Introduction to Probability & Statistics

GENERAL ASSEMBLY

aka How to Lie with Statistics

Agenda

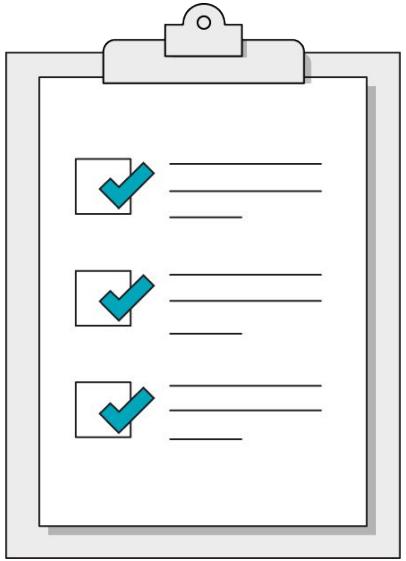
Introduction to Statistics & Probability

Definitions & Key Concepts

Statistical Simulations & Models in Python

Our Learning Goals

- Understand how statistics help us work with data
- Use the common tools and methods used to understand relationships within and between variables
- Deploy fundamental concepts and rules of probability
- Solve probability problems using simulations.



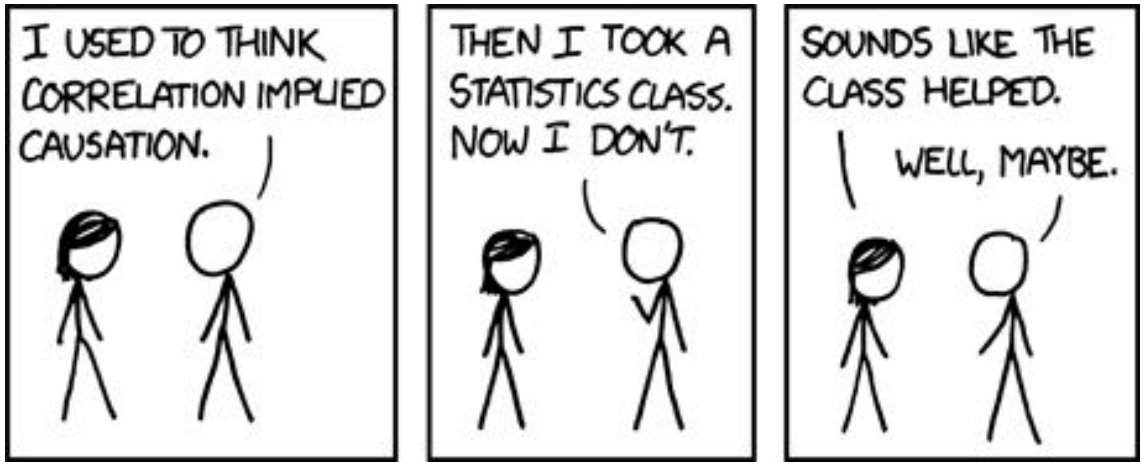
Motivation

- We've already discussed the visual presentation of information, and even exploratory data analysis
- We've thrown around words like 'variable,' 'average,' and 'observation'
- But we haven't yet really talked about data and statistics and how those things are related or even, really, data science

What Are We Going to Do Today?

- We're going to review core concepts in **descriptive statistics**, also called **summary statistics**, and **probability**
- We'll also review how to get and use some of these statistics using python, as well as how to talk about these core tools
- And we're going to do it by being really cynical and looking at how these core concepts can mislead

Obligatory XKCD comic



Credit where credit is due: XKCD comics

How to Lie with Statistics

Let’s Look at Some Data!

Words to Describe a Dataset

- First and foremost: a DataFrame is not a dataset; it is a tool we use to represent a dataset or a part of a dataset
- Each row represents an **observation**: the data we have collected about one single thing
- Each column represents a **variable**: a feature or aspect of the thing that we have measured
- Variables are either **quantitative** or **qualitative**
    - Qualitative variables are unordered, exclusive things; usually categorical
    - Quantitative variables are numerical measures of size, magnitude, quantity
- Knowing the **number of observations** and the **number of variables** is helpful too

Words to Describe Datasets

We'll also talk about quantitative variables being **continuous** or **discrete**

Continuous variables have an infinite number of possible values. Usually, you can tell a continuous variable by the presence of a decimal point. Usually.

Discrete variable, on the other hand, take on a finite number of values. Usually, these are counts, like the number of people in a Zoom room or the number of children in a family

There are also **ordinal variables** which relate to ordered rankings, like the results of a horserace or a customer satisfaction survey.

Words to Describe Datasets

Data is the collection of observations that we, as data scientists, poke, prod, slice, dice, analyze, and use to teach machines to make decisions.

A **statistic** is any of a number of ways that we summarize the data because we aren't very good at keeping vast tables in our heads.

If you dive deep into formal mathematical statistics, you'll also encounter **sample** and **population**. A population is the complete and entire group of things that we're interested in. Sadly, we can usually only ever get a small subset from that population, the sample. The sample is usually our dataset - it's what we can get a hold of to work with.

How to Lie with Statistics

Ways to Summarize Data

People, as a rule, don’t do well with gigantic tables

Summarizing Categorical Variables

There's really only one good way to summarize categorical variables: **counts** and **normalized counts** (percents). Visually, we use barcharts.

Summarizing Quantitative Data

When summarizing quantitative data (usually continuous variables, but not always) we're interested in two things:

- Central Tendency - where the data group together most
- Variability - how much the data moves around that central tendency
- Relationship - if and how variables move together

We are summarizing features - so generally we'll perform these calculations on specific columns in our dataset.

Measures of Central Tendency

There's three measures of central tendency.

Mean is the average of values - add them all up, and divide by the number of values you added.

$$\left(\frac{1}{n}\sum_{i=i}^{n} x_i\right)$$

Where
n = the number of items to sum
∑ = capital sigma; symbol for summation
i = the starting point and index position for *x*
x = the item to sum

Measures of Central Tendency

Median is the middle value in an ordered list of values, calculated by sorting the list of values and grabbing the one in the middle. If the list has an even number of values, average the two in the middle.

Interpretation-wise, when reporting a median value we are saying that half the values in our list are above the median, and half are below the median.

The median tends to be robust against outliers: it doesn't give them additional weight in the calculation.

Measures of Central Tendency

The **mode**, sometimes **modal category**, is just the value (or range of values) that shows up most often. Sometimes there's more than one. We call it a **bimodal distribution** if there's two modes.

Sometimes the mode is useful, sometimes it's not. A histogram here is always your friend.

How to Lie with Statistics

Now for the fun part. I'm going to show you a few slides and create slack threads for each one. In the thread, respond with how the use of a particular summary statistic might be misleading.

Credit where credit is due: the illustrations for this section come from the delightful website Math with Bad Drawings

How to Lie with Statistics

Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

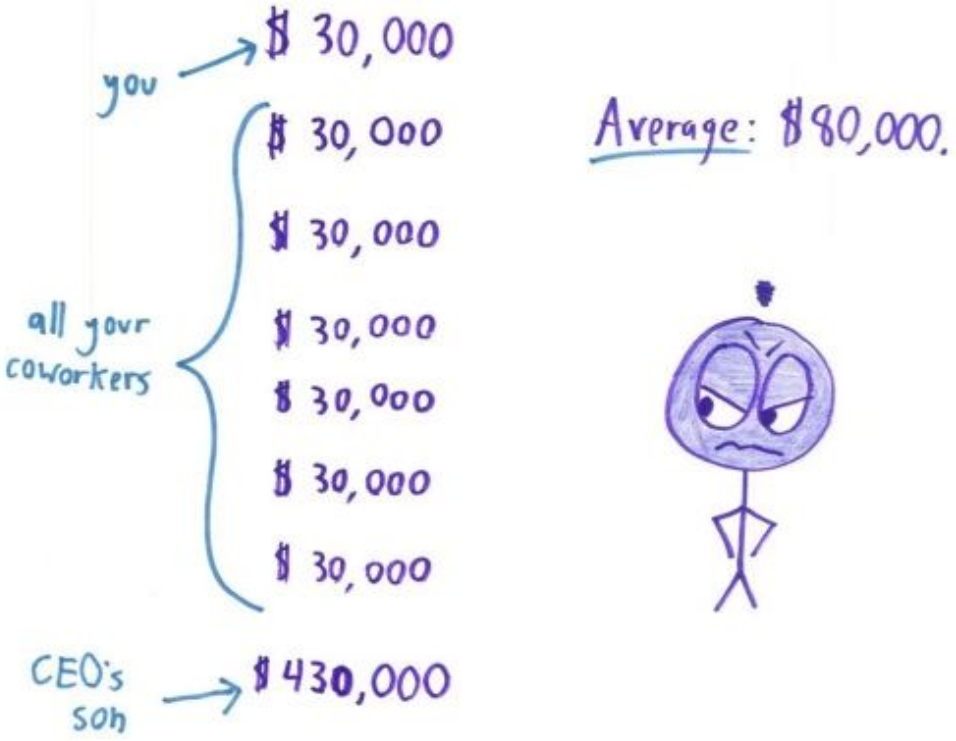


Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

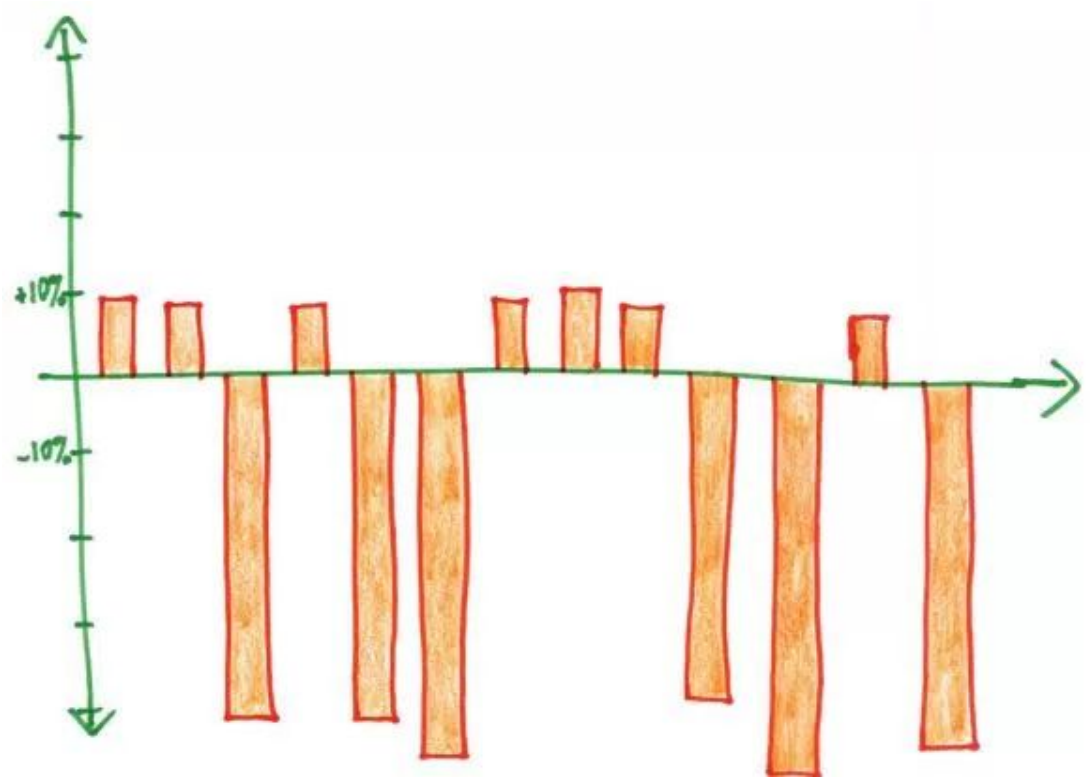


Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

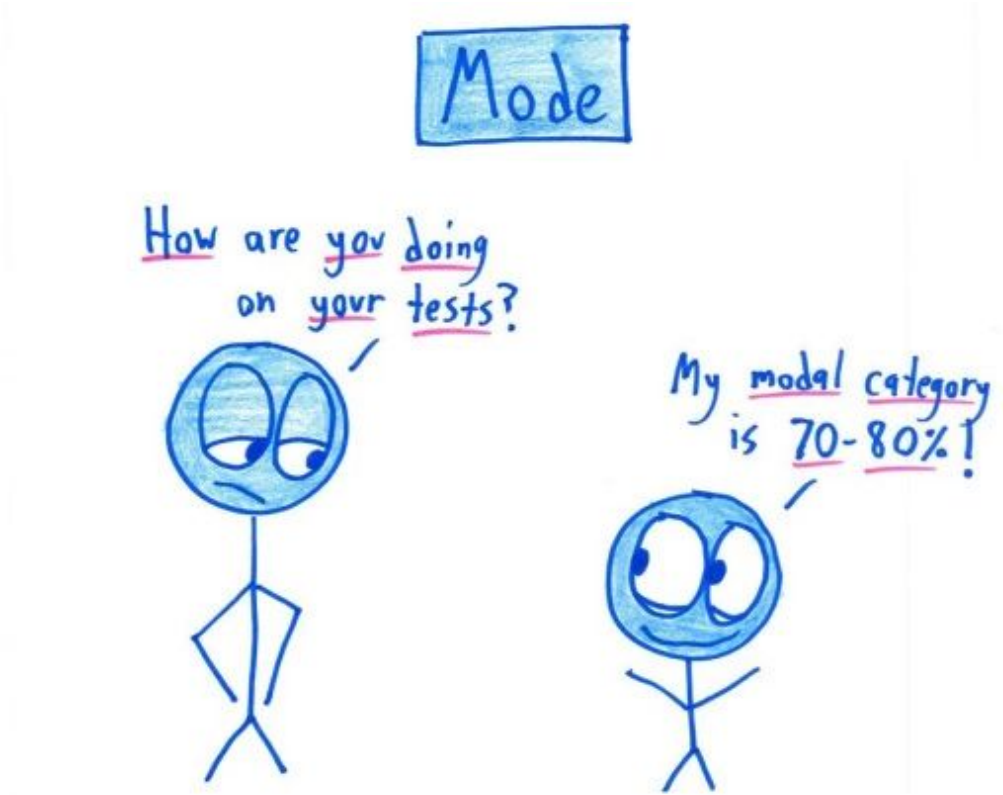


Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

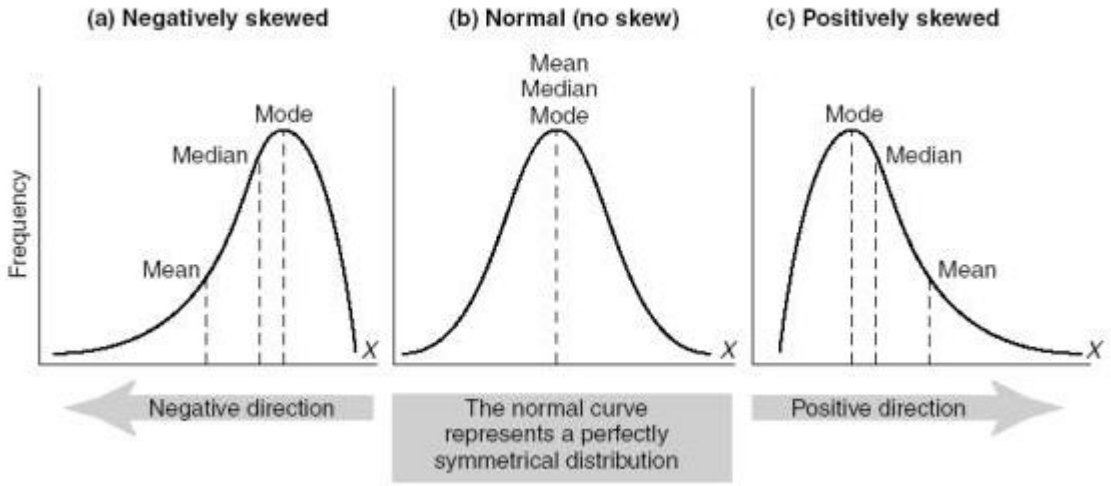
Score Category	Number of Tests
90s	0
80s	0
70s	2
60s	1
50s	1
40s	1
30s	1
20s	1

Illustration from Mathwithbaddrawings.com

Slight Aside: Skew

- It's sadly somewhat rare for data to be symmetric, and for the mean, median, and mode to be equal
- Frequently, data are skewed to the left (aka negative skew) or to the right (aka positive skew). The direction of the skew corresponds to the direction of the 'tail' - those trailing values on either side of the center of the distribution.
- A lot of the algorithms we use, however, assume that the data are normally distributed.

Skew



Symmetric: mean == median == mode
Negative (left) skew: mean < median < mode
Positive (right) skew: mode < median < mean

Measures of Variability: Range

In addition to central tendency, we're interested in how data vary around that central tendency.

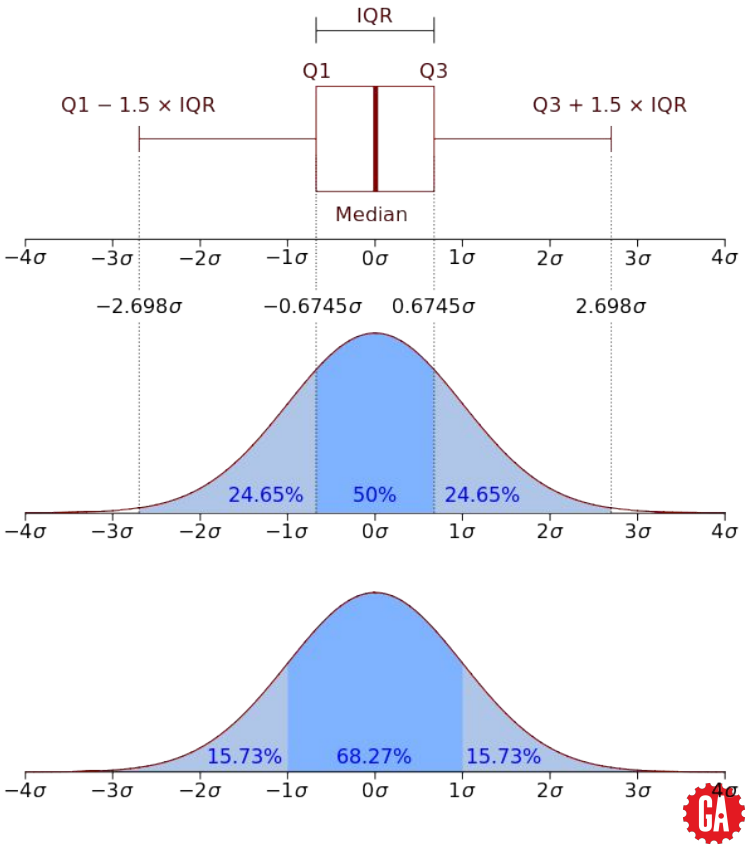
Range tells us the how widely the data are spread. It's just the maximum of the feature minus the minimum of the feature.

Measures of Variability: Quartiles and IQR

We'll frequently also see **quartiles** and the **interquartile range** (IQR), especially in box plots. Quartiles divide the data into four more-or-less equal groups (that is, the same number of values is in each group), providing information about both the center of the data and the spread.

Quartiles and the Interquartile Range (IQR)

- Quartiles divide an ordered data set into four equal parts.
- The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.
- The interquartile range (IQR) is a measure of variability, based on dividing a data set into quartiles. It is the “middle 50” of your data. Also called the H-spread. IQR = Q3-Q1
- Outliers: Q1 - 1.5(IQR), Q3 + 1.5(IQR)



Measures of Variability: Variance and Standard Deviation

Variance and **standard deviation** tell us about the typical distance of our observations from the center of the data, i.e., the mean. The formula for variance is:

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

Where:
σ^2 = lower-case sigma, squared; symbol for variance
$\bar{x}$ = "x-bar"; symbol for 'mean of *x*'

In pythonic pseudocode:

```
X = [some list of values]
total = []
for x in X:
        total.append((x - X.mean())^2)
total_variance = sum(total)/len(total)
```


Measures of Variability: Variance and Standard Deviation

And **standard deviation** is simply the square root of the variance:

$$\sqrt{\sigma^2}$$

Note that standard deviation is expressed in the units of the variable, while variance is in squared units of the variable

How to Lie with Statistics, Part II

Your turn, once again. I'm going to show you a few slides and create slack threads for each one. In the thread, respond with how the use of a particular summary statistic might be misleading.

How to Lie with Statistics

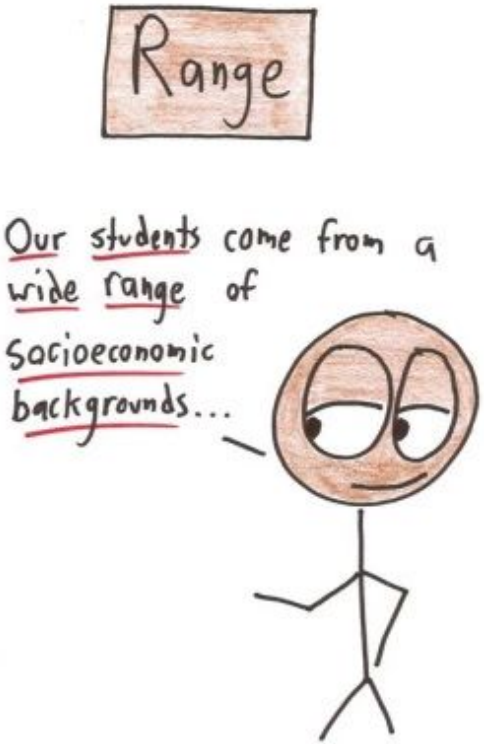


Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

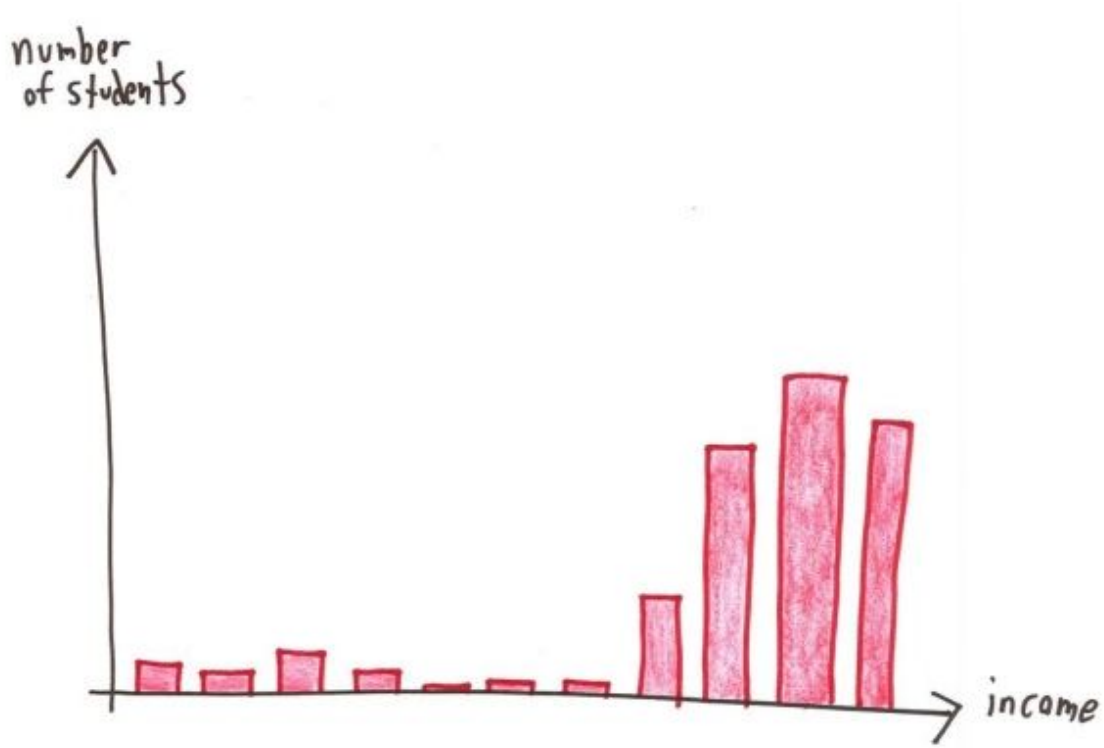


Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

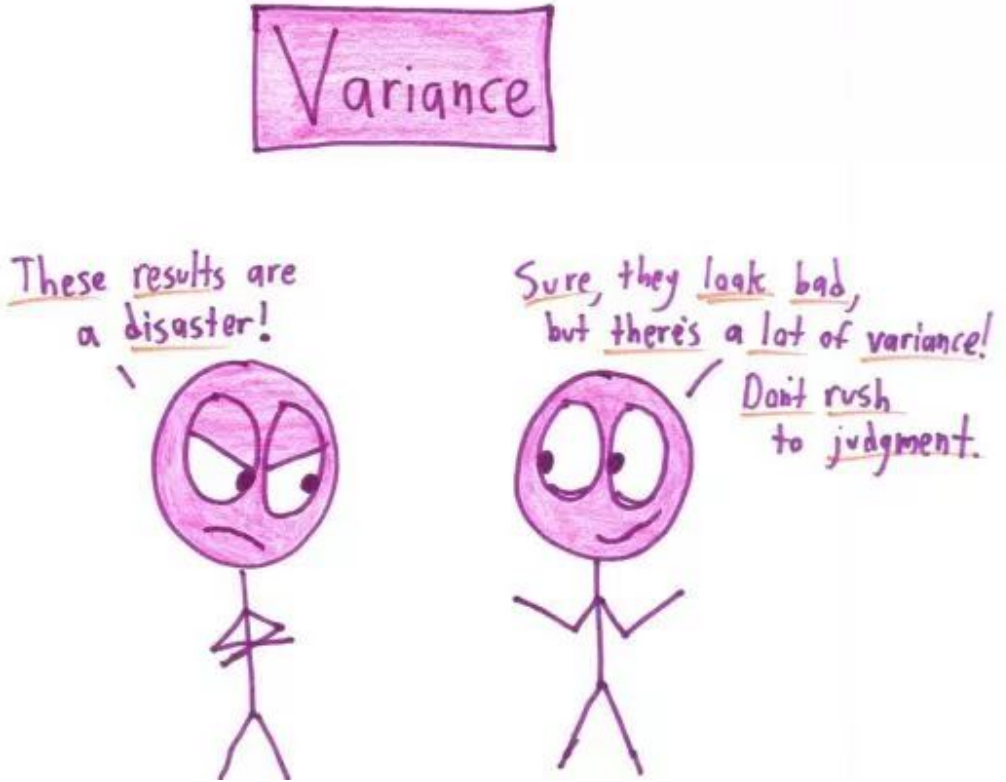


Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

Illustration from Mathwithbaddrawings.com

Measures of Relation: Correlation and Covariance

So far, we've really only talked about summarizing single variables - what we call **univariate** analysis. But dealing with only one variable would get boring. So we're also interested in how two (or more) variables change *together* (**bivariate** and **multivariate** analysis, respectively), which are measured with **covariance** and **correlation.**

Correlation is reported more often than covariance because it's directly interpretable. But covariance is used in the calculation of correlation.

Measures of Relation: Correlation and Covariance

There are several different flavors of correlation coefficient. The most common is **Pearson correlation coefficient** (this is the default in pandas). Without getting into the formula, here's what you need to know:

- Values close to -1 or +1 indicate a strong and linear relationship between the two variables.
- Values close to 0 indicate a weak and/or nonlinear relationship between the two variables.
- Values above 0 indicate a positive relationship between the two variables.
- Values below 0 indicate a negative relationship between the two variables.

Measures of Relation: Correlation and Covariance

Visually, correlation becomes more intelligible.

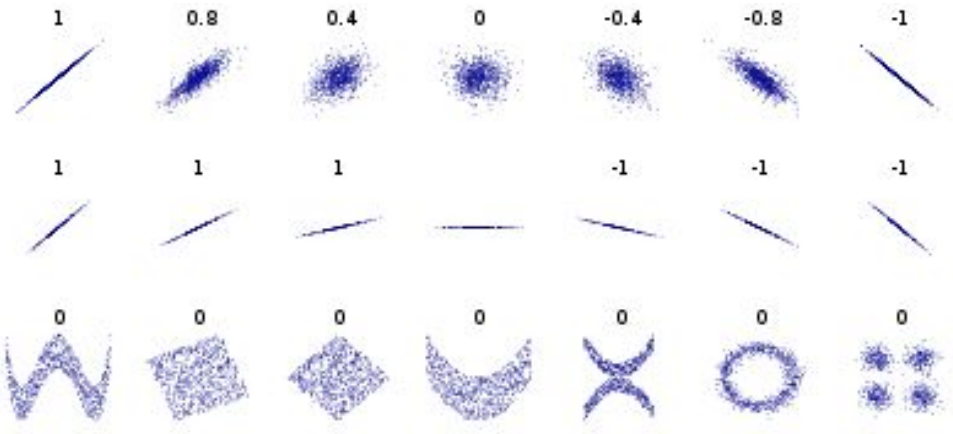


Image from [Wikipedia's article on correlation and dependence]

How to Lie with Statistics, Part III

I’m going to show you one last slide that demonstrates an abuse of correlation. In the thread, explain how this might be misleading.

How to Lie with Statistics

Illustration from Mathwithbaddrawings.com

How to Lie with Statistics

Illustration from Mathwithbaddrawings.com

Summary

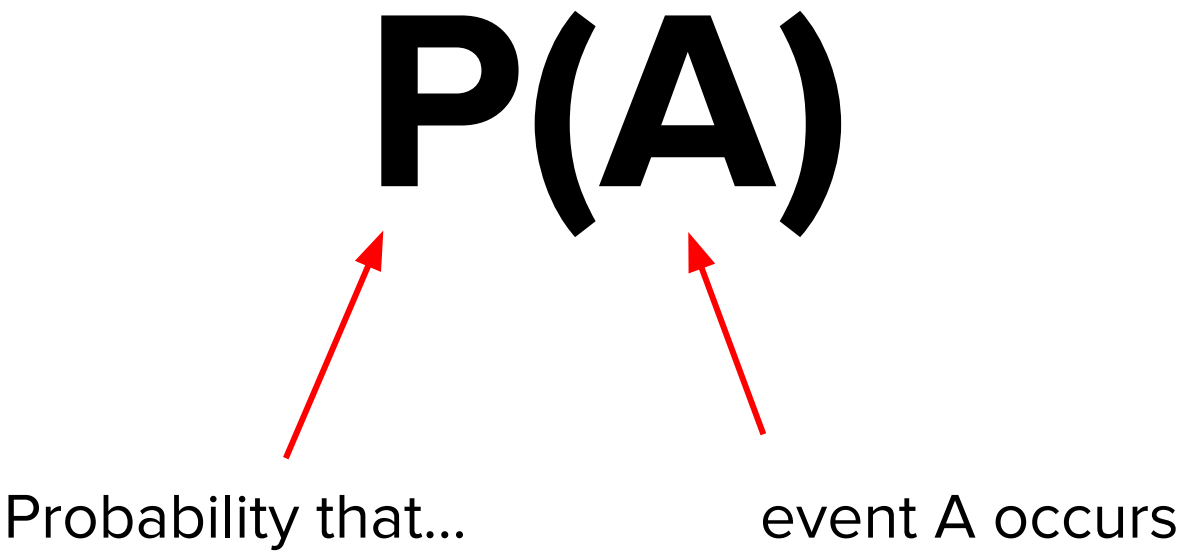
- The statistics of interest to us, mostly, in this course are **descriptive statistics** - the tools we use to slice, dice, and combine our data to understand what's in it
- Primarily, we're interested in
  - **measures of central tendency** (mean, median, mode, quartiles),
  - **measures of variability** (variance, standard deviation),
  - and the **relations between variables** (correlation and covariance)

Introductory Probability

Probability

- We live and operate in an uncertain world: we don't know what is going to happen next.
- **Probability theory** is the mathematical field that has built the tools and language to quantify uncertainty**.**
- The **probability** of an event is simply how likely that event is to happen

Probability



Probability

$$P(A) = \frac{\text{Number of time A happens}}{\text{Number of trials}}$$

Key Probability Facts

1. **P(S) = 1** : The probability that *something* happens is 1

2. **P(∅) = 0** : The probability that *nothing* happens is 0

3. **0 ≤ P(A) ≤ 1** : The probability of any given event is between 0 and 1

Three Kinds of Probability

Marginal probability is the probability of an event for one variable, regardless of the outcomes for other variables. Or, more plainly, the probability for one variable.

Joint probability is the probability of two events happening, denoted **P(A and B)**

Conditional probability is the probability of an event occurring, *given that* some other event has already occurred, denoted **P(A|B)**, read as 'probability of A given B'

Three Kinds of Probability

Let’s take a standard deck of cards.

Three Kinds of Probability

Let's take a standard deck of cards: 13 ranks (A-K) in four suits in two colors, for a total of 52 cards.

The **marginal probability** is the probability of drawing a red card, any red card.

$$P(R) = \frac{13*2}{52} = \frac{26}{52} = 0.5 = 50\%$$

Three Kinds of Probability

Let's take a standard deck of cards.

The **joint probability** is the probability of drawing a red Queen (joint because the card is red and is a Queen.)

$$P(\text{R and Q}) = \frac{2}{52} = 0.03846\ldots \approxeq 3.85\%$$

Three Kinds of Probability

Let's take a standard deck of cards.

The **conditional probability** is the probability of drawing a Queen, given that you have drawn a red card.

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{0.0385}{0.5} = 7.7\%$$

Independence

The concept of independence shows up over and over again in our course.

Two events are said to be **independent** if the probability of one occurring does not affect the probability of the other occurring.

Independence

The concept of independence shows up over and over again in our course.

Two events are said to be **independent** if the probability of one occurring does not affect the probability of the other occurring.

Alternatively, A and B are **independent** if:

$$P(A \mid B) = P(A)$$

Independence

A and B are **independent** if:

$$P(A \mid B) = P(A)$$

So,

$$P(A \cap B) = P(A)P(B)$$

Which of the following are independent?

Scenario	A	B
1	Flipping a heads on a coin	Rolling a 1 on a six-sided die
2	Charlie hits the snooze button on his alarm clock	Someone in Paris, whom Charlie does not know, hits the snooze button on their alarm clock
3	Charlie has pizza for lunch	Charlie has pizza for dinner
4	The amount of shark attacks on a given day is high	The amount of ice cream sales on that same day is high
5	Today's high temperature is 76°	Tomorrow's high temperature is 76°

Which of the following are independent?

Scenario	A	B
1	Flipping a heads on a coin	Rolling a 1 on a six-sided die
2	Charlie hits the snooze button on his alarm clock	Someone in Paris, whom Charlie does not know, hits the snooze button on their alarm clock
3	Charlie has pizza for lunch	Charlie has pizza for dinner
4	The amount of shark attacks on a given day is high	The amount of ice cream sales on that same day is high
5	Today's high temperature is 76°	Tomorrow's high temperature is 76°

Which of the following are independent?

Scenario	A	B	Why are these not independent?
3	Charlie has pizza for lunch	Charlie has pizza for dinner	If I have pizza for lunch, I will probably ~~not~~ have pizza for dinner.
4	The amount of shark attacks on a given day is high	The amount of ice cream sales on that same day is high	While one does not *cause* the other, if the amount of shark attacks is high, it is because the weather is nice, which means ice cream sales are likely to be high
5	Today's high temperature is 76°	Tomorrow's high temperature is 76°	If it's nice today, it will more likely be nice tomorrow, too

When by hand is tough...

When these problems get too hard, why do the math? We can often find the approximate answers via **simulation**.

Why does this work?

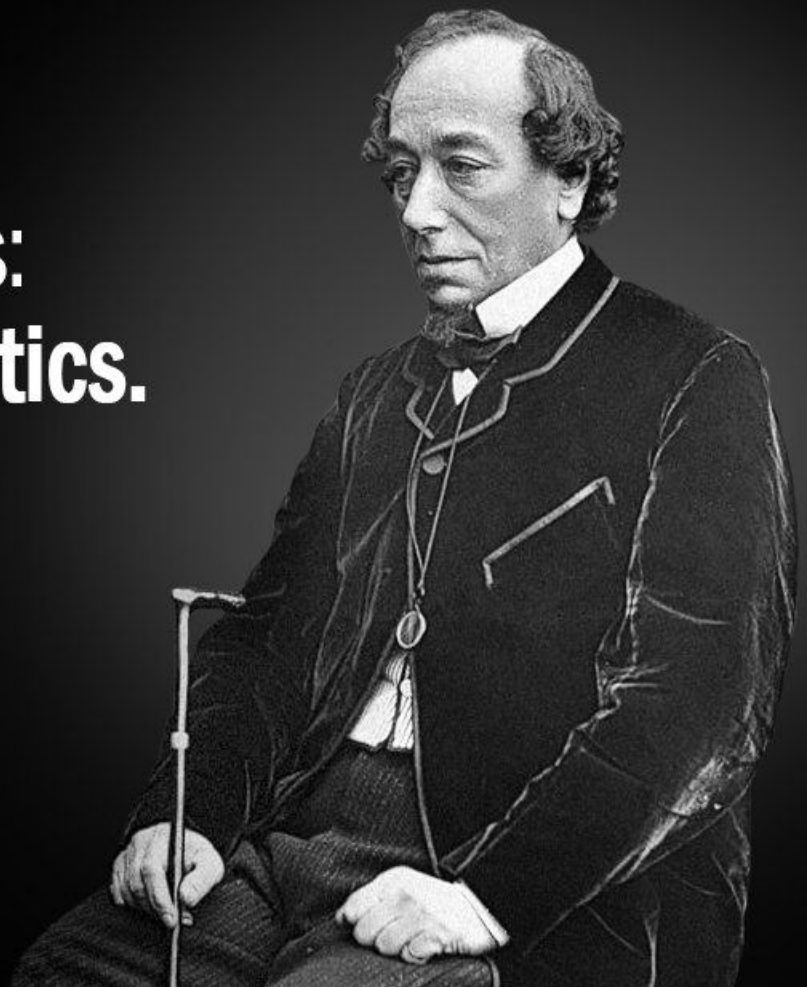
The Law of Large Numbers (LLN)

The **Law of Large Numbers** states that if you were to repeat an experiment infinitely many times and some other conditions which are always met in practice, then the simulated probability of an event approaches the true probability.*

*Ok, the real LLN is more complicated and has many more implications than this, and there are many different kinds of LLNs. This is a good gist, though.

There are three kinds of lies:
lies, damned lies, and statistics.

– Benjamin Disraeli



AZ QUOTES