Statistics

Dr. Carmen Bruni

Centre for Education in Mathematics and Computing
University of Waterloo
http://cemc.uwaterloo.ca

October 12th, 2016

Quote

Quote

"There are three types of lies:

Quote

"There are three types of lies:

- Lies

Quote

"There are three types of lies:

- Lies
- Damned lies

Quote

"There are three types of lies:

- Lies
- Damned lies
- ... and statistics"

(Unclear origin - maybe Lord Courtney [1895] or Mark Twain [1904]).

Quote

"There are three types of lies:

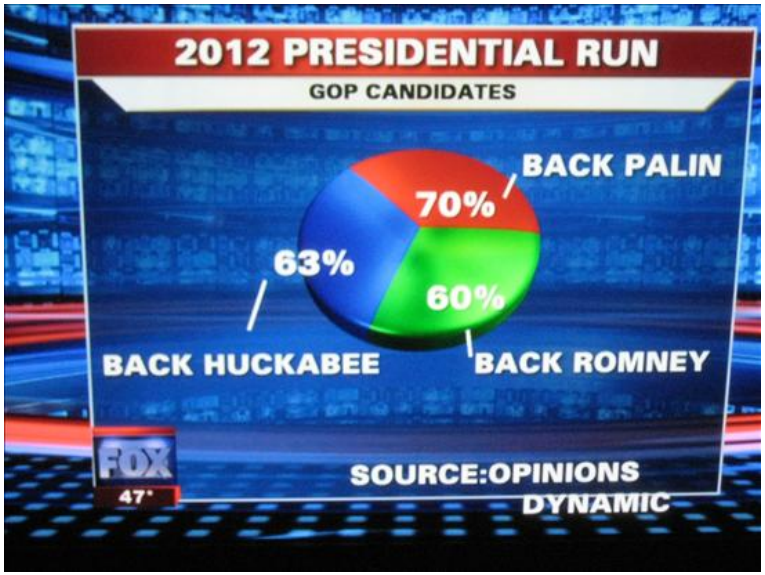
- Lies
- Damned lies
- ... and statistics"

(Unclear origin - maybe Lord Courtney [1895] or Mark Twain [1904]).

How do we read statistics? Can we trust them?

Bad Statistics (Courtesy of Fox News)

Take a look at each of the following graphs and tell me what you think. (Source: Jeff Leek `http://tinyurl.com/ckatyj2`)

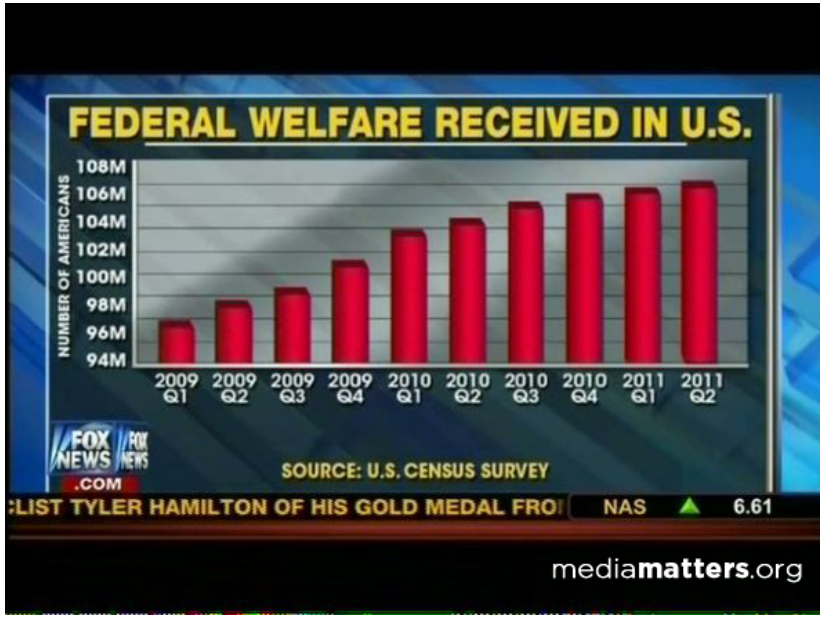


Bad Statistics (Courtesy of Fox News)

Take a look at each of the following graphs and tell me what you think. (Source: Jeff Leek `http://tinyurl.com/ckatyj2`)

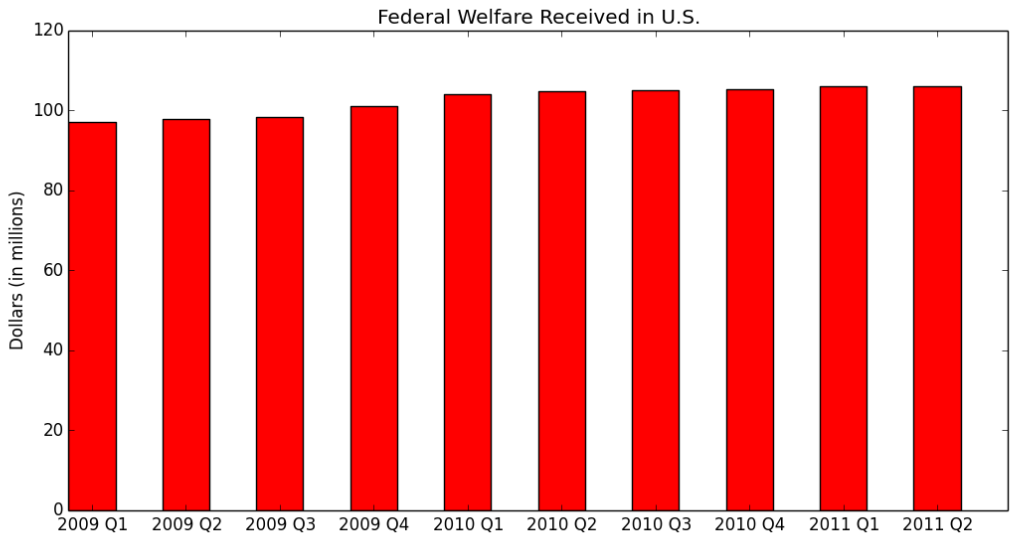
Bad Statistics (Courtesy of Fox News)

Take a look at each of the following graphs and tell me what you think. (Source: Jeff Leek `http://tinyurl.com/ckatyj2`)



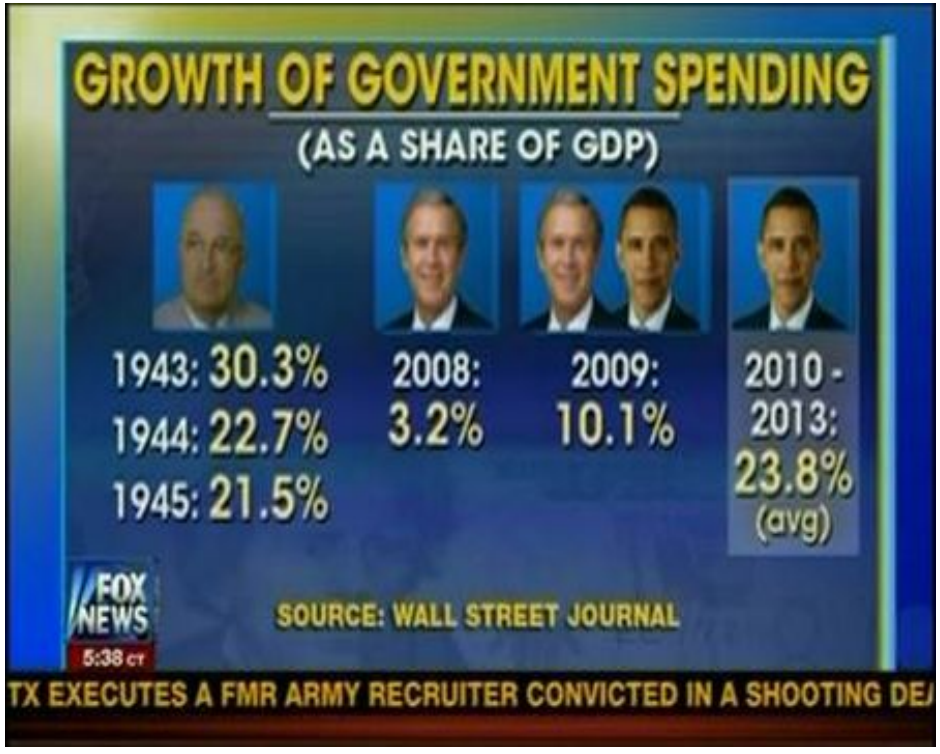
Bad Statistics (Courtesy of Fox News)

Take a look at each of the following graphs and tell me what you think. (Source: Jeff Leek `http://tinyurl.com/ckatyj2`)



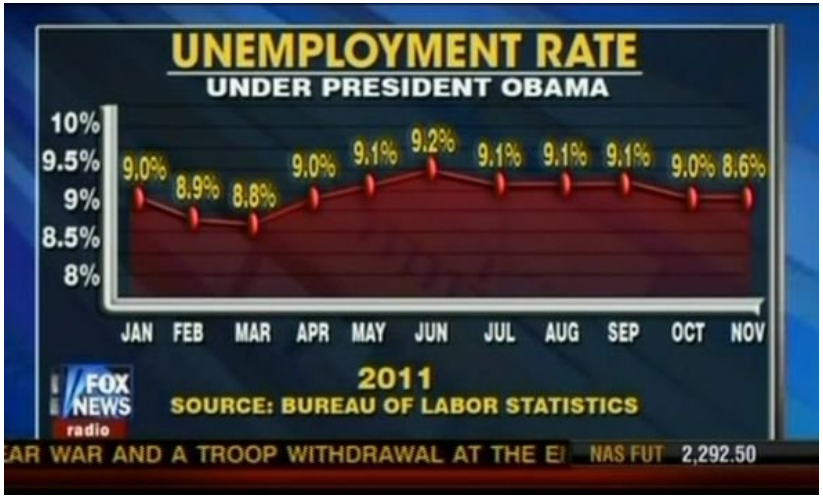
Bad Statistics (Courtesy of Fox News)

Take a look at each of the following graphs and tell me what you think. (Source: Jeff Leek `http://tinyurl.com/ckatyj2`)



Bad Statistics (Courtesy of Fox News)

Take a look at each of the following graphs and tell me what you think. (Source: Jeff Leek `http://tinyurl.com/ckatyj2`)



Summary of Misleading Data

- Items not adding up to 100% (even with a reasonable margin of error)
- Axes not starting at 0
- Incorrectly plotted data points

All of these points are skewings that happened **after** the data was collected! What about errors in the method of collecting data?

Possible Errors in Collecting Data

Possible Errors in Collecting Data

- Poor choice of representative sample population
- Too small of a sample size
- Bias in the questions asked
- Mistakes in identifying appropriate variables (Superbowl winners from the NFC and stock markets).

Errors in Analyzing Data

What errors can we make while analyzing data?

A Puzzle

Consider the two batting averages of the following MLB players

	1995	1996
Derek Jeter	0.250	0.314
David Justice	0.253	0.321

Who is the better overall batter (that is, who has the better overall batting average in both years combined)?

What about this example?

	1995	1996
Joe Brown	$1/1000$	$10/10$
Steve Peters	$0/1$	$198/199$

What about this example?

	1995	1996
Joe Brown	1/1000	10/10
Steve Peters	0/1	198/199

Overall, Joe bats 11/1010 for a batting average of .011. However, Steve is batting an overall 198/200 for a batting average of 0.995, an excellent batting average!

Simpson's Paradox

A shocking answer:

	1995	1996	Total
Derek	$12/48 = 0.250$	$183/582 = 0.314$	$\mathbf{195/630 = 0.310}$
David	$\mathbf{104/411 = 0.253}$	$\mathbf{45/140 = 0.321}$	$149/551 = 0.270$

The original question I gave you could not be answered since we didn't know the respective sample sizes. This is an example where the statistics can be phrased to lead you to believe one thing even though it is incorrect.

Medical Example

A population of 1000 people is suffering from a disease epidemic. Empirical studies have shown that 2% of the population suffer from the disease. There is a test that will determine with 95% accuracy whether you have the disease. More concretely, the test has a 5% false positive rate and a 0% false negative rate.

Suppose that you took the test and the test claims you have the disease. What is the probability that you are **actually** infected?

Medical Example

A population of 1000 people is suffering from a disease epidemic. Empirical studies have shown that 2% of the population suffer from the disease. There is a test that will determine with 95% accuracy whether you have the disease. More concretely, the test has a 5% false positive rate and a 0% false negative rate.

Suppose that you took the test and the test claims you have the disease. What is the probability that you are **actually** infected?

Definition

A **false positive** (crying wolf) occurs when a test returns positive result but the patient is healthy.

A **false negative** (criminal found not guilty in court of law) occurs when a test returns a negative result but the patient is actually infected.

Expectation Table

Let's make a table of the expected outcome of running 1000 tests:

Act.\ Test	Positive	Negative
Positive	20	49
Negative	0	931

Expectation Table

Let's make a table of the expected outcome of running 1000 tests:

Act.\ Test	Positive	Negative
Positive	20	49
Negative	0	931

Note: $49 = 0.05 \cdot 980$

Expectation Table

Let's make a table of the expected outcome of running 1000 tests:

Act.\ Test	Positive	Negative
Positive	20	49
Negative	0	931

Note: $49 = 0.05 \cdot 980$

So if you are diagnosed as positive, this means you have only a $20/(20+49) = 20/69 = 29\%$ chance of **actually being infected!**

Why does this effect happen

The issue here that the test accuracy, while high, is not high enough to match the incidence rate. Having a false positive rate of 5% compared to the fact that only 2% of the population suffers from the disease is a large gap when actually making a diagnosis.

Bayes Theorem

- Bayes Theorem helps us to quantify these observations.
- It is a theorem of probability which helps us to reevaluate probabilities based on new knowledge.
- In Bayes Theorem, after being given certain information, we use this to change the effect on our current probability.

Terminology

- Let A and B be events (these can be say you have a disease or you draw a red ball from a bag of balls etc.)

Terminology

- Let A and B be events (these can be say you have a disease or you draw a red ball from a bag of balls etc.)
- Denote $P(A)$ by the probability of event A occurring (total possible ways divided by total possible outcomes). For example, if A was the event of you picking an even number from the numbers between 1 and 10, this would be $P(A) = 5/10 = 0.5$.

Terminology

- Let A and B be events (these can be say you have a disease or you draw a red ball from a bag of balls etc.)
- Denote $P(A)$ by the probability of event A occurring (total possible ways divided by total possible outcomes). For example, if A was the event of you picking an even number from the numbers between 1 and 10, this would be $P(A) = 5/10 = 0.5$.
- Practice: What is the probability you picked a number with four letters in it? (Call this event B)

Terminology

A Event of an even number from 1 to 10.

B Event of number from 1 and 10 with four letters.

- Denote by $P(A \mid B)$ the probability of A **given that** event B has occurred. For example, if A was as above and B is the event that you picked a number that has four letters in it, then $P(A \mid B) = 1/3$ (only four, five and nine have four letters).

Terminology

A Event of an even number from 1 to 10.
B Event of number from 1 and 10 with four letters.

- Denote by $P(A \mid B)$ the probability of A **given that** event B has occurred. For example, if A was as above and B is the event that you picked a number that has four letters in it, then $P(A \mid B) = 1/3$ (only four, five and nine have four letters).
- Notice that $P(A \mid B) = P(A \text{ and } B)/P(B)$. This is the mathematically formal definition.
- Practice: What is $P(B \mid A)$?

Terminology

A Event of an even number from 1 to 10.

B Event of number from 1 and 10 with four letters.

- Denote by $P(A \mid B)$ the probability of A **given that** event B has occurred. For example, if A was as above and B is the event that you picked a number that has four letters in it, then $P(A \mid B) = 1/3$ (only four, five and nine have four letters).
- Notice that $P(A \mid B) = P(A \text{ and } B)/P(B)$. This is the mathematically formal definition.
- Practice: What is $P(B \mid A)$?
- What is the relationship between $P(A \mid B)$, $P(B \mid A)$, $P(A)$ and $P(B)$?

Bayes Theorem

Let A and B be events. Then

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

Reframing the Disease Problem

Let's rephrase the Disease problem in terms of these variables:

- Let A be the event that you have the disease and A^c be the event that you do not have the disease.

Reframing the Disease Problem

Let's rephrase the Disease problem in terms of these variables:

- Let A be the event that you have the disease and A^c be the event that you do not have the disease.
- Let B be the event that you tested positive for the disease.

Reframing the Disease Problem

Let's rephrase the Disease problem in terms of these variables:

- Let A be the event that you have the disease and A^c be the event that you do not have the disease.
- Let B be the event that you tested positive for the disease.
- Then the question claimed that $P(A) = 2/100$. Further, $P(B \mid A) = 1$ and $P(B \mid A^c) = 1/20$. Then $P(B) = 69/1000$ since $P(B) = P(B \text{ and } A) + P(B \text{ and } A^c)$

Reframing the Disease Problem

Let's rephrase the Disease problem in terms of these variables:

- Let A be the event that you have the disease and A^c be the event that you do not have the disease.
- Let B be the event that you tested positive for the disease.
- Then the question claimed that $P(A) = 2/100$. Further, $P(B \mid A) = 1$ and $P(B \mid A^c) = 1/20$. Then $P(B) = 69/1000$ since $P(B) = P(B \text{ and } A) + P(B \text{ and } A^c)$
- Thus, by Bayes Theorem:

$$P(A \mid B) = P(B \mid A)P(A)/P(B) = \frac{2/100}{69/1000} = \frac{20}{69}$$

Which Test is More Accurate?

1. The test from before
2. A test that always diagnoses a person as healthy

Which Test is More Accurate?

1. The test from before
2. A test that always diagnoses a person as healthy

The test from before was 95% accurate. The new test is actually more accurate as it correctly diagnoses people 98% of the time (it only misses the twenty infected people!)

Accuracy Paradox

Even though the "always not infected" test seems to be more accurate, it is actually far less useful of a test. This emphasizes how important it is to get the test correct because there are many bad tests that give you more accurate results.

References

- http://tinyurl.com/ckatyj2
- https://en.wikipedia.org/wiki/Category:Statistical_paradoxes
- "The Signal and the Noise" - Nate Silver