Statistics

Tushar B. Kute,
http://tusharkute.com

Objectives

- Statistics: Describing a Single Set of Data,
- Correlation,
- Simpson’s Paradox,
- Some Other Correlational Caveats,
- Correlation and Causation

What is statistics?

- Statistics is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.
- In applying statistics to a scientific, industrial, or social problem, it is conventional to begin with a statistical population or a statistical model to be studied.
- Populations can be diverse groups of people or objects such as "all people living in a country" or "every atom composing a crystal". Statistics deals with every aspect of data, including the planning of data collection in terms of the design of surveys and experiments

What is statistics?

- When census data cannot be collected, statisticians collect data by developing specific experiment designs and survey samples.
- Representative sampling assures that inferences and conclusions can reasonably extend from the sample to the population as a whole.
- An experimental study involves taking measurements of the system under study, manipulating the system, and then taking additional measurements using the same procedure to determine if the manipulation has modified the values of the measurements.
- In contrast, an observational study does not involve experimental manipulation.

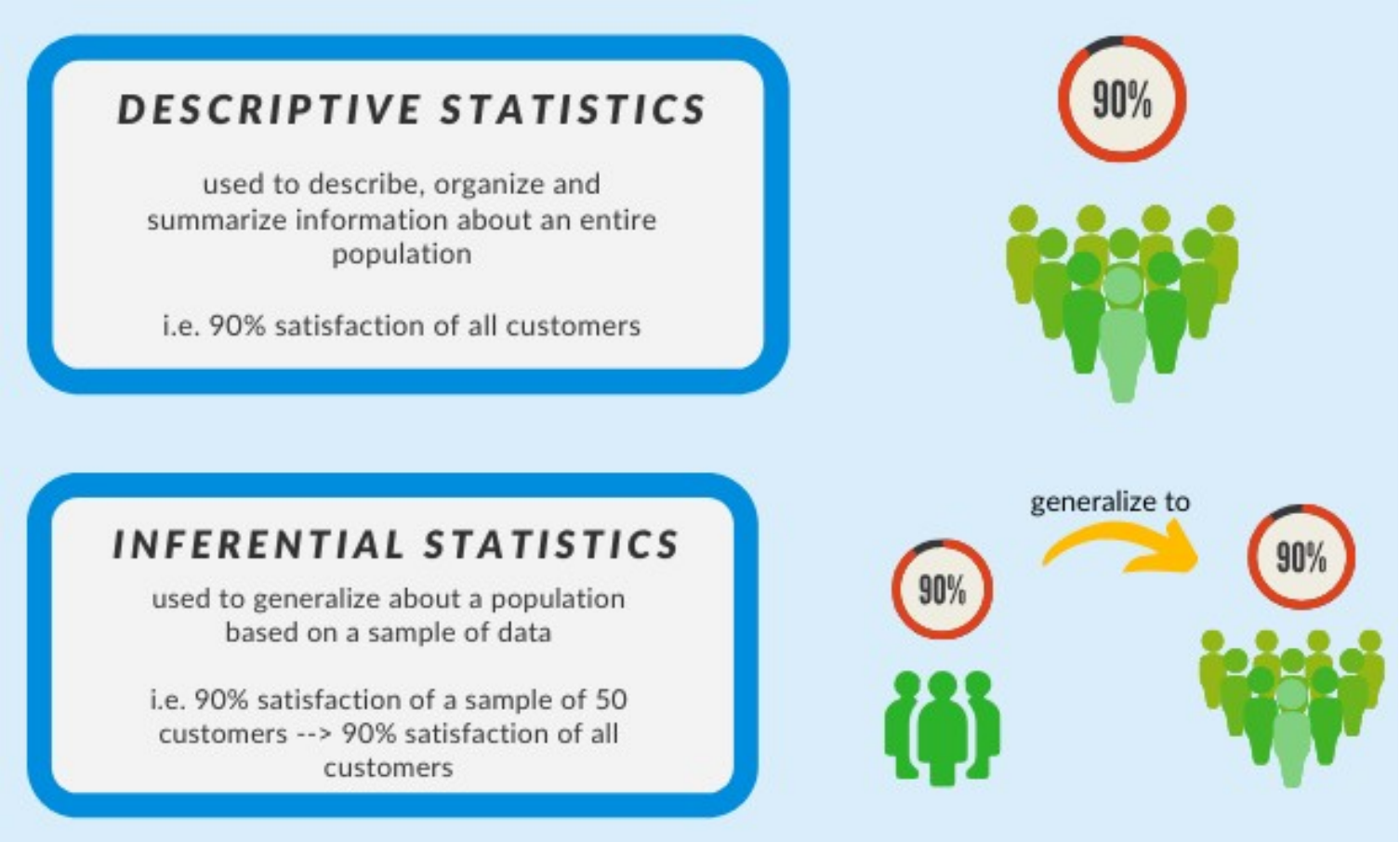
Descriptive Statistics

- A descriptive statistic (in the count noun sense) is a summary statistic that quantitatively describes or summarizes features of a collection of information, while descriptive statistics in the mass noun sense is the process of using and analyzing those statistics.
- Descriptive statistics is distinguished from inferential statistics (or inductive statistics), in that descriptive statistics aims to summarize a sample, rather than use the data to learn about the population that the sample of data is thought to represent.

Inferential Statistics

- Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution.
- Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates.
- It is assumed that the observed data set is sampled from a larger population. Inferential statistics can be contrasted with descriptive statistics.
- Descriptive statistics is solely concerned with properties of the observed data, and it does not rest on the assumption that the data come from a larger population.

Comparing


DESCRIPTIVE STATISTICS
used to describe, organize and summarize information about an entire population
i.e. 90% satisfaction of all customers
90%
INFERENTIAL STATISTICS
used to generalize about a population based on a sample of data
i.e. 90% satisfaction of a sample of 50 customers --> 90% satisfaction of all customers
90%
generalize to
90%

Why Statistics?

- What features are the most important?
- How should we design the experiment to develop our product strategy?
- What performance metrics should we measure?
- What is the most common and expected outcome?
- How do we differentiate between noise and valid data?

From data to knowledge

- In isolation, raw observations are just data. We use descriptive statistics to transform these observations into insights that make sense.
- Then we can use inferential statistics to study small samples of data and extrapolate our findings to the entire population.

Terminologies of statistics

- Population: It is an entire pool of data from where a statistical sample is extracted. It can be visualized as a complete data set of items that are similar in nature.
- Sample: It is a subset of the population, i.e. it is an integral part of the population that has been collected for analysis.
- Variable: A value whose characteristics such as quantity can be measured, it can also be addressed as a data point, or a data item.

Terminologies of statistics

- Distribution: The sample data that is spread over a specific range of values.
- Parameter: It is a value that is used to describe the attributes of a complete data set (also known as ‘population’). Example: Average, Percentage
- Quantitative analysis: It deals with specific characteristics of data- summarizing some part of data, such as its mean, variance, and so on.
- Qualitative analysis: This deals with generic information about the type of data, and how clean or structured it is.

Statistical Machine Learning

- The methods used in statistics are important to train and test the data that is used as input to the machine learning model. Some of these include outlier/anomaly detection, sampling of data, data scaling, variable encoding, dealing with missing values, and so on.
- Statistics is also essential to evaluate the model that has been used, i.e. see how well the machine learning model performs on a test dataset, or on data that it has never seen before.
- Statistics is essential in selecting the final and appropriate model to deal with that specific data in a predictive modelling situation.
- It is also needed to show how well the model has performed, by taking various metrics and showing how the model has fared.

Describing single set of data

- Practically...

Dispersion

- Dispersion refers to measures of how spread out our data is.
- Typically they’re statistics for which values near zero signify not spread out at all and for which large values (whatever that means) signify very spread out.

Variance

- In statistics, the variance is a measure of how far individual (numeric) values in a dataset are from the mean or average value.
- The variance is often used to quantify spread or dispersion. Spread is a characteristic of a sample or population that describes how much variability there is in it.
- A high variance tells us that the values in our dataset are far from their mean. So, our data will have high levels of variability.
- On the other hand, a low variance tells us that the values are quite close to the mean. In this case, the data will have low levels of variability.

Variance

- To calculate the variance in a dataset, we first need to find the difference between each individual value and the mean. The variance is the average of the squares of those differences. We can express the variance with the following math expression:

$$\sigma^2 = \frac{1}{n}\sum_{i=0}^{n-1}(x_i - \mu)^2$$

- In this equation, xi stands for individual values or observations in a dataset. μ stands for the mean or average of those values. n is the number of values in the dataset.
- The term xi - μ is called the deviation from the mean. So, the variance is the mean of square deviations. That's why we denoted it as σ^2.

Variance

Say we have a dataset [3, 5, 2, 7, 1, 3]. To find its variance, we need to calculate the mean which is:

$$(3+5+2+7+1+3)/6 = 3.5$$

Then, we need to calculate the sum of the square deviation from the mean of all the observations. Here's how:

$$(3-3.5)^2 + (5-3.5)^2 + (2-3.5)^2 + (7-3.5)^2 + (1-3.5)^2 + (3-3.5)^2 = 23.5$$

To find the variance, we just need to divide this result by the number of observations like this:

$$23.5/6 = 3.916666667$$

Variance

- That's all. The variance of our data is 3.916666667. The variance is difficult to understand and interpret, particularly how strange its units are.
- For example, if the observations in our dataset are measured in pounds, then the variance will be measured in square pounds.
- So, we can say that the observations are, on average, 3.916666667 square pounds far from the mean 3.5.
- Fortunately, the standard deviation comes to fix this problem

Standard Deviation

- The standard deviation measures the amount of variation or dispersion of a set of numeric values. Standard deviation is the square root of variance σ2 and is denoted as σ. So, if we want to calculate the standard deviation, then all we just have to do is to take the square root of the variance as follows:

$$\sigma=\sqrt{\sigma 2}$$

- Again, we need to distinguish between the population standard deviation, which is the square root of the population variance (σ2) and the sample standard deviation, which is the square root of the sample variance (S2). We'll denote the sample standard deviation as S:

$$S=\sqrt{S2}$$

Standard Deviation

- Low values of standard deviation tell us that individual values are closer to the mean. High values, on the other hand, tell us that individual observations are far away from the mean of the data.
- Values that are within one standard deviation of the mean can be thought of as fairly typical, whereas values that are three or more standard deviations away from the mean can be considered much more atypical. They're also known as outliers.

Standard Deviation

If we're trying to estimate the standard deviation of the population using a sample of data, then we'll be better served using **n - 1** degrees of freedom. Here's a math expression that we typically use to estimate the population variance:

$$\sigma_x = \sqrt{\frac{\sum_{i=0}^{n-1} (x_i - \mu_x)^2}{n-1}}$$

Note that this is the square root of the sample variance with **n - 1** degrees of freedom. This is equivalent to say:

$$S_{n-1} = \sqrt{S_{n-1}^2}$$

Correlation Coefficients

- Correlation coefficients quantify the association between variables or features of a dataset. These statistics are of high importance for science and technology, and Python has great tools that you can use to calculate them. SciPy, NumPy, and Pandas correlation methods are fast, comprehensive, and well-documented.
- We will learn:
  - What Pearson, Spearman, and Kendall correlation coefficients are
  - How to use SciPy, NumPy, and Pandas correlation functions
  - How to visualize data, regression lines, and correlation matrices with Matplotlib

What is correlation ?

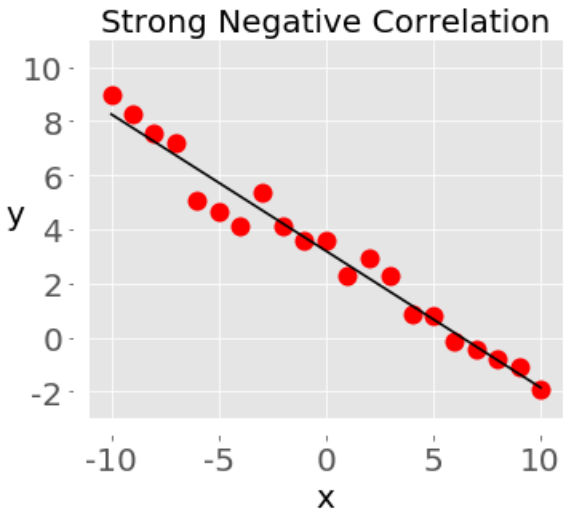
- Statistics and data science are often concerned about the relationships between two or more variables (or features) of a dataset. Each data point in the dataset is an observation, and the features are the properties or attributes of those observations.
- Every dataset you work with uses variables and observations. For example, you might be interested in understanding the following:
  - How the height of basketball players is correlated to their shooting accuracy
  - Whether there's a relationship between employee work experience and salary
  - What mathematical dependence exists between the population density and the gross domestic product of different countries

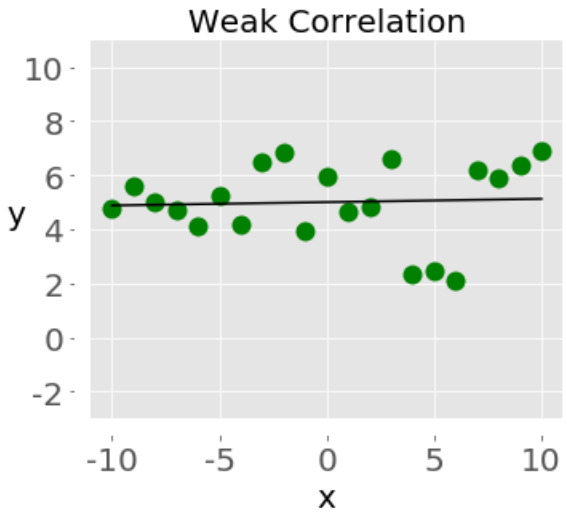
What is correlation ?

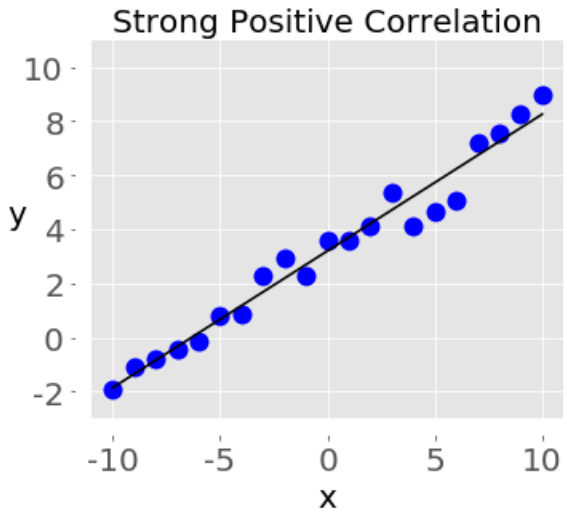
Name	Years of Experience	Annual Salary
Ann	30	120,000
Rob	21	105,000
Tom	19	90,000
Ivy	10	82,000

- In this table, each row represents one observation, or the data about one employee (either Ann, Rob, Tom, or Ivy). Each column shows one property or feature (name, experience, or salary) for all the employees.

Forms of correlation



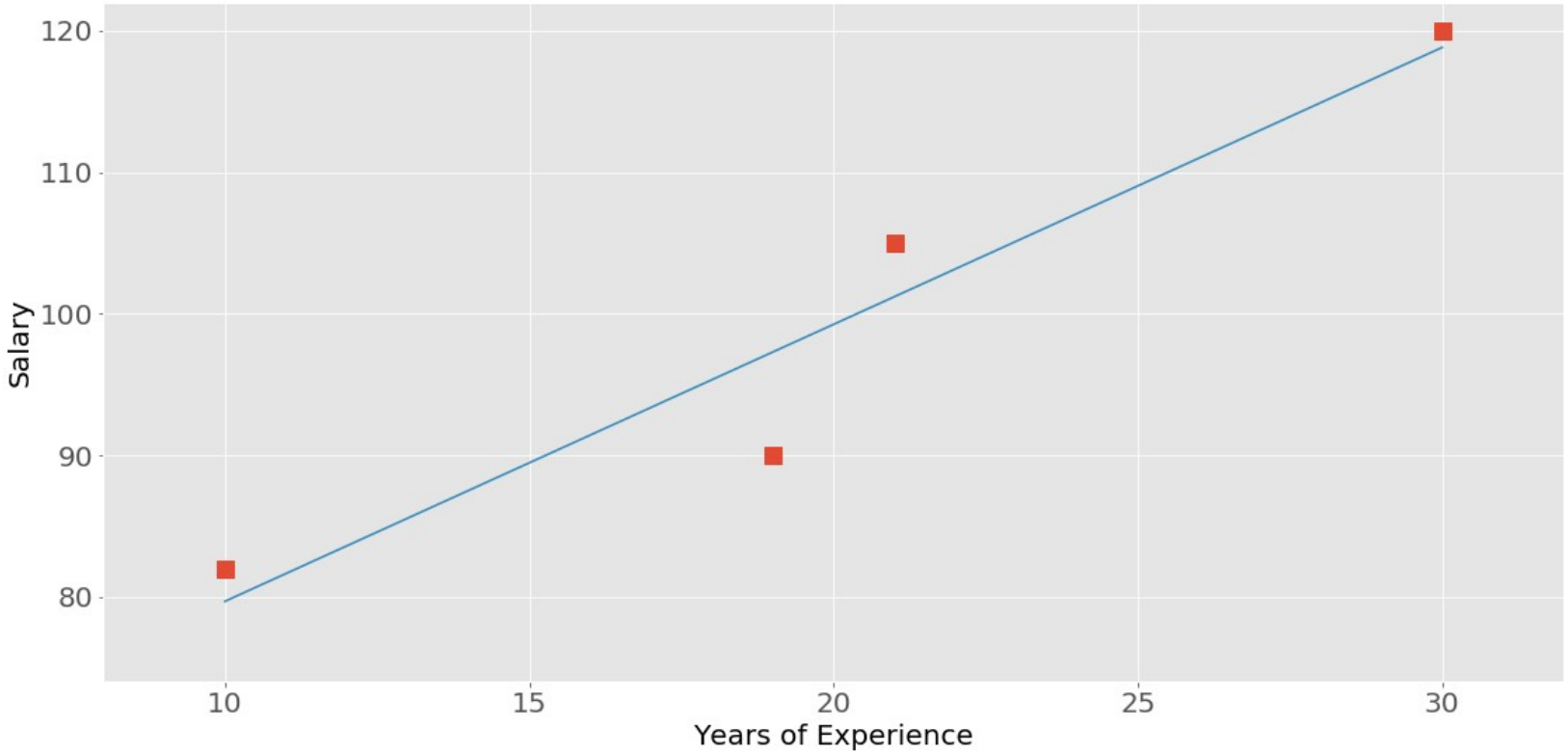




Forms of correlation

- Negative correlation (red dots): In the plot on the left, the y values tend to decrease as the x values increase. This shows strong negative correlation, which occurs when large values of one feature correspond to small values of the other, and vice versa.
- Weak or no correlation (green dots): The plot in the middle shows no obvious trend. This is a form of weak correlation, which occurs when an association between two features is not obvious or is hardly observable.
- Positive correlation (blue dots): In the plot on the right, the y values tend to increase as the x values increase. This illustrates strong positive correlation, which occurs when large values of one feature correspond to large values of the other, and vice versa.

Example: Employee table



Correlation Techniques

- There are several statistics that you can use to quantify correlation. We will be learning about three correlation coefficients:
  - Pearson's r
  - Spearman's rho
  - Kendall's tau
- Pearson's coefficient measures linear correlation, while the Spearman and Kendall coefficients compare the ranks of data.
- There are several NumPy, SciPy, and Pandas correlation functions and methods that you can use to calculate these coefficients.
- You can also use Matplotlib to conveniently illustrate the results.

Basic with numpy

```
import numpy as np
import scipy.stats
x = np.arange(10, 20)
```

```
y = np.array([2, 1, 4, 5, 8, 12, 18, 25, 96, 48])
```

```
r = np.corrcoef(x, y)
```

```
r
```

```
array([[1.        , 0.75864029],
       [0.75864029, 1.        ]])
```

```
r[0,1]
```

```
0.7586402890911867
```

```
r[1,0]
```

```
0.7586402890911869
```

	x	y
x	1.00	0.76
y	0.76	1.00

What sort of correlation ?

- The values on the main diagonal of the correlation matrix (upper left and lower right) are equal to 1.
- The upper left value corresponds to the correlation coefficient for x and x, while the lower right value is the correlation coefficient for y and y. They are always equal to 1.
- However, what you usually need are the lower left and upper right values of the correlation matrix.
- These values are equal and both represent the Pearson correlation coefficient for x and y. In this case, it’s approximately 0.76.

Correlation with scipy

```
import numpy as np
import scipy.stats
```

```
x = np.arange(10, 20)
y = np.array([2, 1, 4, 5, 8, 12, 18, 25, 96, 48])
```

```
scipy.stats.pearsonr(x, y)    # Pearson's r
```

```
(0.7586402890911869, 0.010964341301680829)
```

```
scipy.stats.spearmanr(x, y)   # Spearman's rho
```

```
SpearmanrResult(correlation=0.9757575757575757, pvalue=1.4675461874042197e-06)
```

```
scipy.stats.kendalltau(x, y)  # Kendall's tau
```

```
KendalltauResult(correlation=0.9111111111111111, pvalue=2.9761904761904762e-05)
```

```
scipy.stats.pearsonr(x, y)[0]    # Pearson's r, p-value
```

```
0.7586402890911869
```

Correlation with scipy

```
scipy.stats.pearsonr(x, y)[0]    # Pearson's r, p-value
```

```
0.7586402890911869
```

```
scipy.stats.spearmanr(x, y).correlation
```

```
0.9757575757575757
```

```
scipy.stats.kendalltau(x, y).correlation
```

```
0.9111111111111111
```

```
r, p = scipy.stats.pearsonr(x, y)
```

```
r
```

```
0.7586402890911869
```

Correlation with Pandas

```
import pandas as pd
import numpy as np
x = pd.Series(range(10, 20))
```

```
y = pd.Series([2, 1, 4, 5, 8, 12, 18, 25, 96, 48])
```

```
y.corr(x)
```

```
0.7586402890911869
```

```
x.corr(y, method='spearman')  # Spearman's rho
```

```
0.9757575757575757
```

```
x.corr(y, method='kendall')   # Kendall's tau
```

```
0.9111111111111111
```

Linear Correlation

- Linear correlation measures the proximity of the mathematical relationship between variables or dataset features to a linear function.
- If the relationship between the two features is closer to some linear function, then their linear correlation is stronger and the absolute value of the correlation coefficient is higher.

Pearson Correlation

- Consider a dataset with two features: x and y. Each feature has n values, so x and y are n-tuples. Say that the first value x_1 from x corresponds to the first value y_1 from y, the second value x_2 from x to the second value y_2 from y, and so on. Then, there are n pairs of corresponding values: (x_1, y_1), (x_2, y_2), and so on. Each of these x-y pairs represents a single observation.
- The Pearson (product-moment) correlation coefficient is a measure of the linear relationship between two features. It's the ratio of the covariance of x and y to the product of their standard deviations. It's often denoted with the letter r and called Pearson's r. You can express this value mathematically with this equation:
- $r = \Sigma_i((x_i - \text{mean}(x))(y_i - \text{mean}(y))) \, (\sqrt{\Sigma_i}(x_i - \text{mean}(x))^2 \sqrt{\Sigma_i}(y_i - \text{mean}(y))^2)^{-1}$

Pearson Correlation

- The Pearson correlation coefficient can take on any real value in the range $-1 \leq r \leq 1$.
- The maximum value $r = 1$ corresponds to the case when there's a perfect positive linear relationship between x and y. In other words, larger x values correspond to larger y values and vice versa.
- The value $r > 0$ indicates positive correlation between x and y.
- The value $r = 0$ corresponds to the case when x and y are independent.
- The value $r < 0$ indicates negative correlation between x and y.
- The minimal value $r = -1$ corresponds to the case when there's a perfect negative linear relationship between x and y. In other words, larger x values correspond to smaller y values and vice versa.

Pearson Correlation

Pearson's r Value	Correlation Between x and y
equal to 1	perfect positive linear relationship
greater than 0	positive correlation
equal to 0	independent
less than 0	negative correlation
equal to -1	perfect negative linear relationship

Linear Regression

- Linear regression is the process of finding the linear function that is as close as possible to the actual relationship between features.
- In other words, you determine the linear function that best describes the association between the features. This linear function is also called the regression line.
- You can implement linear regression with SciPy. You'll get the linear function that best approximates the relationship between two arrays, as well as the Pearson correlation coefficient.

Practical Linear Regression

```
import numpy as np
import scipy.stats
x = np.arange(10, 20)
y = np.array([2, 1, 4, 5, 8, 12, 18, 25, 96, 48])
```

```
result = scipy.stats.linregress(x, y)
```

```
result.slope
```

```
7.4363636363636365
```

```
result.intercept
```

```
-85.92727272727274
```

```
result.rvalue
```

```
0.7586402890911869
```

```
result.pvalue
```

```
0.010964341301680825
```

More on Regression

```
xy = np.array([[10, 11, 12, 13, 14, 15, 16, 17, 18, 19],
               [2, 1, 4, 5, 8, 12, 18, 25, 96, 48]])
```

```
scipy.stats.linregress(xy)
```

```
LinregressResult(slope=7.4363636363636365, intercept=-85.92727272727
274, rvalue=0.7586402890911869, pvalue=0.010964341301680825, stderr=
2.257878767543913)
```

```
xy.T;
```

```
scipy.stats.linregress(xy.T)
```

```
LinregressResult(slope=7.4363636363636365, intercept=-85.92727272727
274, rvalue=0.7586402890911869, pvalue=0.010964341301680825, stderr=
2.257878767543913)
```

```
scipy.stats.linregress(np.arange(3), np.array([2, np.nan, 5]))
```

```
LinregressResult(slope=nan, intercept=nan, rvalue=nan, pvalue=nan, s
tderr=nan)
```

Using Multidimensional Data

```
xy = np.array([[10, 11, 12, 13, 14, 15, 16, 17, 18, 19],
               [2, 1, 4, 5, 8, 12, 18, 25, 96, 48]])
```

```
np.corrcoef(xy)
```

```
array([[1.        , 0.75864029],
       [0.75864029, 1.        ]])
```

```
xyz = np.array([[10, 11, 12, 13, 14, 15, 16, 17, 18, 19],
                [2, 1, 4, 5, 8, 12, 18, 25, 96, 48],
                [5, 3, 2, 1, 0, -2, -8, -11, -15, -16]])
```

```
np.corrcoef(xyz)
```

```
array([[ 1.        ,  0.75864029, -0.96807242],
       [ 0.75864029,  1.        , -0.83407922],
       [-0.96807242, -0.83407922,  1.        ]])
```

Using Pandas

```
x = pd.Series(range(10, 20))
```

```
y = pd.Series([2, 1, 4, 5, 8, 12, 18, 25, 96, 48])
```

```
z = pd.Series([5, 3, 2, 1, 0, -2, -8, -11, -15, -16])
```

```
xy = pd.DataFrame({'x-values': x, 'y-values': y})
```

```
xy.corr()
```

	x-values	y-values
x-values	1.00000	0.75864
y-values	0.75864	1.00000

Using Pandas

```
xyz = pd.DataFrame({'x-values': x, 'y-values': y, 'z-values': z})
```

```
xyz.corr()
```

	x-values	y-values	z-values
x-values	1.000000	0.758640	-0.968072
y-values	0.758640	1.000000	-0.834079
z-values	-0.968072	-0.834079	1.000000

Using Pandas

```
corr_matrix = xy.corr()
```

```
corr_matrix
```

	x-values	y-values
x-values	1.00000	0.75864
y-values	0.75864	1.00000

```
corr_matrix.at['x-values', 'y-values']
```

```
0.7586402890911869
```

```
corr_matrix.iat[0, 1]
```

```
0.7586402890911869
```

Using corrwith

```
xy.corrwith(z)
```

```
x-values   -0.968072
y-values   -0.834079
dtype: float64
```

```
x.corr(y, method='spearman')
```

```
0.9757575757575757
```

```
xyz.corr(method='spearman')
```

	x-values	y-values	z-values
x-values	1.000000	0.975758	-1.000000
y-values	0.975758	1.000000	-0.975758
z-values	-1.000000	-0.975758	1.000000

```
xy.corrwith(z, method='spearman')
```

```
x-values   -1.000000
y-values   -0.975758
```

```
xy.corrwith(z, method='spearman')
```

```
x-values   -1.000000
y-values   -0.975758
dtype: float64
```

```
x.corr(y, method='kendall')
```

```
0.911111111111111
```

```
xy.corr(method='kendall')
```

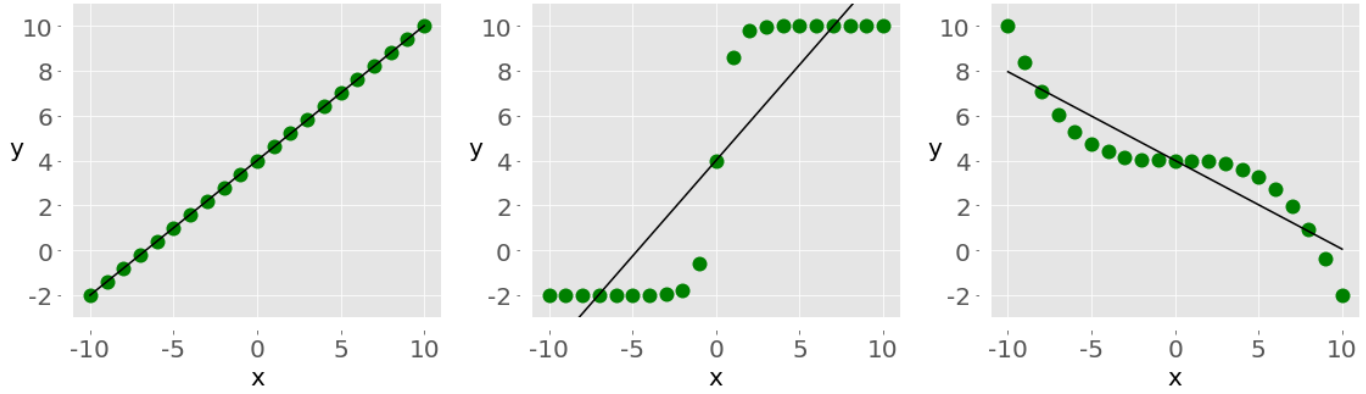
	x-values	y-values
x-values	1.000000	0.911111
y-values	0.911111	1.000000

```
xyz.corr(method='kendall')
```

Rank Correlation

- Rank correlation compares the ranks or the orderings of the data related to two variables or dataset features.
- If the orderings are similar, then the correlation is strong, positive, and high. However, if the orderings are close to reversed, then the correlation is strong, negative, and low.
- In other words, rank correlation is concerned only with the order of values, not with the particular values from the dataset.

Rank Correlation



- The left plot has a perfect positive linear relationship between x and y, so r = 1. The central plot shows positive correlation and the right one shows negative correlation. However, neither of them is a linear function, so r is different than –1 or 1.
- When you look only at the orderings or ranks, all three relationships are perfect! The left and central plots show the observations where larger x values always correspond to larger y values. This is perfect positive rank correlation. The right plot illustrates the opposite case, which is perfect negative rank correlation.

The Spearman Correlation

- The Spearman correlation coefficient between two features is the Pearson correlation coefficient between their rank values.
- It's calculated the same way as the Pearson correlation coefficient but takes into account their ranks instead of their values. It's often denoted with the Greek letter rho (ρ) and called Spearman's rho.
- Say you have two n-tuples, x and y, where (x_1, y_1), (x_2, y_2), … are the observations as pairs of corresponding values.
- You can calculate the Spearman correlation coefficient ρ the same way as the Pearson coefficient. You'll use the ranks instead of the actual values from x and y.

The Spearman Correlation

- It can take a real value in the range $-1 \leq \rho \leq 1$.
- Its maximum value $\rho = 1$ corresponds to the case when there's a monotonically increasing function between x and y. In other words, larger x values correspond to larger y values and vice versa.
- Its minimum value $\rho = -1$ corresponds to the case when there's a monotonically decreasing function between x and y. In other words, larger x values correspond to smaller y values and vice versa.

Kendall Correlation Coefficient

- Let's start again by considering two n-tuples, x and y. Each of the x-y pairs (x_1, y_1), (x_2, y_2), … is a single observation. A pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, will be one of three things:
  - concordant if either ($x_i > x_j$ and $y_i > y_j$) or ($x_i < x_j$ and $y_i < y_j$)
  - discordant if either ($x_i < x_j$ and $y_i > y_j$) or ($x_i > x_j$ and $y_i < y_j$)
  - neither if there's a tie in x ($x_i = x_j$) or a tie in y ($y_i = y_j$)
- The Kendall correlation coefficient compares the number of concordant and discordant pairs of data.
- This coefficient is based on the difference in the counts of concordant and discordant pairs relative to the number of x-y pairs. It's often denoted with the Greek letter tau (τ) and called Kendall's tau.

Kendall Correlation Coefficient

- It can take a real value in the range $-1 \leq \tau \leq 1$.
- Its maximum value $\tau = 1$ corresponds to the case when the ranks of the corresponding values in x and y are the same. In other words, all pairs are concordant.
- Its minimum value $\tau = -1$ corresponds to the case when the rankings in x are the reverse of the rankings in y. In other words, all pairs are discordant.

Ranking with scipy

```
import numpy as np
import scipy.stats
x = np.arange(10, 20)
y = np.array([2, 1, 4, 5, 8, 12, 18, 25, 96, 48])
z = np.array([5, 3, 2, 1, 0, -2, -8, -11, -15, -16])
```

```
scipy.stats.rankdata(x)
```

```
array([ 1.,  2.,  3.,  4.,  5.,  6.,  7.,  8.,  9., 10.])
```

```
scipy.stats.rankdata(y)
```

```
array([ 2.,  1.,  3.,  4.,  5.,  6.,  7.,  8., 10.,  9.])
```

```
scipy.stats.rankdata(z)
```

```
array([10.,  9.,  8.,  7.,  6.,  5.,  4.,  3.,  2.,  1.])
```

```
scipy.stats.rankdata([8, 2, 0, 2])
```

```
array([4. , 2.5, 1. , 2.5])
```

Implementation

```
result = scipy.stats.kendalltau(x, y)
```

```
result
```

```
KendalltauResult(correlation=0.9111111111111111, pvalue=2.9761904761904762e-05)
```

```
result.correlation
```

```
0.9111111111111111
```

```
tau, p = scipy.stats.kendalltau(x, y)
```

```
tau
```

```
0.9111111111111111
```

Scipy with 2D

```
xyz = np.array([[10, 11, 12, 13, 14, 15, 16, 17, 18, 19],
                [2, 1, 4, 5, 8, 12, 18, 25, 96, 48],
                [5, 3, 2, 1, 0, -2, -8, -11, -15, -16]])
```

```
corr_matrix = np.corrcoef(xyz).round(decimals=2)
```

```
corr_matrix
```

```
array([[ 1.  ,  0.76, -0.97],
       [ 0.76,  1.  , -0.83],
       [-0.97, -0.83,  1.  ]])
```

Scipy with 2D

```
corr_matrix, p_matrix = scipy.stats.spearmanr(xyz, axis=1)
```

```
corr_matrix
```

```
array([[ 1.        ,  0.97575758, -1.        ],
       [ 0.97575758,  1.        , -0.97575758],
       [-1.        , -0.97575758,  1.        ]])
```

```
p_matrix
```

```
array([[6.64689742e-64, 1.46754619e-06, 6.64689742e-64],
       [1.46754619e-06, 6.64689742e-64, 1.46754619e-06],
       [6.64689742e-64, 1.46754619e-06, 6.64689742e-64]])
```

Using pandas

```
xy.corr(method='kendall')
```

	x-values	y-values
x-values	1.000000	0.911111
y-values	0.911111	1.000000

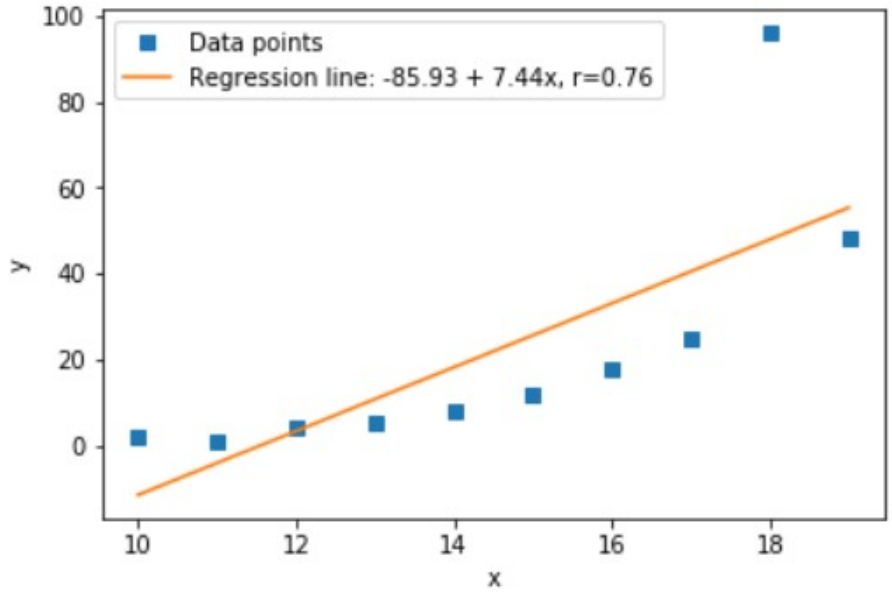
```
xyz.corr(method='kendall')
```

	x-values	y-values	z-values
x-values	1.000000	0.911111	-1.000000
y-values	0.911111	1.000000	-0.911111
z-values	-1.000000	-0.911111	1.000000

Visualization of Regression Line

```
import matplotlib.pyplot as plt
fig, ax = plt.subplots()
ax.plot(x, y, linewidth=0, marker='s', label='Data points')
ax.plot(x, intercept + slope * x, label=line)
ax.set_xlabel('x')
ax.set_ylabel('y')
ax.legend(facecolor='white')
```

```
<matplotlib.legend.Legend at 0x7f03ce882668>
```



Simpson’s Paradox

- Simpson’s Paradox refers to a situation where you believe you understand the direction of a relationship between two variables, but when you consider an additional variable, that direction appears to reverse.

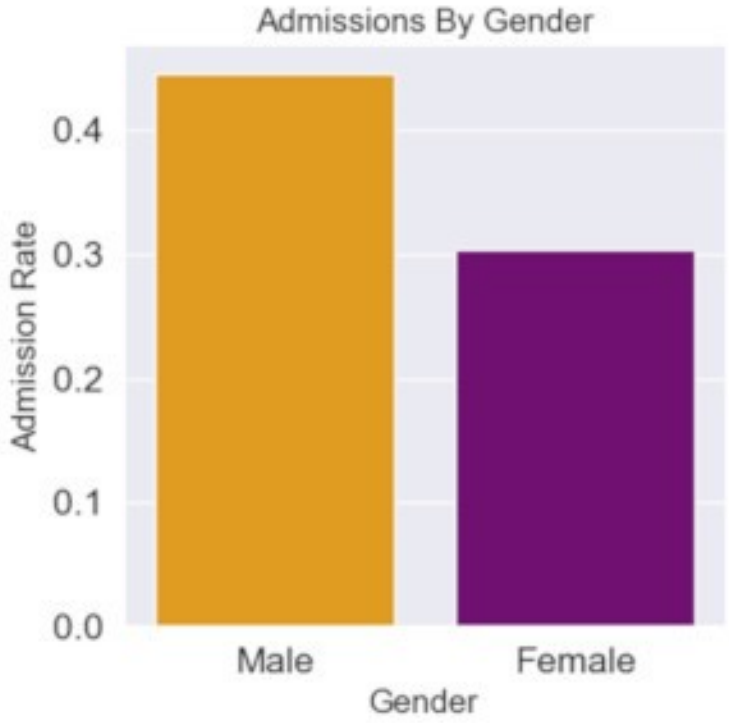
Simpson's Paradox : Why?

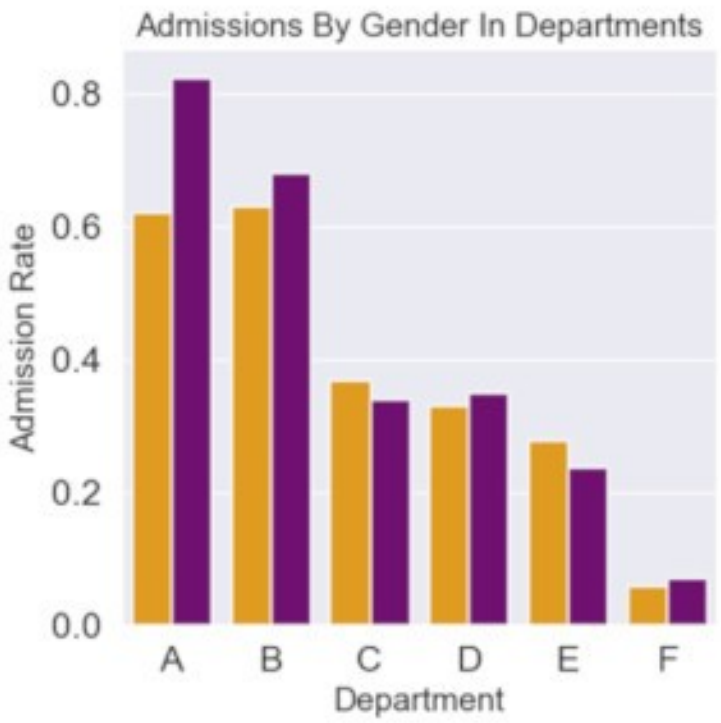
- Simpson's Paradox happens because disaggregation of the data (e.g., splitting it into subgroups) can cause certain subgroups to have an imbalanced representation compared to other subgroups.
- This might be due to the relationship between the variables, or simply due to the way that the data has been partitioned into subgroups.

Example #1: Admissions

- A famous example of Simpson's Paradox appears in the admissions data for graduate school at UC Berkeley in 1973.
- In this example, when looking at the graduate admissions data overall, it appeared that men were more likely to be admitted than women (gender discrimination!), but when looking at the data for each department individually, men were less likely to be admitted than women in most of the departments.

Example #1: Admissions





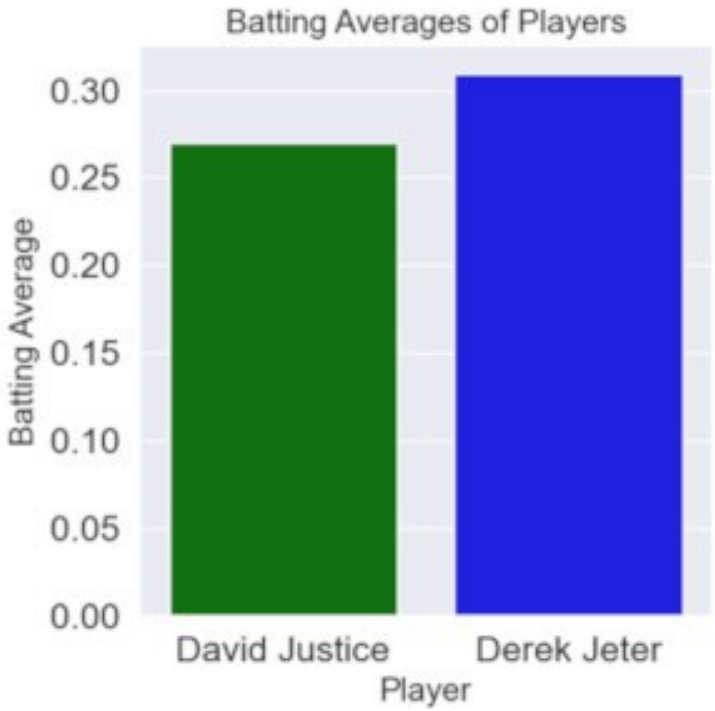
Why?

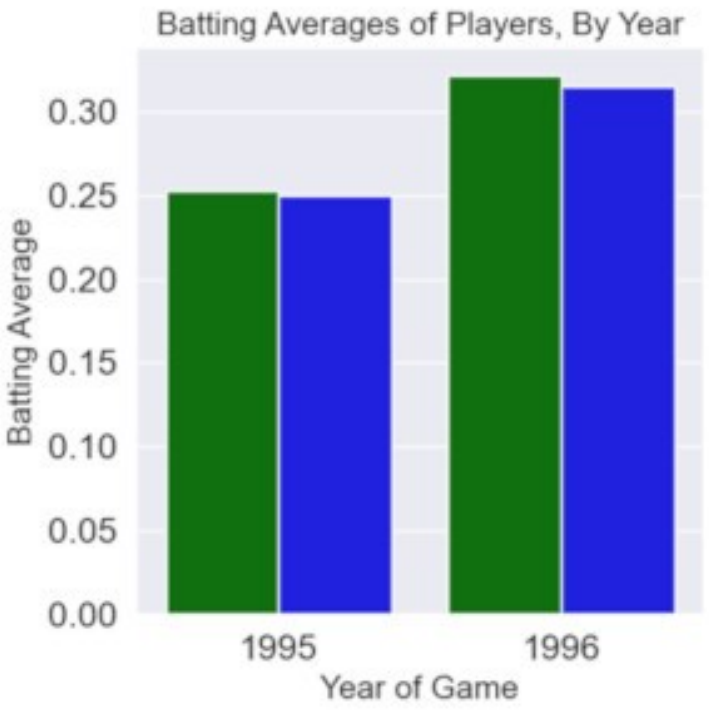
- Here is an explanation of why this happens:
  - Different departments had very different acceptance rates (some were much “harder” to get into than others)
  - More females applied to the “harder" departments
  - Therefore, females had a lower acceptance rate in aggregate
- This leads us to ask: which view is the correct view? Do men or women have a higher acceptance rate? Is there a gender bias in admissions at this university?
- In this case, it seems most reasonable to conclude that looking at the admissions rates by department makes more sense, and the disaggregated view is correct.

Example #2 : Baseball

- Another example of Simpson's Paradox can be found in the batting averages of two famous baseball players, Derek Jeter and David Justice, from 1995 and 1996.
- David Justice had a higher batting average in both 1995 and 1996 individually, but Derek Jeter had a higher batting average over the two years combined.

Example #2


Batting Averages of Players
Batting Average
0.30
0.25
0.20
0.15
0.10
0.05
0.00
David Justice
Derek Jeter
Player


Batting Averages of Players, By Year
Batting Average
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1995
1996
Year of Game

Why?

- Here is an explanation of why this happens:
  - Both players had significantly higher batting averages in 1996 than in 1995
  - Derek Jeter had significantly more at-bats in 1996; David Justice had significantly more in 1995
  - Therefore, Derek Jeter had a higher batting average in aggregate

What to do?

- Without enough domain knowledge, it's hard to know which view of the relationship between two variables makes more sense – the one with or without the third variable.
- But before we think about how to deal with Simpson's Paradox, we need to find a way to efficiently detect it in a dataset.
- As mentioned earlier, it's possible to find an instance of Simpson's Paradox (a "Simpson's Pair") simply by disaggregating a contingency table or a plot of data points and studying the results.
- However, there are other ways we can find Simpson's Pairs using models

What to do?

  - By building decision trees and comparing the distributions, or
  - By building regression models and comparing the signs of the coefficients
- There are benefits to both, however, this can get difficult very quickly, especially when working with big datasets.
- It's hard to know which variables in the dataset may reverse the relationship between two other variables, and it can be hard to check all possible pairs of variables manually.
- Imagine we have a dataset with only 20 variables: we'd need to check almost 400 pairs to be sure to find all cases of Simpson's Paradox.

Some other correlation caveats

- A correlation of zero indicates that there is no linear relationship between the two variables. However, there may be other sorts of relationships. For example, if:

  x = [-2, -1, 0, 1, 2]

  y = [2, 1, 0, 1, 2]
- then x and y have zero correlation. But they certainly have a relationship — each element of y equals the absolute value of the corresponding element of x .
- What they don't have is a relationship in which knowing how x_i compares to mean(x) gives us information about how y_i compares to mean(y) . That is the sort of relationship that correlation looks for.

Some other correlation caveats

- In addition, correlation tells you nothing about how large the relationship is. The variables:

  x = [-2, 1, 0, 1, 2]

  y = [99.98, 99.99, 100, 100.01, 100.02]

- are perfectly correlated, but (depending on what you're measuring) it's quite possible that this relationship isn't all that interesting.

Causation

- Causation means that one event causes another event to occur.
- Causation can only be determined from an appropriately designed experiment.
- In such experiments, similar groups receive different treatments, and the outcomes of each group are studied.
- We can only conclude that a treatment causes an effect if the groups have noticeably different outcomes.

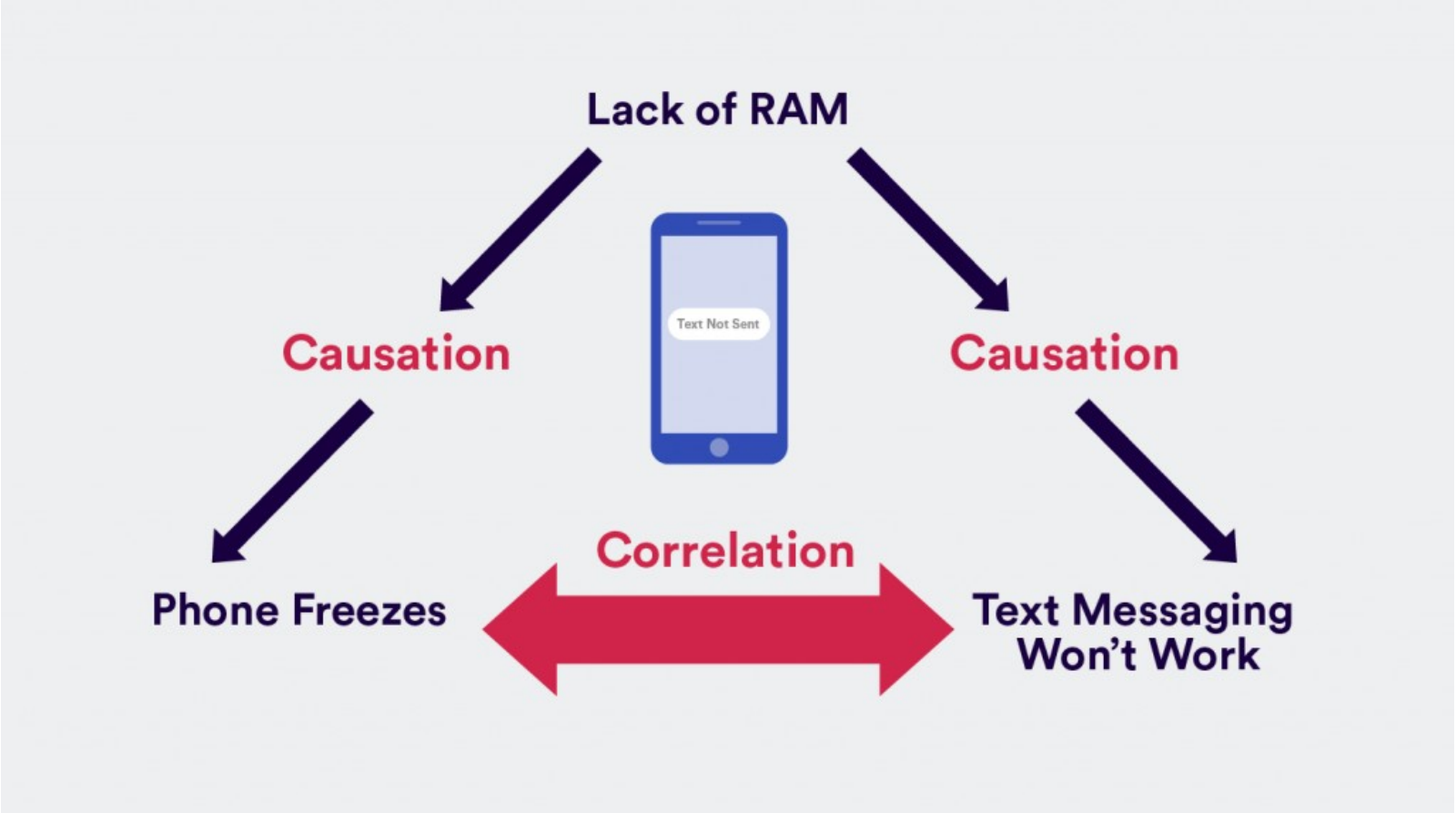
Why?

- Describing a relationship between variables
- Identifying statements consistent with the relationship between variables
- Identifying valid conclusions about correlation and causation for data shown in a scatterplot
- Identifying a factor that could explain why a correlation does not imply a causal relationship

Example:

- My mother-in-law recently complained to me: “Whenever I try to text message, my phone freezes.”
- A quick look at her smartphone confirmed my suspicion: she had five game apps open at the same time plus Facebook and YouTube.
- The act of trying to send a text message wasn’t causing the freeze, the lack of RAM was.
- But she immediately connected it with the last action she was doing before the freeze.

Example:


Lack of RAM
Causation
Causation
Text Not Sent
Correlation
Phone Freezes
Text Messaging
Won't Work

Causation

- Causation is implying that A and B have a cause-and-effect relationship with one another. You're saying A causes B.
- Causation is also known as causality.
  - Firstly, causation means that two events appear at the same time or one after the other.
  - And secondly, it means these two variables not only appear together, the existence of one causes the other to manifest.

Why does correlation means causation?

- Even if there is a correlation between two variables, we cannot conclude that one variable causes a change in the other. This relationship could be coincidental, or a third factor may be causing both variables to change.
- For example, Liam collected data on the sales of ice cream cones and air conditioners in his hometown. He found that when ice cream sales were low, air conditioner sales tended to be low and that when ice cream sales were high, air conditioner sales tended to be high.
  - Liam can conclude that sales of ice cream cones and air conditioner are positively correlated.
  - Liam can't conclude that selling more ice cream cones causes more air conditioners to be sold. It is likely that the increases in the sales of both ice cream cones and air conditioners are caused by a third factor, an increase in temperature!

Thank you

@mitu_skillologies

/mITuSkillologies

@mitu_group

/company/mitu-skillologies

MITUSkillologies

Web Resources
https://mitu.co.in
http://tusharkute.com

contact@mitu.co.in

tushar@tusharkute.com