

# Paradoxes

by Sophia Tutorial



## WHAT'S COVERED

This tutorial will cover the topic of statistical paradoxes. Our discussion breaks down as follows:

1. Paradoxes
2. Benford's Law
3. Exponential Growth
4. Simpson's Paradox

## 1. Paradoxes

**Paradoxes** are apparent contradictions in what you see versus what you expected to see. Specifically, the one that we're going to talk about in this tutorial is called Benford's Law, covered in the next section.

Statistics allows us to draw conclusions about things that we see. Sometimes, though, the phenomena that we see are counter to what we thought would happen. These seeming contradictions are called paradoxes. If we understand them better, we can improve as statistical thinkers.

🔗 **EXAMPLE** Suppose that you were going to create a phony checking account and you wanted to set it up so that you could steal some money from people. To do so, you would need to create a checking account number for this phony account. For your fake account number, would it matter what number you selected as the first number?

You probably think that, no, it really wouldn't matter which number you choose as the first in your fake account number. All the numbers one through nine are equally likely to be selected for the first number, so if the account number is randomly selected, it wouldn't really matter which number you selected first. That's your intuition.

However, your intuition that all the numbers 1-9 are equally likely to be selected for the first number is actually wrong. What's really the case is that checking account numbers are most likely to start with 1. In published data, checking account numbers and many other kinds of numbers are most likely to start with 1.



## TERM TO KNOW

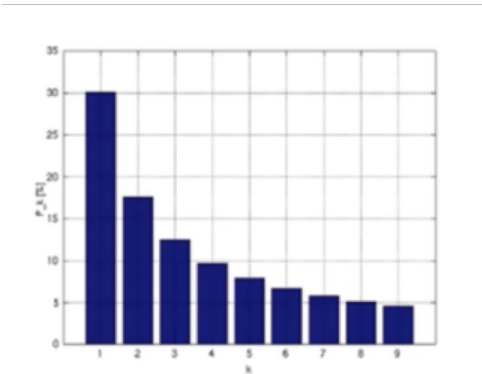
### Paradox

## 2. Benford's Law

**Benford's Law**, also called the First Digit Law, says that the first number of most any real-life data, including financial reports, follow a pattern with the number 1 being the most likely, 2 being the next most likely, and so on in a specific order.

This law shows that only about 10% of account numbers will start with a four, whereas about 30% will start with a one.

Digit	Likelihood
1	30.1%
2	17.6%
3	12.5%
4	9.7%
5	7.9%
6	6.7%
7	5.8%
8	5.1%
9	4.6%



People who try to steal identities are likely to use more 4's, 5's, and 6's, because they think those are the middle. In reality, it's the number 1 that's the most likely as a lead number.

Benford examined many different sets of data, including:

- The area of rivers
- Populations of different countries
- Constants that are used in physics
- Numbers that happen to appear in the newspaper
- The specific heat values of different things

Benford looked at these different values and saw that almost across the board, 1 is the most likely lead number, 2 is the next most likely lead number, and 9 is the least likely lead number.

Data	Samples	1	2	3	4	5	6	7	8	9
Rivers, Area	335	31.0%	16.4%	10.7%	11.3%	7.2%	8.6%	5.5%	4.2%	5.1%
Population	3259	33.9%	20.4%	14.2%	8.1%	7.2%	6.2%	4.1%	3.7%	2.2%
Constants	104	41.3%	14.4%	4.8%	8.6%	10.6%	5.8%	1.0%	2.9%	10.6%
Newspapers	100	30.0%	18.0%	12.0%	10.0%	8.0%	6.0%	6.0%	5.0%	5.0%

Specific Heat	1389	24.0%	18.4%	16.2%	14.6%	10.6%	4.1%	3.2%	4.8%	4.1%
Pressure	703	29.6%	18.3%	12.8%	9.8%	8.3%	6.4%	5.7%	4.4%	4.7%
Mol. Wgt.	1800	26.7%	25.2%	15.4%	10.8%	6.7%	5.1%	4.1%	2.8%	3.2%



#### THINK ABOUT IT

Ask yourself, if the lead number follows that law, then does that mean that the second digit must also follow that law? If you thought yes, then your intuition again led you astray. While the first digit follows Benford's Law, the second digit does not.

#### Benford's Law: Expected Occurrence of First and Second Digits

Digit	First Digit	Second Digit
0	---	12.0%
1	30.1%	11.4%
2	17.6%	10.8%
3	12.5%	10.4%
4	9.7%	10.0%
5	7.9%	9.7%
6	6.7%	9.3%
7	5.8%	9.0%
8	5.2%	8.8%
9	4.6%	8.5%

As you can see in the image above, the second digit is approximately equally likely to be any of the numbers 0-9. There is a slight favoring of the lower numbers, but all are about 10%. The second digit has an equal frequency.



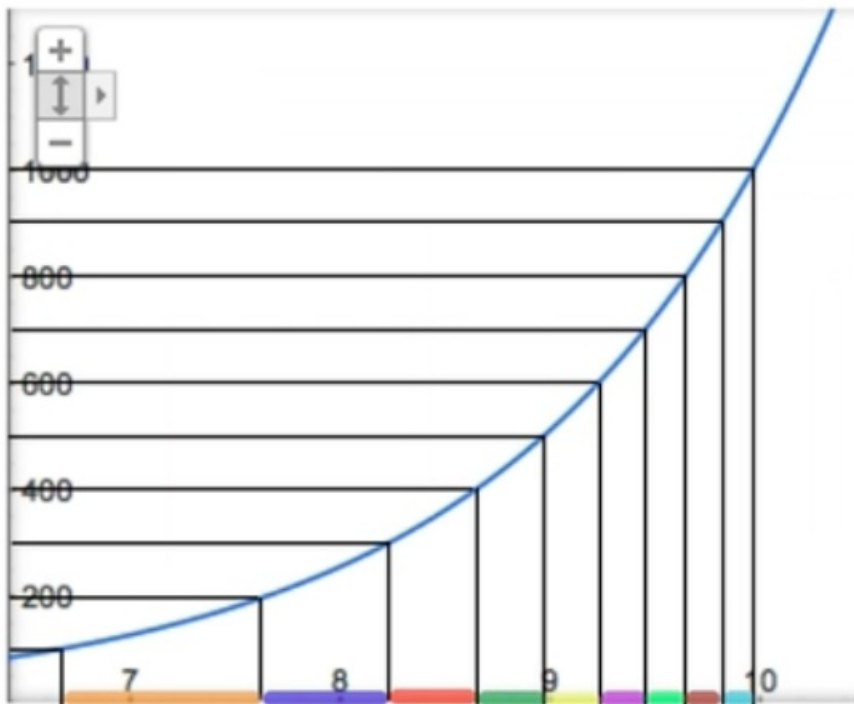
#### TERM TO KNOW

#### Benford's Law

A law that shows that most of the numbers that are published, regardless of topic, begin with smaller numbers, and very few of them lead with larger numbers. The most common first digit is 1; the least common is 9.

## 3. Exponential Growth

The reason for a phenomenon like the one you saw in the above examples has to do with exponential growth, which looks like this:



As you can see, the 100-200, 200-300, and 300-400 ranges are equally spaced. However, there are more numbers on the x-axis that create a value 100-200 versus ones that create a number 200-300. That amount diminishes as you move along to the right of the x-axis.

## 4. Simpson's Paradox

There are many kinds of paradoxes, and **Simpson's Paradox** is just one of them. Simpson's Paradox is a relationship that's present in groups but reversed when the groups are combined.



### TERM TO KNOW

#### Simpson's Paradox

When two sets of data are subdivided, the means for the first data set can be consistently higher than the second, but when looked at as a whole, the mean of the second set is higher than the first.

A very famous example of Simpson's Paradox took place in 1973. That year, UC Berkeley had a sex discrimination lawsuit filed against them that asserted that UCB was favoring men over women substantially in the admissions process for their grad schools. Here is the data:

Men		Women	
Applied	Accepted	Applied	Accepted
977	492	400	148
50.3%		37%	

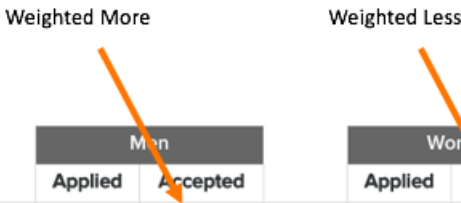
As you can see, it looks like 977 men applied and 492 were accepted, which is a little over half. In contrast, of the 400 women who applied, well under half, only 148, were accepted. In fact, the proportions are 50.3% versus 37%.

The difference between 37% and 50.3% is huge, which is why the lawsuit was filed. To see exactly where the women were being discriminated against, the lawyers looked into the admissions by department. You would expect that there would be a large discrepancy in certain departments. For this tutorial, we will look at the data for two departments, which we are calling the Engineering and English (though the true numbers within these departments may have been different in the real case).

	Men			Women		
	Applied	Accepted		Applied	Accepted	
Engineering	560	353	63.0%	25	17	68.0%
English	417	139	33.3%	375	131	34.9%
	977	492	50.4%	400	148	37.0%

For the Engineering department, you can see that about 63% percent of men were accepted to the Engineering department versus 68% for women. Women were accepted at higher rates to the Engineering department. Therefore, the discrepancy was not present in the Engineering department. You might then assume that the discrepancy occurs in the English department. However, women were accepted at higher rates to the English department as well--34.9% versus 33.3%.

Women were accepted at higher rates to the Engineering and English departments, but much lower overall. Examining how the men's application rates were distributed, their 63% was weighted for a lot more into the weighted average of admissions rates versus the 68% for the women.



	Men			Women		
	Applied	Accepted		Applied	Accepted	
Engineering	560	353	63.0%	25	17	68.0%
English	417	139	33.3%	375	131	34.9%
	977	492	50.4%	400	148	37.0%

Only 25 of the 400 applicants to the Engineering department were women. That's not very many. Therefore, that 68%, even though it's a high percentage, doesn't count nearly as much in the weighted average as the 34.9% does. So, the 63% is weighted heavily for the men versus the 68%, which is weighted hardly at all for the women. This is why you see that reversal of association.



## SUMMARY

A paradox is a seeming contradiction between what you think should happen versus what's actually happening. The First Digit Law, which is Benford's Law, is one of these paradoxes. We thought that we would find a uniform distribution among the first digits of certain numbers that we see, but 1 is a much more common lead number than any other. Not all numbers occur with the equal frequency as the lead digit, which is related to exponential growth. Simpson's Paradox is an association that the

data show when you group that data in specific ways, causing the association to be reversed when the groups are combined. There are several paradoxes like this, of which Simpson's paradox is just one.

Good luck!

Source: Adapted from Sophia tutorial by Jonathan Osters.



## TERMS TO KNOW

### **Benford's Law**

A law that shows that most of the numbers that are published, regardless of topic, begin with smaller numbers, and very few of them lead with larger numbers. The most common first digit is 1, the least common is 9.

### **Paradox**

An apparent contradiction between what our intuition tells us, and what is true in reality.

### **Simpson's Paradox**

When two sets of data are subdivided, the means for the first data set can be consistently higher than the second, but when looked at as a whole, the mean of the second set is higher than the first.