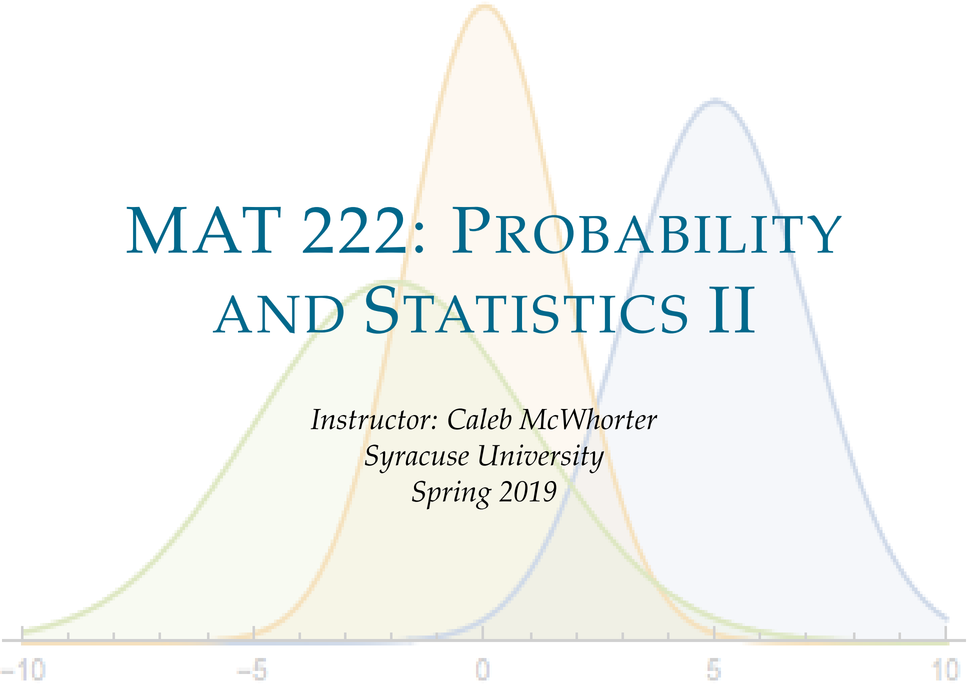
MAT 222: Probability and Statistics II

Instructor: Caleb McWhorter
Syracuse University
Spring 2019



What is Statistics?

Definition (Statistics)

Science of collecting, organizing, analyzing, interpreting, and presenting data.

Statistics is the Science of Data

Why do we care?

Statistics is *EVERYWHERE*!

- Accounting
- Advertising
- Aerospace Engineering
- Anthropology
- Architecture
- Biology
- Chemistry
- Child & Family Policy
- Civil Engineering
- Cognitive Science
- Communication
- Computer Art & Animation
- Computer Game Design
- Computer Science
- Cybersecurity
- Drama
- Earth Science
- Economics
- Education Studies
- Electrical Engineering
- Environmental Engineering
- Exercise Science
- Fashion Design
- Film
- Finance
- Fine Arts
- Food Science
- Forensic Science
- Geography
- Gerontology
- Graphic Design
- Health & Exercise Science
- Health & Wellness
- History
- Industrial Design
- Information Technology
- International Business
- International Relations
- Journalism
- Law
- LGBTQIA+ Studies
- Linguistic Studies
- Management Studies
- Marketing
- Mathematics
- Mechanical Engineering
- Music
- Musical Theater
- Neuroscience
- Nutritional Science
- Photography
- Physics
- Policy Study
- Political Science
- Psychology
- Public Communication
- Public Relations
- Real Estate
- Religion
- Renewable Energy
- Retail Management
- Social Welfare & Social Work
- Sociology
- Sport Analytics
- Sport Management
- Systems & Information Science
- TVR
- Women's and Gender Studies
- Writing & Rhetoric

Furthermore, Statistics is used for…

- …making policy and funding decisions.
- …determining insurance rates.
- …predicting stock values and financial trends.
- …evaluating effectiveness of drugs and treatments in medical fields.
- …estimating fundamental constants and values in the Sciences.
- …examining human behaviors.
- …finding trends in History.

What Types of Things Can Statistics Do?

Explain Patterns

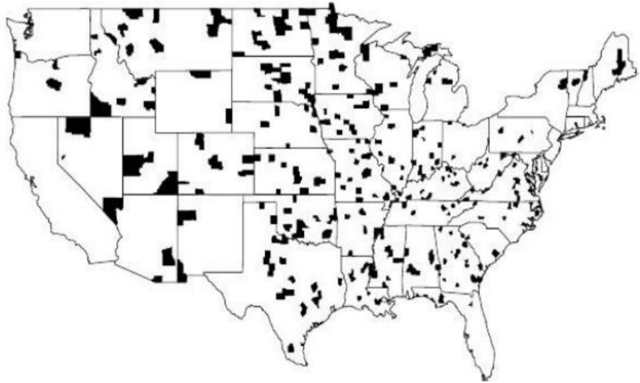


Figure 1: Counties with the highest 10% age-standardized death rates of kidney/ureter cancer in males, 1980–1989.

A. Gelman and D. Nolan (2017). *Teaching Statistics: A Bag of Tricks*. Oxford University Press.

Make Data Driven Decisions

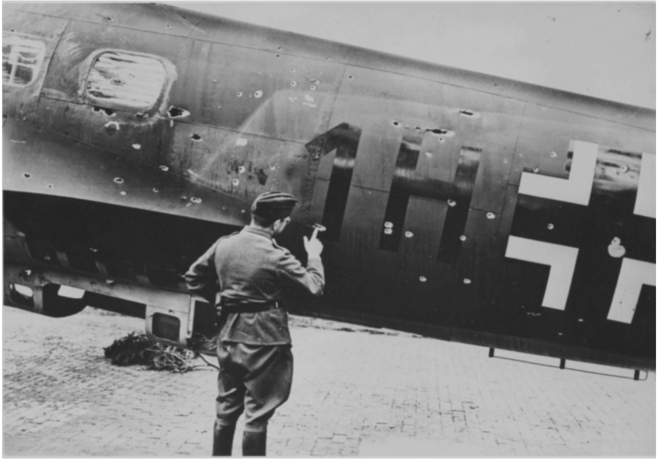


Figure 2: Soldier examing bullet holes on a bomber.

Evgeniy. "The German officer is studying bullet holes on the fuselage of the bomber He.111, returning from combat sortie." War Thunder, 01/25/2016. http://waralbum.ru/274297

Reveal Human Nature

Figure 3: Monty Hall in 'Let's Make a Deal'.

Frauenfelder, Mark. "Monty Hall." BoingBoing, 10/01/2017. https://boingboing.net/2017/10/01/monty-hall-1921-2017.html. Accessed 05/20/2018

Describe the Universe

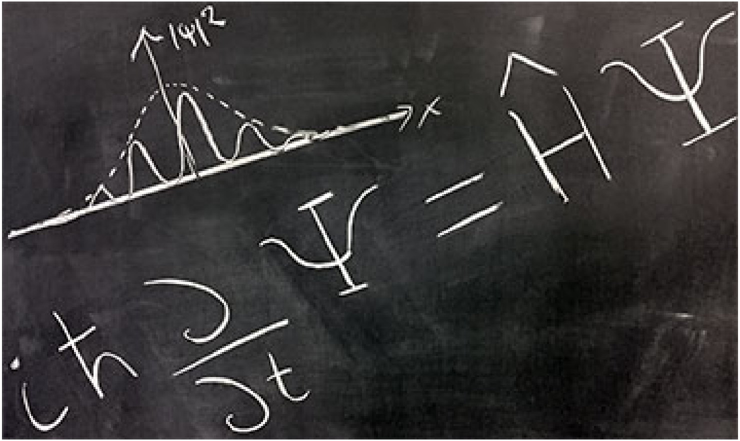


Figure 4: Quantum Mechanics on a blackboard.

edX. "Quantum Mechanics: Quantum physics in 1D Potentials." Online video clip. https://www.youtube.com/watch?time_continue=2&v=0jmW2PeQ-oQ. YouTube. Youtube, 10/02/2017. Web 05/20/2018

Inform & Misinform

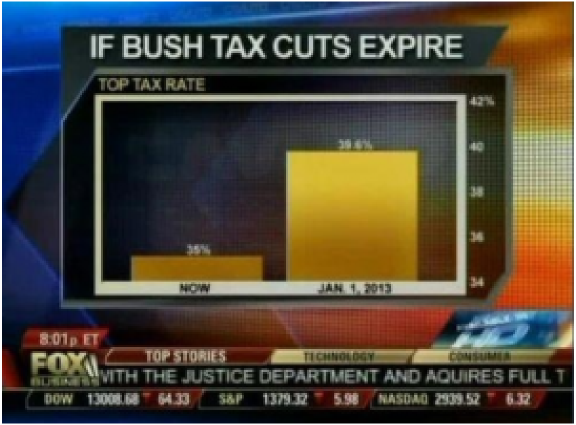


Figure 5: If Bush tax cuts expire.

Robbins, Naomi. "Another Misleading Graph of Romney's Tax Plan." 08/04/2012. https://www.forbes.com/sites/naomirobbins/2012/08/04/another-misleading-graph-of-romneys-tax-plan/#b71ed4133b89. Accessed 05/20/2018

In Fact: Statistics Can Be “The Science of Bullshit”

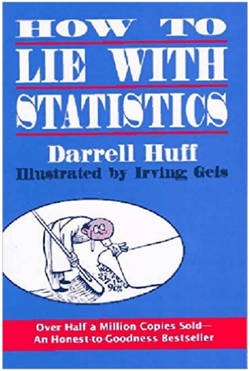

HOW TO
LIE WITH
STATISTICS
Darrell Huff
Illustrated by Irving Geis
Over Half a Million Copies Sold—
An Honest-to-Goodness Bestseller

Figure 6: How to lie with Statistics.

Amazon. “How to lie with Statistics.” https://www.amazon.com/How-Lie-Statistics-Darrell-Huff/dp/0393310728. Accessed 05/21/2018

Detect Fraud

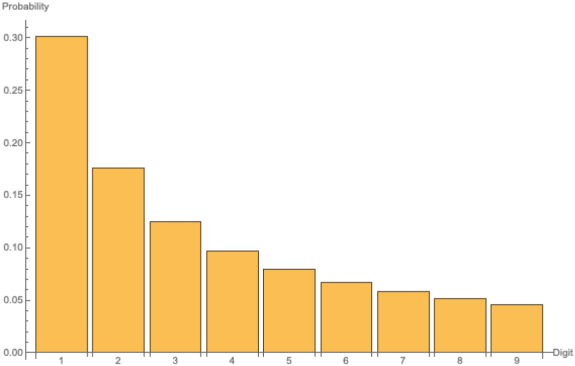


Figure 7: Distribution of of percentage of leading digits according to Benford's Law.

Two Main Types of Statistics

Definition (≈Descriptive Statistics)

Describes data, e.g. how one presents data or interesting characteristics: mean, median, min/max, standard deviation, skew, kurtosis, etc..

Definition (Inferential Statistics)

Draws inferences from data, e.g. z-statistic, t-statistic, χ^2, F-ratio, etc..

How does one approach Statistics?

“Statistical Method”

1. Ask a question.
2. Collect data.
3. Analyze the data.
4. Interpret the results.
5. Present the conclusions

“Statistical Method”

1. Ask a question.
2. Collect data.
3. Analyze the data.
4. Interpret the results.
5. Present the conclusions

VS

Scientific Method

1. Observation
2. Question
3. Hypothesis
4. Experiment
5. Analysis
6. Conclusion

1. Ask a question

Ask a Question

- “Does opening needle exchange clinics reduce rates of HIV contractions in a region?”
- “Are there more ‘extreme weather’ events now than in decades past?”
- “Does taking a daily aspirin in old age reduce rates of heart attack or stroke?”
- “Is there a difference between crime rates for natural born citizens compared to legal/illegal immigrants?”
- “Does gun legislation actually reduce rates of murder or violent crime?”
- “Are there differences in police shooting rates for different races?”

2. Design an experiment/Collect data

Design an experiment/Collect Data

This is what makes Statistics difficult!

Design an experiment/Collect Data

This is what makes Statistics difficult!

- Given perfect data, the mathematics is 'simple' and *always* gives you predictions with *exact* measures of error.

Design an experiment/Collect Data

This is what makes Statistics difficult!

- Given perfect data, the mathematics is ‘simple’ and *always* gives you predictions with *exact* measures of error.
- But perfect experiments and data do not exist!
- No amount of mathematics can make up for poor experimentation or bad data.

Design an experiment/Collect Data

This is what makes Statistics difficult!

- Given perfect data, the mathematics is ‘simple’ and *always* gives you predictions with *exact* measures of error.
- But perfect experiments and data do not exist!
- No amount of mathematics can make up for poor experimentation or bad data.
- Designing experiments carefully and checking assumptions carefully are key to good statistics.
- Good Statistics always makes all these factors clear and discusses them!

Why is experimentation & data collection difficult?

Question Design is Difficult

Example (Loaded Questions)

1. Do you believe the President's choice of an inexperienced candidate should be confirmed?
2. Do you believe there should be locations to inject illegal drugs?
3. "Should a smack as part of good parental correction be a criminal offence in New Zealand?"[7]

[7] Newshub. "Anti-smacking debate goes to referendum.". 06/15/2009. Accessed 05/21/2018.

Question Design is Difficult

Example (Leading Questions)

1. Do you believe responsible parents should discipline their children?
2. How good was the new film?
3. Do you have problems with your employer?

Question Design is Difficult

Example (Double Barreled Questions)

1. Are you satisfied or dissatisfied are you with your current level of salary and benefits?
2. When was the last time you showered or shaved?
3. Is the new textbook useful for students and teachers?

Question Design is Difficult

Example (Absolutes)

1. Do you believe the President ever lies?
2. Do you always shower in the morning?
3. Should one never terminate a pregnancy?

Question Design is Difficult

Example (Language/Knowledge/Experience Considerations)

1. Should CPS receive more state and/or federal funding?
2. Do you believe individuals convicted of illicit treatment of animals have predilections which may indicate they will participate in future in other felonious behavior?
3. Which suite of cards is your favorite?

Question Design is Difficult

Example (Poor Phrasing/Question Design)

1. Does it seem possible or impossible that the Nazi extermination of the Jews never happened?
2. Did you grow up with a family with a mother and father or only a mother and father?
3. In the past 30 days, when you might have consumed chewing gum, did you share it with anyone?

Costs of Data Collection

- Pay people to design, distribute, collect, and analyze survey/experiment.

- Experiment could involve large and expensive apparatuses.

- Does one offer incentives? [This could create other issues.]

Practical vs Statistical Significance

Definition (Statistical Significance)

When a result is unlikely to have occurred by chance.

Example

ProCare Industries offered Gender Choice to increase chances a baby would be born with a desired sex. It increased the chance to 52% of girls when desired. [Study on 10,000 couples, 5200 had girls. 0.003% chance of this happening.] But this is not practically significant. [Note: actual probability of girl 48.8%.]

Sampling Issues

- A liberal news website asks whether visitors are satisfied with the current president.
- A survey about a women's rights issue that mostly surveyed male individuals or a drug is given to only female subjects and not male subjects.
- A otherwise well designed survey where participants can see the results as they arrive.
- An internet poll about computer usage.
- A survey about health where individuals are asked about their weight.

Other Important Considerations

- The survey/experiment should actually text the question at hand without bias.

- Replication: The experiment/survey should be able to be repeated.

- Blinding: Should the subjects know the point of the experiment or whether they are receiving treatment? One must be aware of the *placebo effect*. Whenever possible (especially in medical studies), double-blinding should be used whenever possible.

Double-blinding is when *both* the subjects and the experimenters do not know who is or is not receiving treatment.

Other Important Considerations

- Randomization: Where individuals are assigned to different groups through a process of random selection.
- Timeframe: How long will the data collection take? Will it be useful/relevant by then? What are the costs with the timeframe? What if participants drop out or disappear? Will short term data be useful?
- Confusing Responses/Non-response: What is a survey response is (partially) illegible, confusing, or otherwise problematic? What if some of the responses are missing or incomplete?

Non-response should always be considered a source of bias.

Bad Experiment/Survey = Bad Data

Good Experiment/Survey $\neq$ Good Data

3. Analyze the data.

Analyze the Data

This is the ‘easy’ part! [But this part can be ruined by bad data.]

Remark

Analyzing the data should *always* begin by graphically examining the data (whenever physically possible). This is especially important to check normality assumptions.

4. Interpret the results.

Interpret the results.

Interpret the results, being careful to consider and discuss the assumptions required for statistical validity. But there are even more issues to consider!

Remark

Even in a well designed and executed study, the data and analysis may not be sufficient.

Definition (Confounding Variable)

A variable that influences the dependent and independent variable, causing an association. Also called a lurking or hidden variable.

Example (Confounding Variables)

1. Birth order and Presence of Down’s Syndrome: Here a hidden variable is the mother’s age.
2. Exercise and Weight Gain by sex: Possible confounding variables are age, amount eaten, occupation, etc..
3. Murder rate and ice cream consumption: Possible confounding variable is weather.

Remark

One must be careful when examining and interpreting data.

Correlation $\neq$ Causation

EXAMPLE

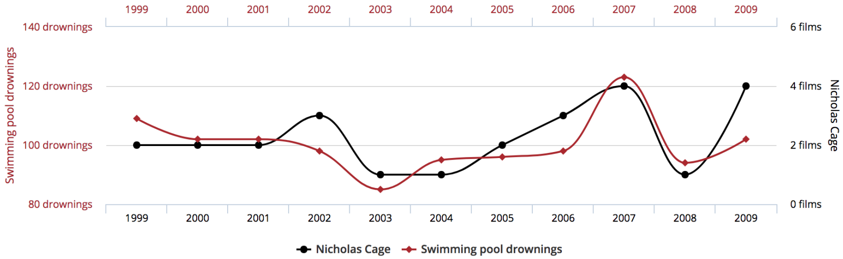


Figure 8: Number of people who drowned by falling into a pool correlates with films Nicholas Cage appeared in ($r = 0.666004$).

T. Vigen. *Spurious Correlations*. http://www.tylervigen.com/spurious-correlations. Accessed 05/21/2018

Example

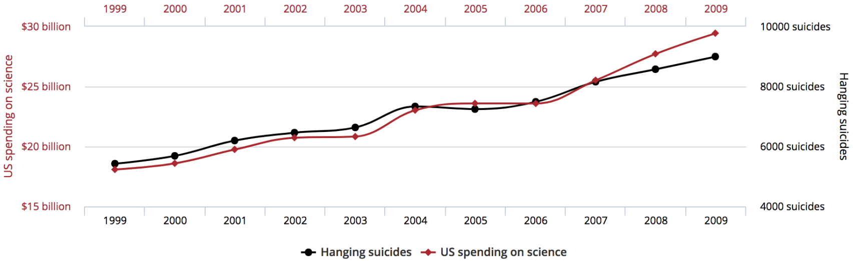


Figure 9: US spending on science, space, and technology correlates with suicides by hanging, strangulation, and suffocation ($r = 0.99789126$).

T. Vigen. *Spurious Correlations*. http://www.tylervigen.com/spurious-correlations. Accessed 05/21/2018

Example

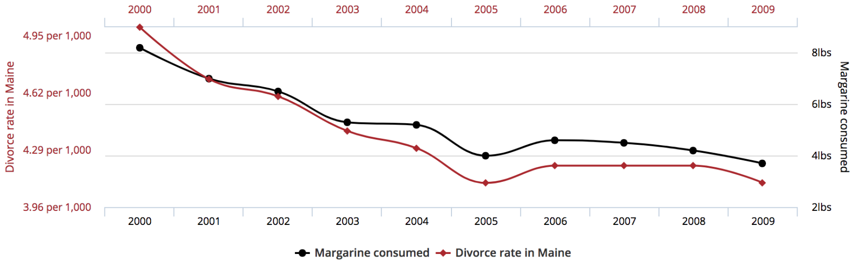


Figure 10: Divorce rate in Maine correlates with per capita consumption of margarine ($r = 0.992558$).

T. Vigen. *Spurious Correlations*. http://www.tylervigen.com/spurious-correlations. Accessed 05/21/2018

EXAMPLE

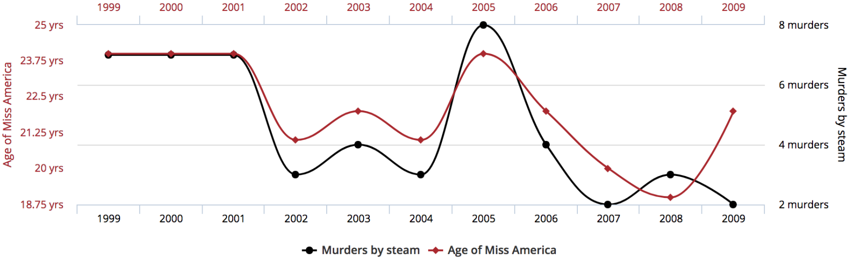


Figure 11: Age of Miss America correlates with murders by steam, hot vapors, and hot objects ($r = 0.870127$).

T. Vigen. *Spurious Correlations*. http://www.tylervigen.com/spurious-correlations. Accessed 05/21/2018

Example

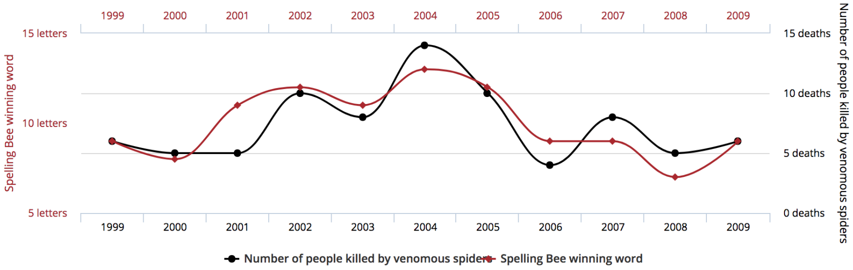


Figure 12: Letters in winning word of Scripps National Spelling Bee correlates with number of people killed by venomous spiders ($r = 0.8057$).

T. Vigen. *Spurious Correlations*. http://www.tylervigen.com/spurious-correlations. Accessed 05/21/2018

Example (More Seriously)

Example

At young ages, there is a strong correlation between IQ and astrological sign.

Example (More Seriously)

Example

At young ages, there is a strong correlation between IQ and astrological sign. This is because early on your age may determine whether you have started school or not, i.e. early/late starters. Children born in certain months are more likely to be attending school, and thus will have higher IQ scores than children who have not or have had less years of schooling. But after a few years, this difference is 'corrected' by more years of schooling for each group of children.

5. Present the conclusions.

Presenting the Conclusions

- This includes discussing experimental design.
- Present the analysis of the data.
- Discuss what the math does/does not say about the data.
- When appropriate, use graphical representation of the data.

Graphs Should Lead, Not Mislead

Example

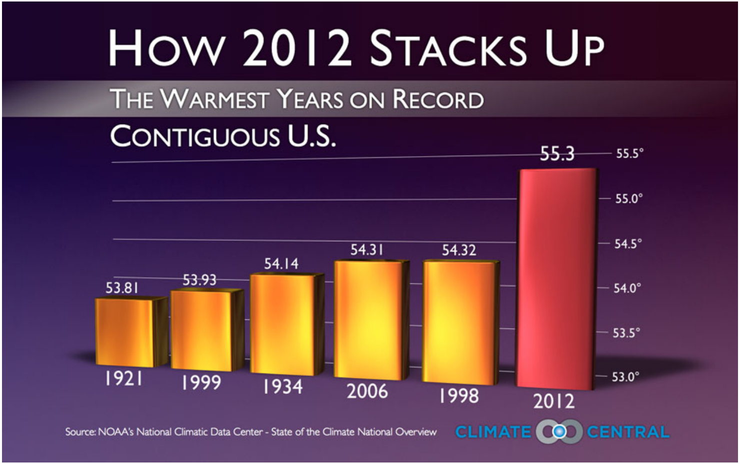
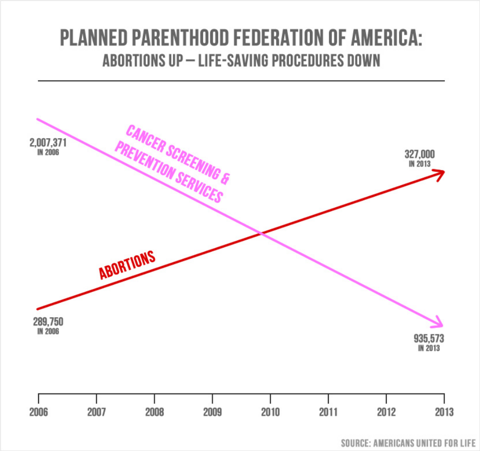


Figure 13: Climate data.

A. Gelman, "How 2012 stacks up: The worst graph on record?". 01/08/2013. http://themonkeycage.org/2013/01/how-2012-stacks-up-the-worst-graph-on-record/. Accessed 05/22/2018

Example



SOURCE: AMERICANS UNITED FOR LIFE

Figure 14: Planned Parenthood services.

L. Qiu, "Chart shown at Planned Parenthood hearing is misleading and 'ethically wrong"'. 10/01/2015. http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/. Accessed
05/22/2018

Example (Corrected)

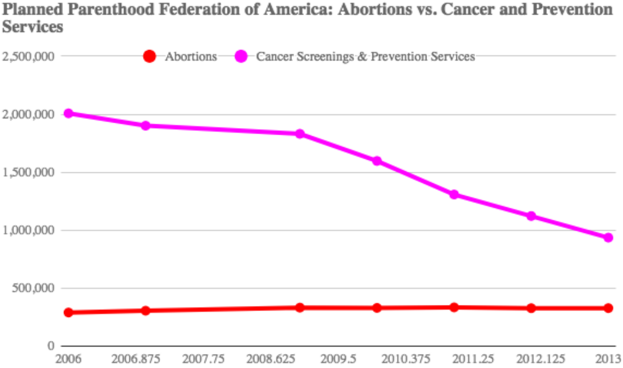


Figure 15: Planned Parenthood services.

L. Qiu, “Chart shown at Planned Parenthood hearing is misleading and ’ethically wrong”’. 10/01/2015. http://www.politifact.com/truth-o-meter/statements/2015/oct/01/jason-chaffetz/chart-shown-planned-parenthood-hearing-misleading-/. Accessed
05/22/2018

Example

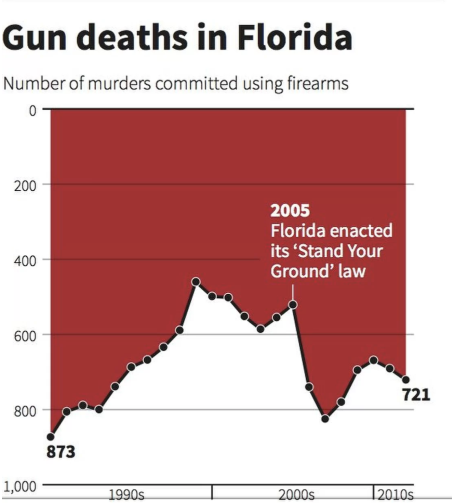


Figure 16: Stand your ground (Source: Florida Department of Law Enforcement).

M. Lallanilla. "Misleading Gun-Death Chart Draws Fire". 04/23/2014. https://www.livescience.com/45083-misleading-gun-death-chart.html. Accessed 05/22/2018

Example (Corrected)

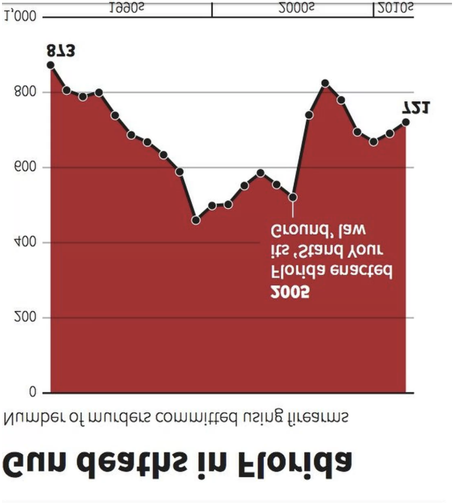
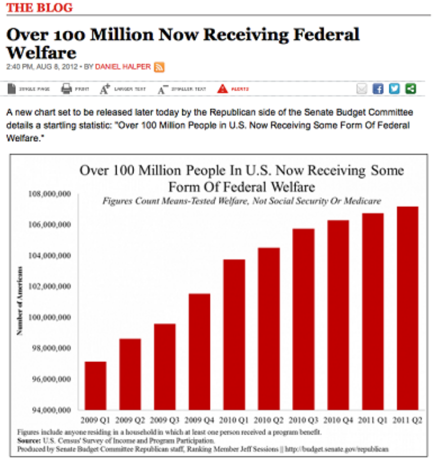


Figure 17: Stand your ground (Source: Florida Department of Law Enforcement).

M. Lallanilla. “Misleading Gun-Death Chart Draws Fire”. 04/23/2014. https://www.livescience.com/45083-misleading-gun-death-chart.html. Accessed 05/22/2018

EXAMPLE



THE BLOG

Over 100 Million Now Receiving Federal Welfare

2:40 PM, AUG 8, 2012 • BY DANIEL HALPER

A new chart set to be released later today by the Republican side of the Senate Budget Committee details a startling statistic: "Over 100 Million People in U.S. Now Receiving Some Form Of Federal Welfare."

Figures include anyone residing in a household in which at least one person received a program benefit.
Source: U.S. Census' Survey of Income and Program Participation.
Produced by Senate Budget Committee Republican staff, Ranking Member Jeff Sessions || http://budget.senate.gov/republican

Figure 18: Welfare.

Stephanie. "Misleading Graphs". 01/24/2014. http://www.statisticshowto.com/misleading-graphs/. Accessed 05/22/2018.

Example

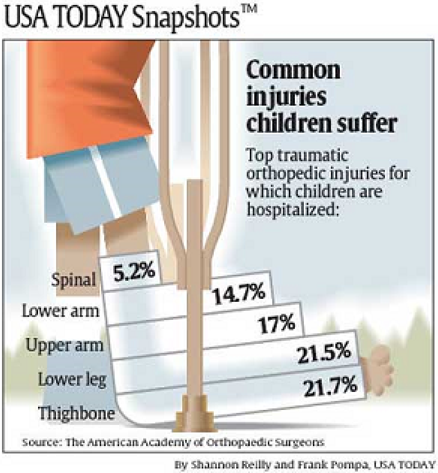


Figure 19: Common injuries.

Stephanie. "Misleading Graphs". 01/24/2014. http://www.statisticshowto.com/misleading-graphs/. Accessed 05/22/2018.

EXAMPLE

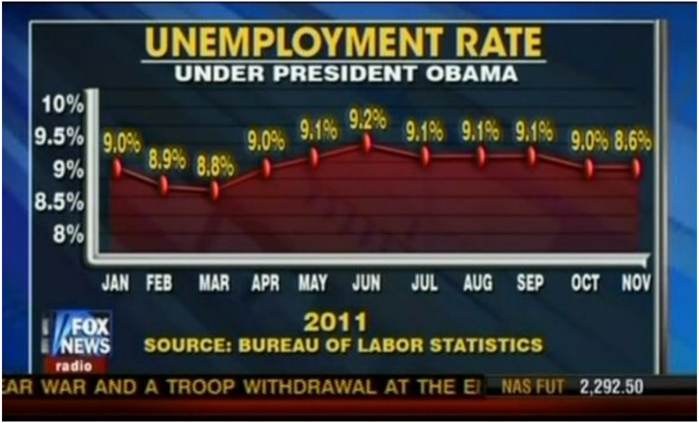


Figure 20: Unemployment rate.

Stephanie. “Misleading Graphs”. 01/24/2014. http://www.statisticshowto.com/misleading-graphs/. Accessed 05/22/2018.

EXAMPLE (CORRECTED)

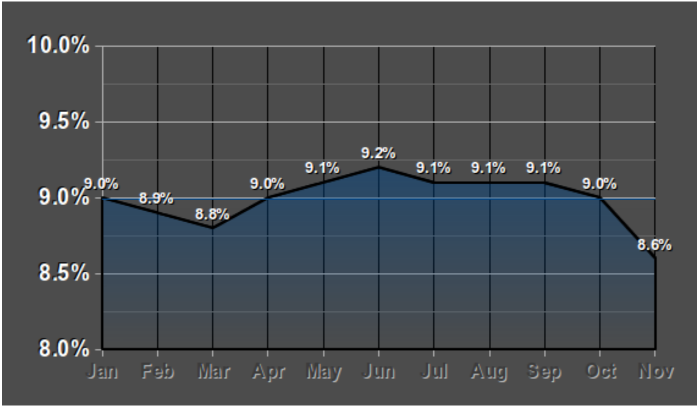


Figure 21: Unemployment rate.

Stephanie. “Misleading Graphs”. 01/24/2014. http://www.statisticshowto.com/misleading-graphs/. Accessed 05/22/2018.

EXAMPLE

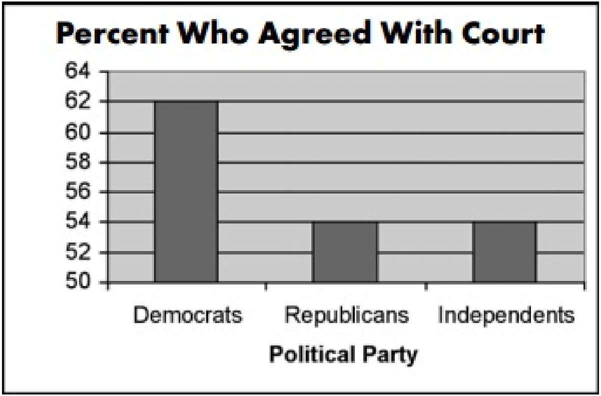


Figure 22: Should Terry Schiavo be removed from life support?

Stephanie. “Misleading Graphs”. 01/24/2014. http://www.statisticshowto.com/misleading-graphs/. Accessed 05/22/2018.

Example

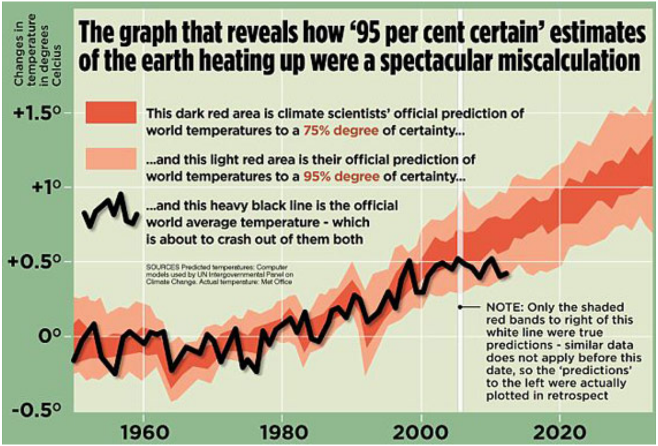


Figure 23: Global Warming.

Stephanie. "Misleading Graphs". 01/24/2014. http://www.statisticshowto.com/misleading-graphs/. Accessed 05/22/2018.

Should we trust Statistics?

Scientific Studes: Last Week Tonight with John Oliver