**.
- It is often applied to categorical data, such as colors.
- Consider a situation when we have two groups of people - some that strongly prefer red, and others who prefer blue.
- If we code colors by numbers, the mean value for a favorite color would be somewhere in the orange-green spectrum, which does not indicate the actual preference on neither group.
- However, the mode would be either one of the colors, or both colors, if the number of people voting for them is equal (in this case we call the sample **multimodal**).

Covariance and Correlation

- One of the things Data Science does is finding relations between data.
- We say that two sequences **correlate** when they exhibit the similar behavior at the same time, i.e. they either rise/fall simultaneously, or one sequence rises when another one falls and vice versa.
- In other words, there seems to be some relation between two sequences.

> Correlation does not necessarily indicate causal relationship between two sequences; sometimes both variables can depend on some external cause, or it can be purely by chance the two sequences correlate. However, strong mathematical correlation is a good indication that two variables are somehow connected.

Covariance and Correlation

- Mathematically, the main concept that shows the relation between two random variables is **covariance**, that is computed like this: $\mathrm{Cov}(X,Y) = \mathbf{E}[(X-\mathbf{E}(X))(Y-\mathbf{E}(Y))]$.
- We compute the deviation of both variables from their mean values, and then product of those deviations.
- If both variables deviate together, the product would always be a positive value, that would add up to positive covariance.
- If both variables deviate out-of-sync (i.e. one falls below average when another one rises above average), we will always get negative numbers, that will add up to negative covariance.
- If the deviations are not dependent, they will add up to roughly zero.

Covariance and Correlation

- The absolute value of covariance does not tell us much on how large the correlation is, because it depends on the magnitude of actual values.
- To normalize it, we can divide covariance by standard deviation of both variables, to get **correlation**.
- The good thing is that correlation is always in the range of [-1,1], where 1 indicates strong positive correlation between values, -1 - strong negative correlation, and 0 - no correlation at all (variables are independent).

Covariance and Correlation

Example: We can compute correlation between weights and heights of baseball players from the dataset mentioned above:

```
print(np.corrcoef(weights,heights))
```

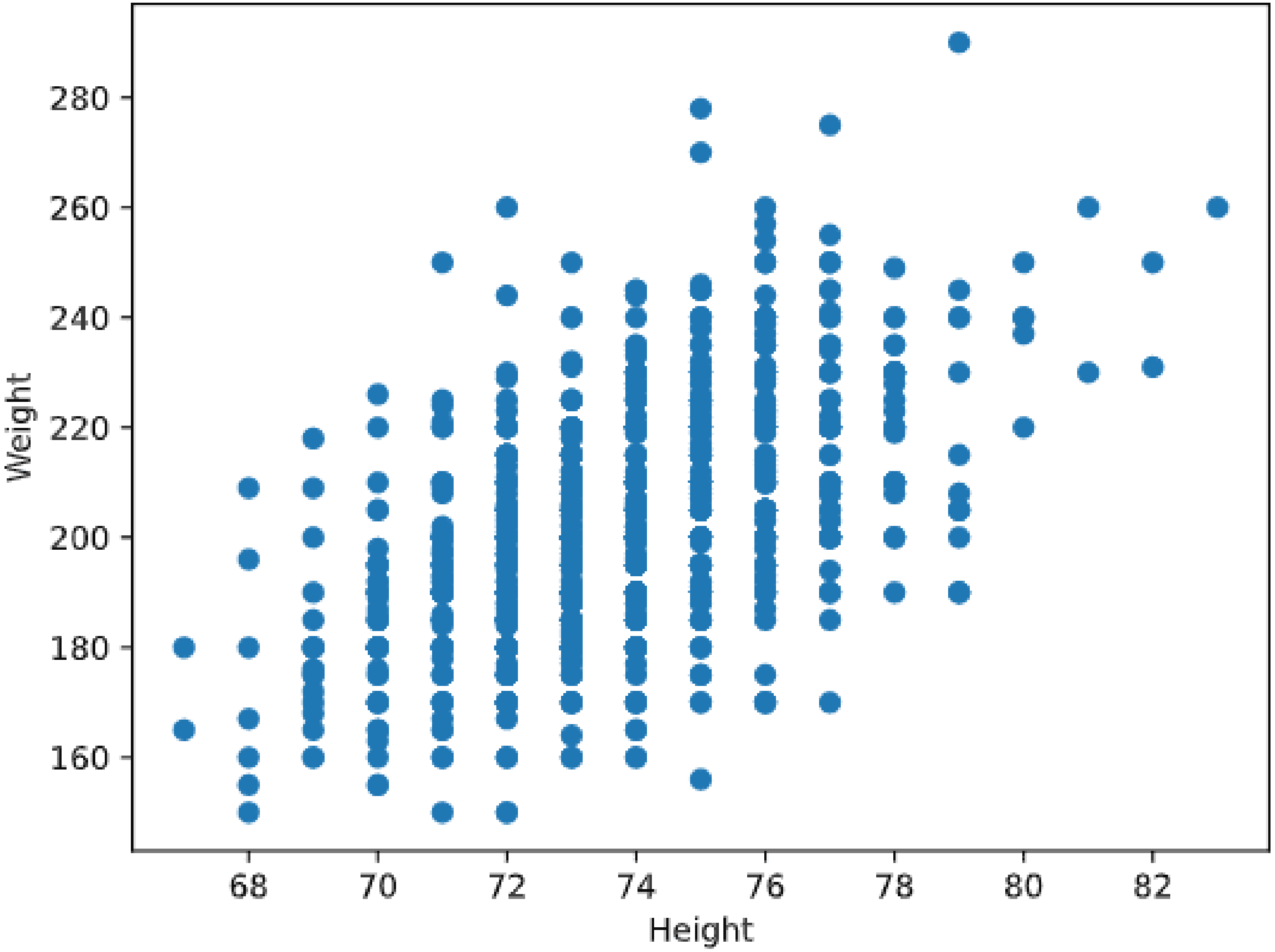
As a result, we get **correlation matrix** like this one:

```
array([[1.        , 0.52959196],
       [0.52959196, 1.        ]])
```

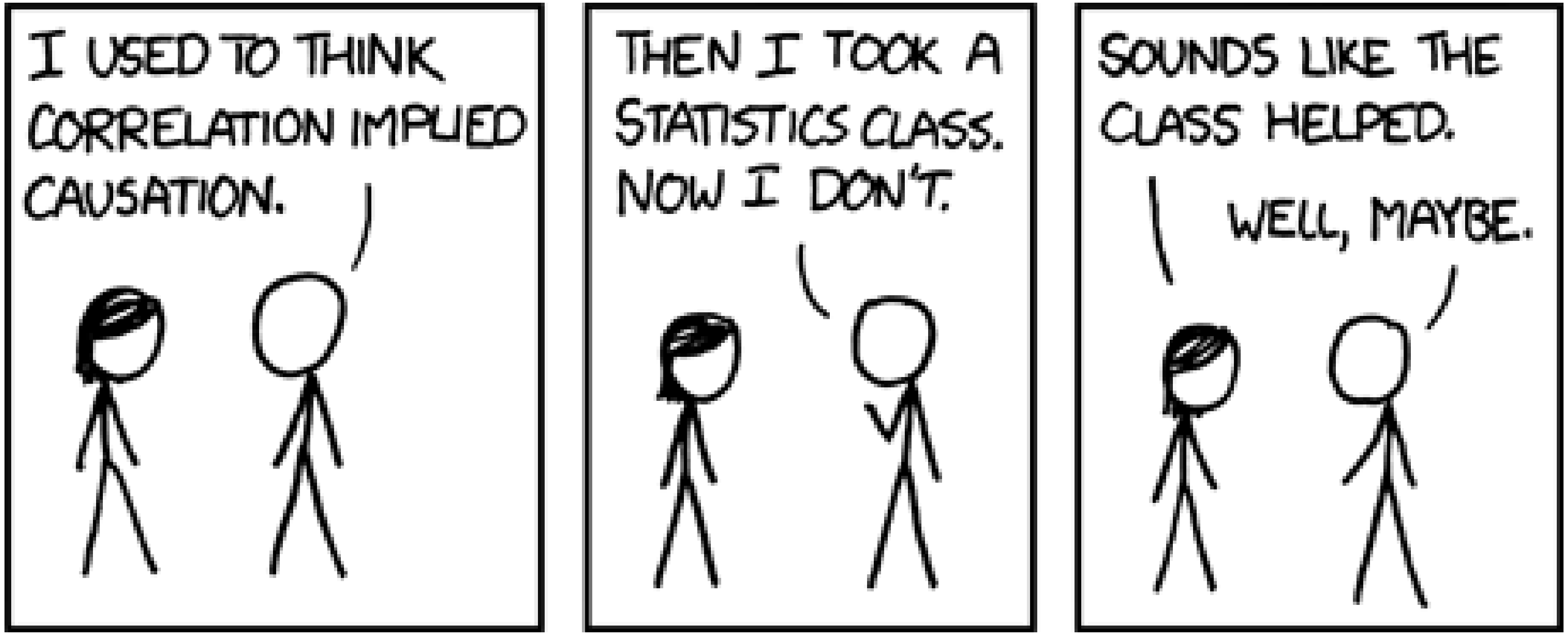
Covariance and Correlation

> Correlation matrix C can be computed for any number of input sequences $S_1, \ldots, S_n$. The value of C_{ij} is the correlation between S_i and S_j, and diagonal elements are always 1 (which is also auto-correlation of S_i).

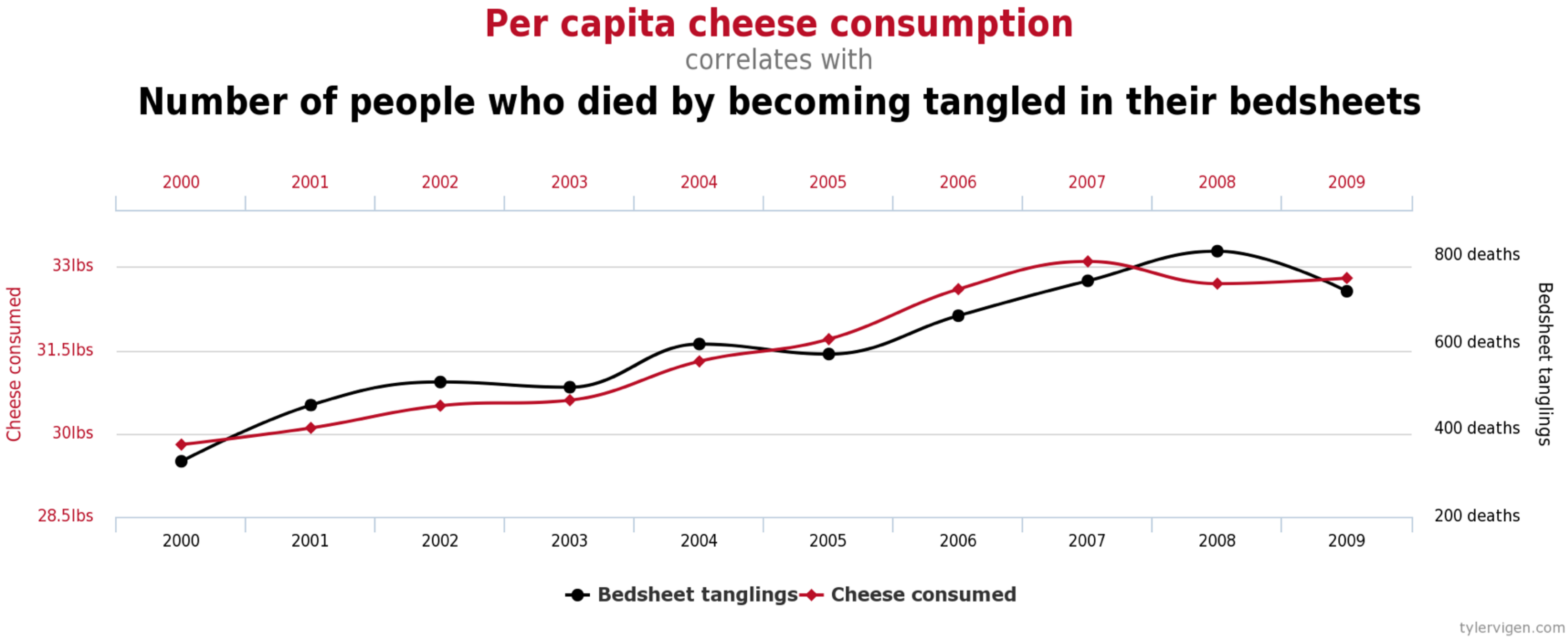
In our case, the value 0.53 indicates that there is some correlation between weight and height of a person. We can also make the scatter plot of one value against the other to see the relationship visually:


Weight
280
260
240
220
200
180
160
Height
68
70
72
74
76
78
80
82

Correlation vs Causation


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Spurious Correlations I (www.tylervigen.com)



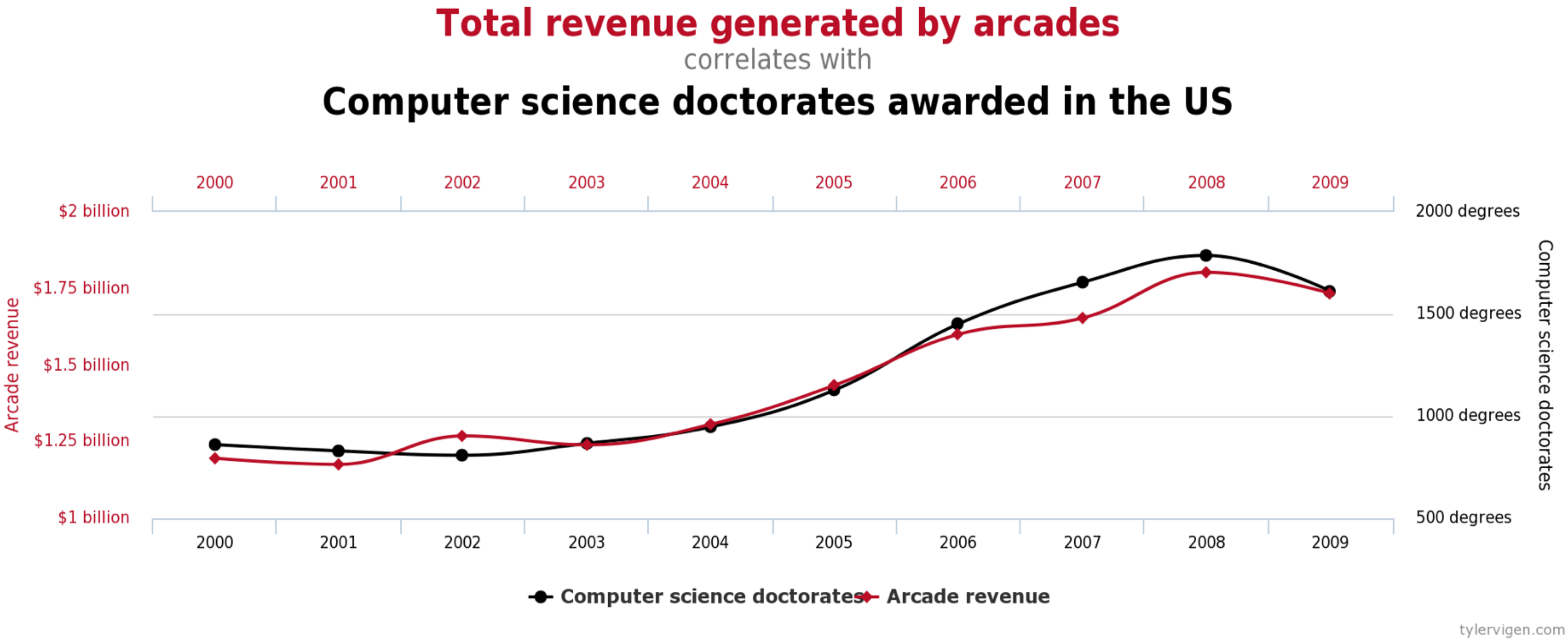
Per capita cheese consumption

correlates with

Number of people who died by becoming tangled in their bedsheets

tylervigen.com

Spurious Correlations II (www.tylervigen.com)



Confounding: example

- Suppose we are given city statistics covering a four-month summer period.
- We observe that swimming pool deaths tend to increase on days when more ice cream is sold.
- Should we conclude that ice cream is the killer?
    - No!
- As astute analysts, we identify average daily temperature as a confounding variable: on hotter days, people are more likely to both buy ice cream and visit swimming pools.
- Regression methods can be used to statistically control for this confounding variable, eliminating the direct relationship between ice cream sales and swimming pool deaths.

> **source**: Jacob Westfall and Tal Yarkoni, Statistically Controlling for Confounding Constructs Is Harder than You Think, PLOS One (2016). link

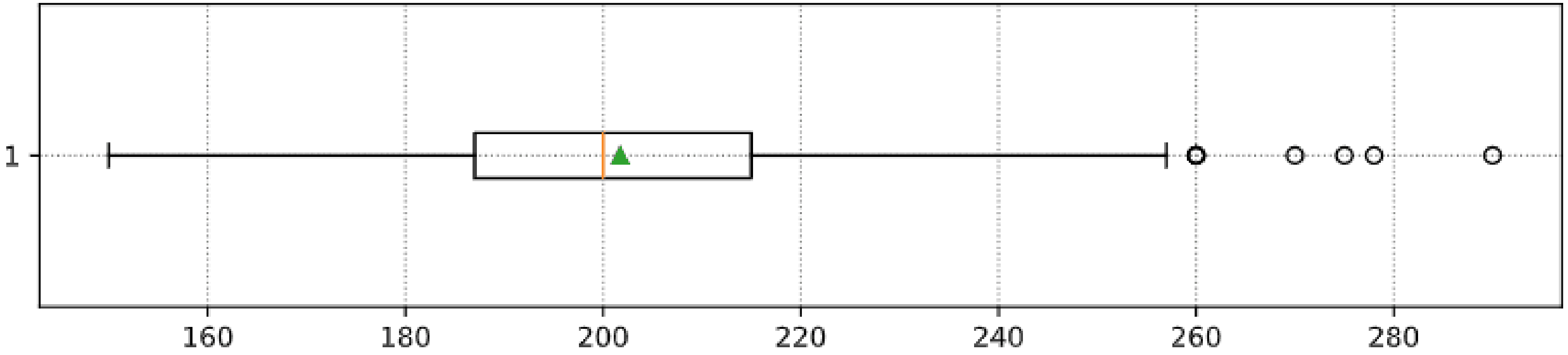
Real-world Data

- When we analyze data from real life, they often are not random variables as such, in a sense that we do not perform experiments with unknown result.
- For example, consider a team of baseball players, and their body data, such as height, weight and age.
- Those numbers are not exactly random, but we can still apply the same mathematical concepts.
- For example, a sequence of people's weights can be considered to be a sequence of values drawn from some random variable.
- Below is the sequence of weights of actual baseball players from Major League Baseball, taken from this dataset (for your convenience, only first 20 values are shown):

```
[180.0, 215.0, 210.0, 210.0, 188.0, 176.0, 209.0, 200.0, 231.0, 180.0, 188.0, 180.0, 185.0, 160.0, 180.0, 185.0, 197.0, 189.0, 185.0, 219.0]
```

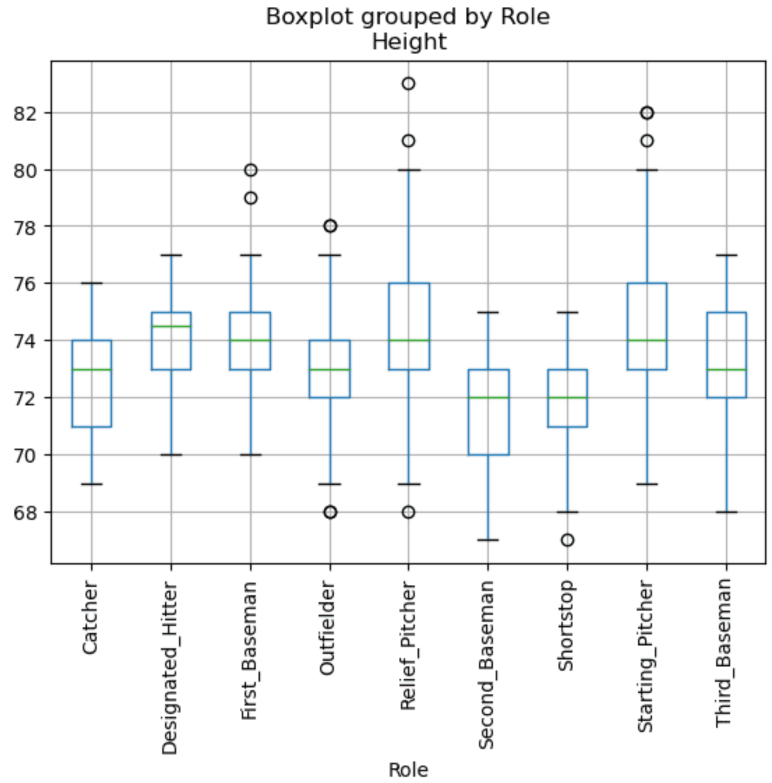
Real-world Data

Here is the box plot showing mean, median and quartiles for our data:



Since our data contains information about different player **roles**, we can also do the box plot by role - it will allow us to get the idea on how parameters values differ across roles. This time we will consider height:

Real-world Data



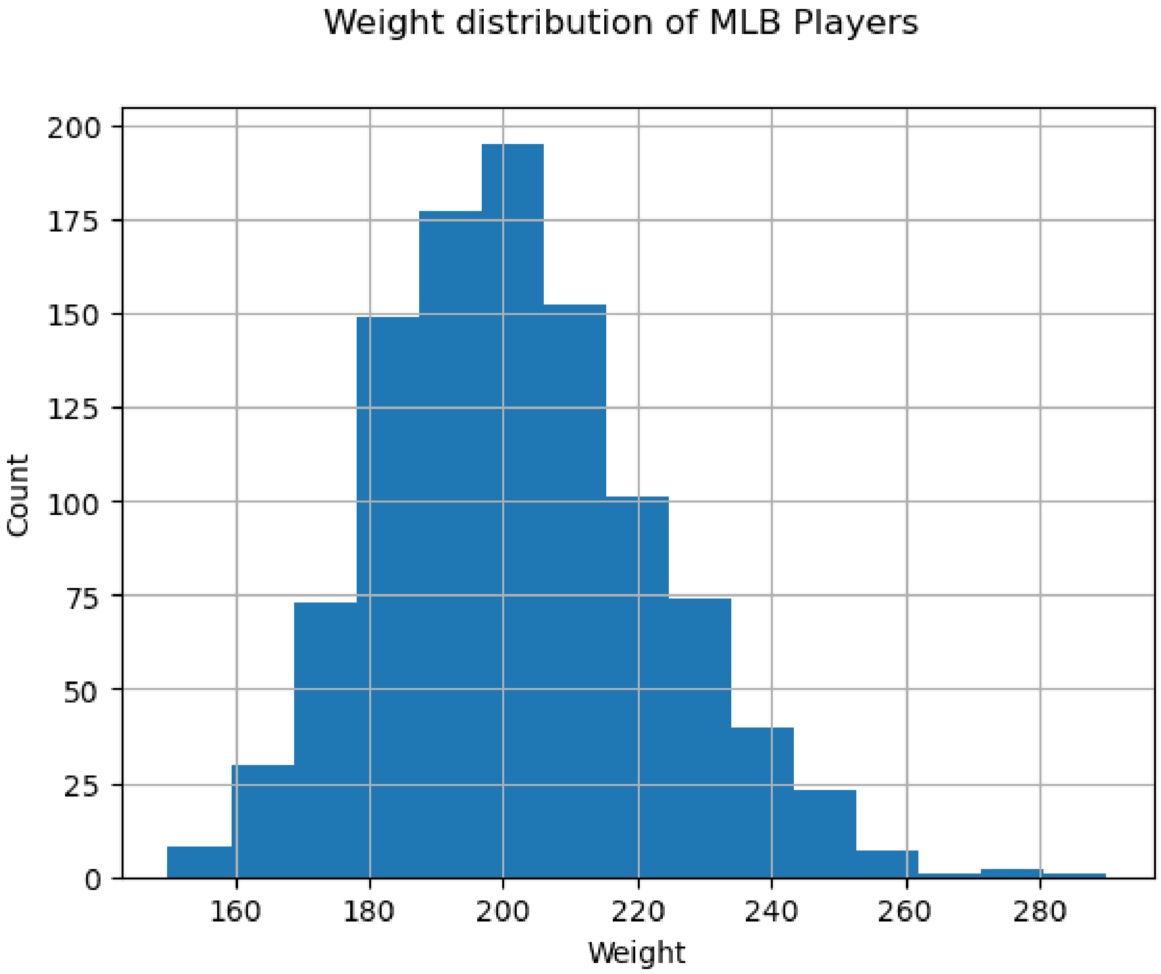
This diagram suggests that, on average, height of first basemen is higher that height of second basemen. Later in this lesson we will learn how we can test this hypothesis more formally, and how to demonstrate that our data is statistically significant to show that.

Real-world Data

> When working with real-world data, we assume that all data points are samples drawn from some probability distribution. This assumption allows us to apply machine learning techniques and build working predictive models.

- To see what the distribution of our data is, we can plot a graph called a **histogram**.
- X-axis would contain a number of different weight intervals (so-called **bins**), and the vertical axis would show the number of times our random variable sample was inside a given interval.

Real-world Data



Real-world Data

- From this histogram you can see that all values are centered around certain mean weight, and the further we go from that weight - the fewer weights of that value are encountered.
- I.e., it is very improbable that the weight of a baseball player would be very different from the mean weight.
- Variance of weights show the extent to which weights are likely to differ from the mean.

> If we take weights of other people, not from the baseball league, the distribution is likely to be different. However, the shape of the distribution will be the same, but mean and variance would change. So, if we train our model on baseball players, it is likely to give wrong results when applied to students of a university, because the underlying distribution is different.

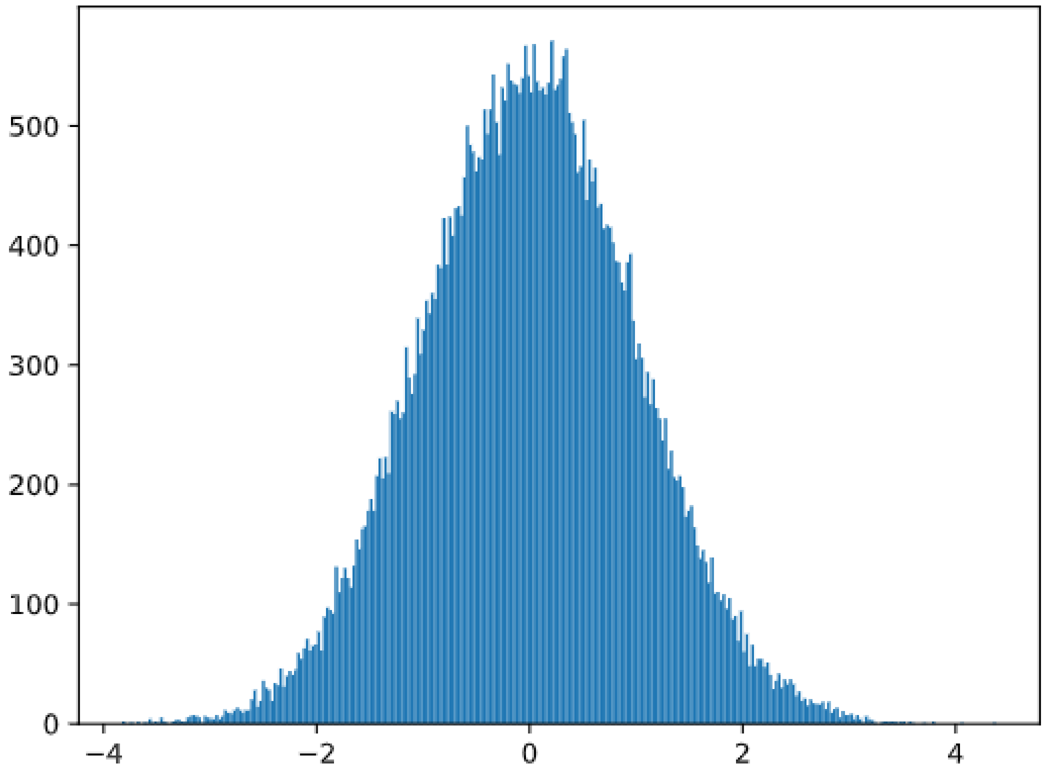
Normal Distribution

- The distribution of weights that we have seen above is very typical, and many measurements from real world follow the same type of distribution, but with different mean and variance.
- This distribution is called **normal distribution**, and it plays a very important role in statistics.
- Using normal distribution is a correct way to generate random weights of potential baseball players.
- Once we know mean weight `mean` and standard deviation `std`, we can generate 1000 weight samples in the following way:

```
samples = np.random.normal(mean,std,1000)
```

Normal Distribution

If we plot the histogram of the generated samples we will see the picture very similar to the one shown above. And if we increase the number of samples and the number of bins, we can generate a picture of a normal distribution that is more close to ideal:



Normal Distribution with mean=0 and std.dev=1

Confidence Intervals

- When we talk about weights of baseball players, we assume that there is certain **random variable** W that corresponds to ideal probability distribution of weights of all baseball players (so-called **population**).
- Our sequence of weights corresponds to a subset of all baseball players that we call **sample**. An interesting question is, can we know the parameters of distribution of W, i.e. mean and variance of the population?
- The easiest answer would be to calculate mean and variance of our sample.
- However, it could happen that our random sample does not accurately represent complete population.
- Thus it makes sense to talk about **confidence interval**.

Confidence Intervals

> **Confidence interval** is the estimation of true mean of the population given our sample, which is accurate is a certain probability (or **level of confidence**).

- Suppose we have a sample $X_1, X_2, \ldots, X_n$ from our distribution.
- Each time we draw a sample from our distribution, we would end up with different mean value μ;.
- Thus μ can be considered to be a random variable.
- A **confidence interval** with confidence p is a pair of values (L_p, R_p), such that $(L_p \leq \mu \leq R_p) = p$, i.e. a probability of measured mean value falling within the interval equals to p.

Confidence Intervals

- It does beyond our short intro to discuss in detail how those confidence intervals are calculated.
- Some more details can be found on Wikipedia.
- In short, we define the distribution of computed sample mean relative to the true mean of the population, which is called **student distribution**.

> **Interesting fact**: Student distribution is named after mathematician William Sealy Gosset, who published his paper under the pseudonym "Student". He worked in the Guinness brewery, and, according to one of the versions, his employer did not want general public to know that they were using statistical tests to determine the quality of raw materials.

Confidence Intervals

- If we want to estimate the mean μ of our population with confidence p, we need to take $(1-p)/2$-th percentile of a Student distribution A, which can either be taken from tables, or computer using some built-in functions of statistical software (eg. Python, R, etc.).
- Then the interval for μ would be given by $X \pm A \times D/\sqrt{n}$, where X is the obtained mean of the sample, D is the standard deviation.

> **Note**: We also omit the discussion of an important concept of degrees of freedom, which is important in relation to Student distribution. You can refer to more complete books on statistics to understand this concept deeper.

An example of calculating confidence interval for weights and heights is given in the accompanying notebooks.

p	Weight mean
0.85	201.73±0.94
0.90	201.73±1.08
0.95	201.73±1.28

Notice that the higher is the confidence probability, the wider is the confidence interval.

Role	Height	Weight	Count
Catcher	72.723684	204.328947	76
Designated_Hitter	74.222222	220.888889	18
First_Baseman	74.000000	213.109091	55
Outfielder	73.010309	199.113402	194
Relief_Pitcher	74.374603	203.517460	315
Second_Baseman	71.362069	184.344828	58
Shortstop	71.903846	182.923077	52
Starting_Pitcher	74.719457	205.163636	221
Third_Baseman	73.044444	200.955556	45

Hypothesis Testing

We can notice that the mean heights of first basemen is higher than that of second basemen. Thus, we may be tempted to conclude that **first basemen are higher than second basemen**.

> This statement is called **a hypothesis**, because we do not know whether the fact is actually true or not.

- However, it is not always obvious whether we can make this conclusion.
- From the discussion above we know that each mean has an associated confidence interval, and thus this difference can just be a statistical error.
- We need some more formal way to test our hypothesis.

Let's compute confidence intervals separately for heights of first and second basemen:

Confidence	First Basemen	Second Basemen
0.85	73.62..74.38	71.04..71.69
0.90	73.56..74.44	70.99..71.73
0.95	73.47..74.53	70.92..71.81

We can see that under no confidence the intervals overlap. That proves our hypothesis that first basemen are higher than second basemen.

More formally, the problem we are solving is to see if **two probability distributions are the same**, or at least have the same parameters. Depending on the distribution, we need to use different tests for that. If we know that our distributions are normal, we can apply **Student t-test**.

- In Student t-test, we compute so-called **t-value**, which indicates the difference between means, taking into account the variance.
- It is demonstrated that t-value follows **student distribution**, which allows us to get the threshold value for a given confidence level p (this can be computed, or looked up in the numerical tables).
- We then compare t-value to this threshold to approve or reject the hypothesis.

In Python, we can use the **SciPy** package, which includes `ttest_ind` function. It computes the t-value for us, and also does the reverse lookup of confidence p-value, so that we can just look at the confidence to draw the conclusion.

For example, our comparison between heights of first and second basemen give us the following results:

```python
from scipy.stats import ttest_ind

tval, pval = ttest_ind(
  df.loc[df['Role']=='First_Baseman',['Height']],
  df.loc[df['Role']=='Designated_Hitter',['Height']],
  equal_var=False)
print(f"T-value = {tval[0]:.2f}\nP-value: {pval[0]}")
```

```
T-value = 7.65
P-value: 9.137321189738925e-12
```

In our case, p-value is very low, meaning that there is strong evidence supporting that first basemen are taller.

Hypothesis Testing

There are also different other types of hypothesis that we might want to test, for example:

- To prove that a given sample follows some distribution. In our case we have assumed that heights are normally distributed, but that needs formal statistical verification.
- To prove that a mean value of a sample corresponds to some predefined value
- To compare means of a number of samples (eg. what is the difference in happiness levels among different age groups)

Law of Large Numbers and Central Limit Theorem

- One of the reasons why normal distribution is so important is so-called **central limit theorem**.
- Suppose we have a large sample of independent N values $X_1, X_2, \ldots, X_N$, sampled from any distribution with mean μ; and variance σ^2.
- Then, for sufficiently large N (in other words, when $N \to \infty$), the mean $\sum_i X_i$ would be normally distributed, with mean μ and variance σ^2/N.

> Another way to interpret the central limit theorem is to say that regardless of distribution, when you compute the mean of a sum of any random variable values you end up with normal distribution.

From the central limit theorem it also follows that, when $N \to \infty$, the probability of the sample mean to be equal to μ becomes 1. This is known as **the law of large numbers**.

Central Limit Theorem

One of the reasons that the normal distribution is **so important** is the following theorem.

Central Limit Theorem. Let $\{X_1, \dots, X_n\}$ be a sample of n random variables chosen identically and independently from a distribution with mean μ and finite variance σ^2. If n is 'large', then

- the sum of the variables $\sum_{i=1}^{n} X_i$ is also a random variable and is approximately **normally** distributed with mean $n\mu$ and variance $n\sigma^2$ and
- the mean of the variables $\frac{1}{n}\sum_{i=1}^{n} X_i$ is also a random variable and is approximately **normally** distributed with mean μ and variance $\frac{\sigma^2}{n}$.

How can we use the central limit theorem (CLT)?

- A binomial random variable is the sum of n Bernoulli random variables.
- So the CLT tells us that if n is large, binomial random variables will be distributed approximately normally.
- That is, if we flip a coin many times, the number of heads that we're likely to see is described by a normal distribution.
- This provides a different (easier) way to answer the question: `How unlikely is it to flip a fair coin 1000 times and see 545 heads?`

How can we use the central limit theorem (CLT)?

Suppose we flip a fair ($p = 0.5$) coin 1000 times.

Question: How many heads do we expect to see?

The CLT says that the number of heads (= sum of Bernoulli r.v. = binomial r.v.) is approximately normally distributed with mean

$$n\mu = np = 1000 * 0.5 = 500$$

and variance

$$n\sigma^2 = np(1-p) = 1000 * 0.5 * 0.5 = 250.$$

Summary of hypothesis testing and the z-test

- Identify the parameter of interest and describe it in the context of the problem.
- Determine the null and alternative hypotheses.
- Choose a significance level α.
- Find the formula for the computed value of the test statistic, *e.g.*, $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$ (use the CLT). (Note as we saw with the coin tosses, if the data is binary and H_A involves a proportion, applying the CLT to a sum of Bernoulli's gives $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ where $\hat{p}$ is the sample proportion.)
- Using the sampled data, compute the P-value, *e.g.*, $F(z)$
- Compare the significance level to the P-value to decide whether or not the null hypothesis should be rejected and state the conclusion in the problem context. Report the P-value!

One- and two- sided hypothesis testing:

Depending on the null and alternative hypothesis, the P-value will be different integrals of the 'bell curve'. This is called one- and two- sided hypothesis testing.

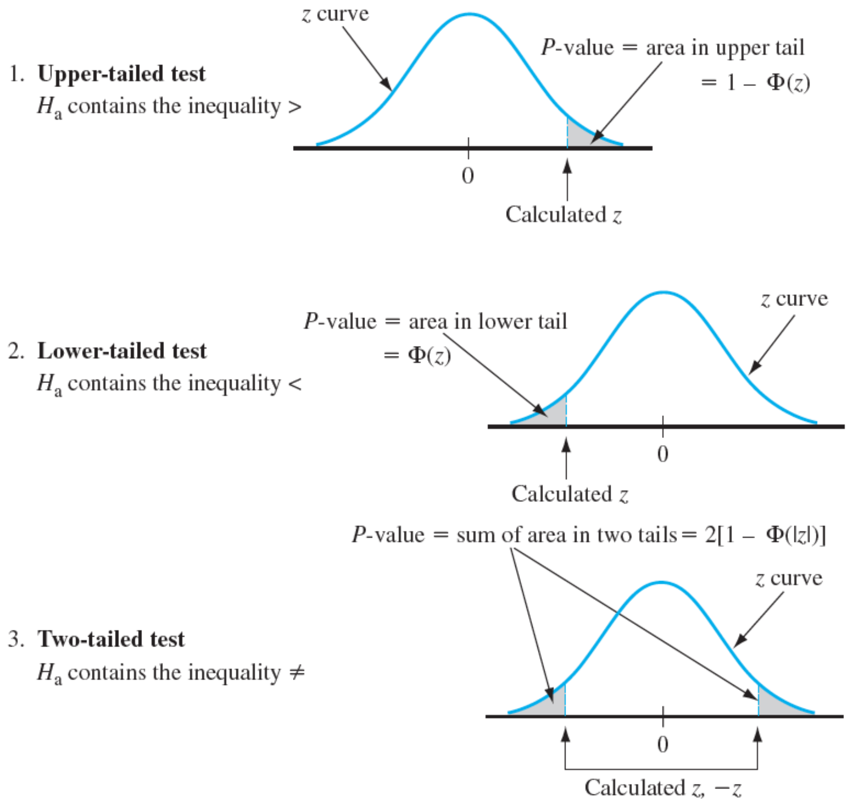

1. Upper-tailed test
H_a contains the inequality >
z curve
P-value = area in upper tail
= 1 − Φ(z)
0
Calculated z
2. Lower-tailed test
H_a contains the inequality <
P-value = area in lower tail
= Φ(z)
z curve
0
Calculated z
3. Two-tailed test
H_a contains the inequality ≠
P-value = sum of area in two tails = 2[1 − Φ(|z|)]
z curve
0
Calculated z, −z

Figure 8.4 Determination of the *P*-value for a *z* test

Types of error in hypothesis testing

In hypothesis testing, there are two types of errors.

- A **type I error** is the incorrect rejection of a true null hypothesis (a "false positive").
- A **type II error** is incorrectly accepting a false null hypothesis (a "false negative").

Depending on the application, one type of error can be more consequential than the other.

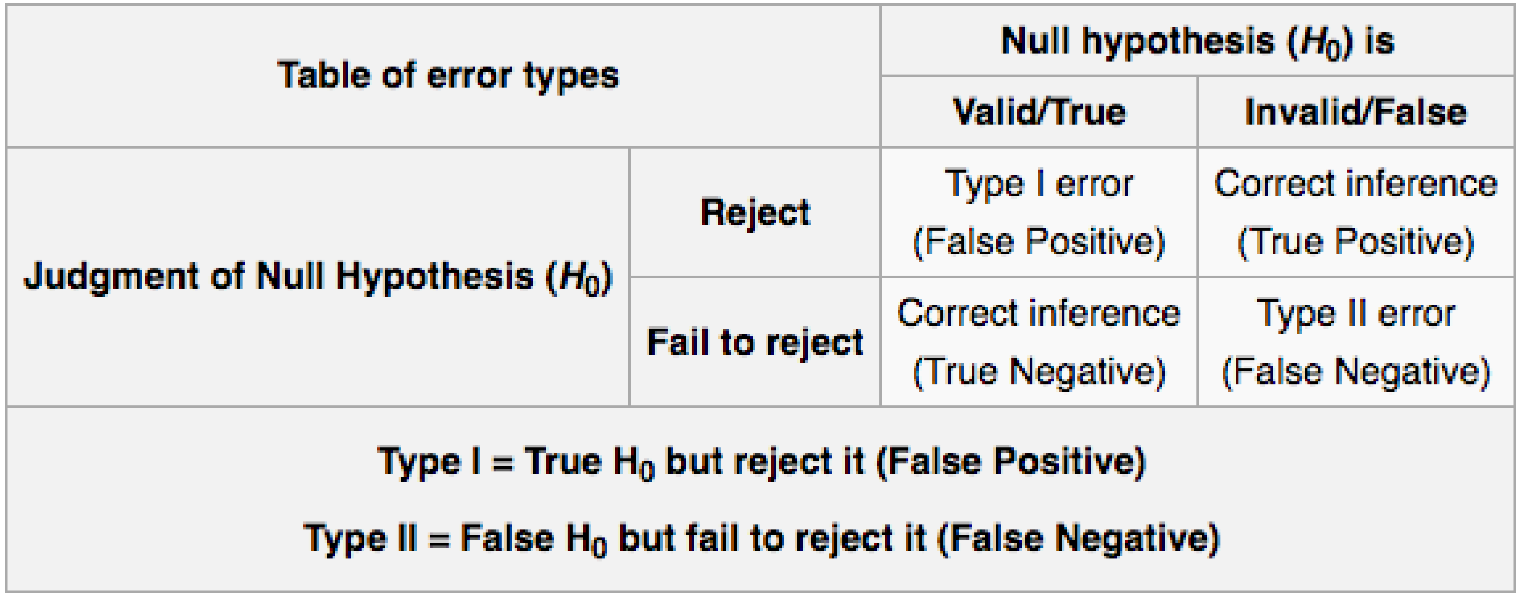


Table of error types		Null hypothesis (H_0) is	
		Valid/True	Invalid/False
Judgment of Null Hypothesis (H_0)	Reject	Type I error (False Positive)	Correct inference (True Positive)
	Fail to reject	Correct inference (True Negative)	Type II error (False Negative)
Type I = True H_0 but reject it (False Positive) Type II = False H_0 but fail to reject it (False Negative)			

Types of error in hypothesis testing

- The probability of making a type I (false positive) error is the significance level α.
- The probability of making a type II (false negative) error is more difficult to calculate.

Examples

(1) In drug testing, we take the null hypothesis H_0: "This drug has no effect on the disease." A type I error detects an effect (the drug cures the disease) that is not present. A type II error fails to detect an effect (the drug cures the disease) that is present.

(2) In a trial, we take the null hypothesis H_0: "This man is innocent." A type I error convicts an innocent person. A type II error lets a guilty person go free.

P hacking

Some comments about the p-value:

- A p-value is a probability calculated assuming that H_0 is true.
- The smaller the p-value, the stronger the evidence against H_0.
- A p-value is not the probability that the null hypothesis is true or false. It is the probability that an erroneous conclusion is reached.

Recently the *misuse* of hypothesis testing (p-values) has raised considerable controversy. Basically, if you do enough hypothesis tests, eventually you'll have a Type I (false positive) error. This is sometimes referred to as Data dredging. This is a real problem in a world with tons of data in which it is easy to do many, many hypothesis tests automatically. One method to avoid this is called *cross-validation*, which we'll discuss later.

Descriptive vs. Inferential Statistics

- Descriptive statistics quantitatively describe or summarize features of a dataset.
- Inferential statistics attempts to learn about the population from which the data was sampled.

> **Example**: The week before a US presidential election, it is not possible to ask every voting person who they intend to vote for. Instead, a relatively small number of individuals are surveyed. The *hope* is that we can determine the population's preferred candidate from the surveyed results.

- Often, we will model a population characteristic as a *probability distribution*.
- *Inferential statistics* is deducing properties of an underlying probability distribution from sampled data.

Conclusion

In this section, we have learnt:

- basic statistical properties of data, such as mean, variance, mode and quartiles
- different distributions of random variables, including normal distribution
- how to find correlation between different properties
- how to use sound apparatus of math and statistics in order to prove some hypotheses,
- how to compute confidence intervals for random variable given data sample

While this is definitely not exhaustive list of topics that exist within probability and statistics, it should be enough to give you a good start into this course.

Extra: A/B testing

A/B testing is a method of comparing two or more versions of an advertisement, webpage, app, etc. We set up an experiment where the variants are shown to users at random and statistical analysis is used to determine which is best. AB testing is the *de facto* test for many business decisions.

Example.

A/B testing was extensively used by President Obama during his 2008 and 2012 campaigns to develop

- optimized fund-raising strategies,
- get-out-the-vote programs that would be most beneficial, and
- target ads to the most susceptible audiences.

Extra: A/B testing

Example.

Suppose your company is developing an advertisement. The art department develops two internet ads: "Ad A" and "Ad B". Your job is to figure out which is better.

You decide to do an experiment: You use Google ads to randomly show 1000 internet users Ad A and 1000 internet users Ad B.

It turns out that 500 Ad A viewers click on the ad while 550 Ad B viewers click on the ad? Obviously Ad B did better, but is the difference "significant" enough to say that Ad B is better? Or perhaps Ad B just got lucky in this test?

Credits

This lesson has been authored by Dmitry Soshnikov