TECHNISCHE UNIVERSITÄT CHEMNITZ

How to Lie with Statistics

Source: http://photos1.blogger.com/x/blogger/4986/984/1600/882132/Irving_Geis.jpg

Jessica Dheskali, M.A.
TU Chemnitz
English and American Studies
English Language and Linguistics
Reichenhainer Straße 39, room 218
09126 Chemnitz

email: jessica.dheskali@phil.tu-chemnitz.de

TECHNISCHE UNIVERSITÄT CHEMNITZ

1. Introduction

When are you exposed to statistical information in your everyday life?

Are statistics reliable?

"99% of people that drink water die within 100 years." => **water kills.**

Misleading Statistics:

- are simply the misuse – purposeful or not – of numerical data
- results provide misleading information to the receiver, who then believes something wrong if he or she does not notice the error or does not have the full data picture
- numbers don't lie, but they can in-fact be used to mislead with half-truths

(Source: https://www.datapine.com/blog/misleading-statistics-and-data/)

TECHNISCHE UNIVERSITÄT CHEMNITZ

2. Background Information

Why are statistics used in academic & journalistic writing and in advertisements?

TECHNISCHE UNIVERSITÄT CHEMNITZ

Why are statistics used in academic & journalistic writing and in advertisements?

- to summarize results
- to describe a data set or e.g. a certain group of people
- to make arguments (seem) more scientific and credible
- to make statements (appear) more definite
- to highlight the importance of something
- to manipulate the audience/readers (to make them e.g. buy a certain product)
- to provide convincing information/ to add factual weight to an argument
- to enable writers to draw conclusions and argue for/against something without sounding speculative or vague
- to mislead us
- to provide a type of evidence that is difficult to refute
- …

Descriptive Statistics in Academic Writing

- require you to grapple with numbers in a real world context, to describe observations using numbers, and to use the numbers in your own analyses and arguments
- ask you to draw conclusions based on numerical or other quantitative evidence, which is either supplied or which you must develop
- help you get an in-depth understanding of all the variables involved in a study
- allow you to summarize a big data set and your findings and to show only what is meaningful or significant
- show e.g. whether or not a single variable has an impact or not on other aspects or whether or not the groups involved are similar or different from each other
- give meaning to your data set and justify whether you have achieved the purpose of your study or not and if your results are significant
- paint a picture of the distribution of your data by e.g. stating the mean, the mode, the median, the range (**central tendency**), and the standard deviation (one form of **variability**)
- are used because in most cases, it is not possible to present all of your data in any form that your reader will be able to quickly interpret

Source: https://www.aresearchguide.com/writing-with-descriptive-statistics.html

TECHNISCHE UNIVERSITÄT CHEMNITZ

However, sometimes

- sources or numbers are given more credibility than they should be
- key information needed to evaluate the numbers is missing
- numbers are taken out of context
- correlation and causation are ignored or assumed
- graphs are displayed wrongly or funky graphics are used
- issues are defined in different ways
- percent and percentage points are used (wrongly) to mislead
- aspects were excluded on purpose
- big or total numbers are used as shock values
- the “average” is not displayed or calculated correctly
- “apples are compared with oranges”

TECHNISCHE UNIVERSITÄT CHEMNITZ

Question evidence!!!

- statistics do not exist without people

So, ask yourself:

- What is the source of the statement and/or data?
- Is the sample of an adequate size?
- Is the sample representative?
- Who collected the data and who counted?
- What was counted, compared or classified?
- Why/ with which purpose was it counted?
- Can these aspects, groups etc. even be compared?
- Were all aspects needed taken into consideration?
- How is the information reported?
- Was a certain form of visualization used to mislead the receiver?

3. Discussion: Example 1

Picture 1: *Trust in the printed media*, statistics presented by Dušica L.

TECHNISCHE UNIVERSITÄT CHEMNITZ

Example 2

(Huff 2010: 16)

Group 2:

"According to the *New York Sun*, the average Yaleman makes $ 25,109 a year. "

What stands out here?
Which aspects could you question?
Which background information do you need?

Is this sample representative?

TECHNISCHE UNIVERSITÄT CHEMNITZ

“Every 11 minutes, a single falls in love through Parship.”

“Unstatistik des Monats” (12/2015, http://www.rwi-essen.de/unstatistik/50/)

Immer noch Konfusion bei Kriminalität

Unstatistik vom 27.04.2018

Die Unstatistik April 2018 ist die Zahl 14.864. So viele erfasste Straftaten pro 100.000 Einwohner gab es im Jahr 2017 in Frankfurt am Main. Die Stadt führt damit die Kriminalitätsliga in Deutschland an, melden unter anderem die Frankfurter Rundschau und der Tagesspiegel.

Aber tut sie das wirklich? Zunächst einmal gibt es große Unterschiede über Raum und Zeit bei der Erfassung von Kriminalität. In der einen Gemeinde schaut man bei Rauschgiftdelikten lieber weg, in der anderen wird ermittelt. Zudem pendeln in Frankfurt rund 300.000 Menschen täglich zur Arbeit ein, rund 60 Millionen Fluggäste kamen 2017 auf dem Flughafen Frankfurt an oder flogen ab. Alle von diesen Menschen verübten oder durch diese Menschen erlittenen Straftaten gehen auf das Konto der Stadt Frankfurt. In München dagegen gehört der Flughafen den Landkreisen Erding und Freising an.

Für einen sinnvollen Vergleich der Kriminalität über Gemeinden oder Länder hinweg wäre es also besser, die Zahl der Straftaten auf die Zahl der potentiellen Opfer und Täter und nicht auf die gemeldeten Einwohner zu beziehen.

Source: http://www.rwi-essen.de/unstatistik/78/

TECHNISCHE UNIVERSITÄT CHEMNITZ

"Never forget the 6-foot-tall man who drowned crossing the stream that was 5 feet deep on average."

-> 6-ft. tall and drowning in river of 5-ft average depth. How is this possible?

A river of average depth 5-feet is not deep uniformly. It can be 2-ft deep at one location and 7-ft at another. It can be 4-ft at one place and 12 or more feet at others.

Group work I: handout mean, median, mode, range

Task:
Define these four terms. Explain them to everyone as group!
Find the mean etc. for the given example.

4. Definitions of Key Terms: Measures of Central Tendency and Other Commonly Used Descriptive Statistics

- The mean, median, and the mode are all measures of central tendency and are all different forms of 'the average.'
- When writing statistics, avoid saying 'average' because it is difficult, if not impossible, for your reader to understand if you are referring to the mean, the median, or the mode.
- For a large number of statistics, it is best to visualize them in form of a table or a graph and to capture the main statistics in the text to avoid describing all results in your text.

TECHNISCHE UNIVERSITÄT CHEMNITZ

“99% of people that drink water die within 100 years.” => water kills.

Group work II:

Together with your neighbour, come up with a similar statement and let the others explain why it is wrong.

TECHNISCHE UNIVERSITÄT CHEMNITZ

Causation and Correlation, Cause and Effect

Is it true that A causes B because A is correlated to B?
Does B go up if A goes up?

Examples:

(1) WhatsApp message – phone freezes

(2) violent TV shows -> kids become more violent

(3) head lies make you healthier

(4) smoking -> bad grades

(5) vegetarians have higher income than meat-eaters

TECHNISCHE UNIVERSITÄT CHEMNITZ

Causation and Correlation, Cause and Effect

Correlation is a statistical measure (expressed as a number) that describes the size and direction of a relationship between two or more variables. A correlation between variables, however, does not automatically mean that the change in one variable is the cause of the change in the values of the other variable.

Causation indicates that one event is the result of the occurrence of the other event; i.e. there is a causal relationship between the two events. This is also referred to as cause and effect.

Causality is the area of statistics that is commonly misunderstood and misused by people in the mistaken belief that because the data shows a correlation that there is necessarily an underlying causal relationship.

Source: https://www.abs.gov.au

TECHNISCHE UNIVERSITÄT CHEMNITZ

Percent, Percentage Points: Examples

Suppose Bisera has $1000 and Irina has $1500.

o Irina has 50% more money than Bisera,
o Bisera has 67% as money as Irina.
o Bisera has 40% of the total ($2500)
o Irina has 60% of the the total.
o ….

- drop out rate (2017: 5%, 2018: 10%)

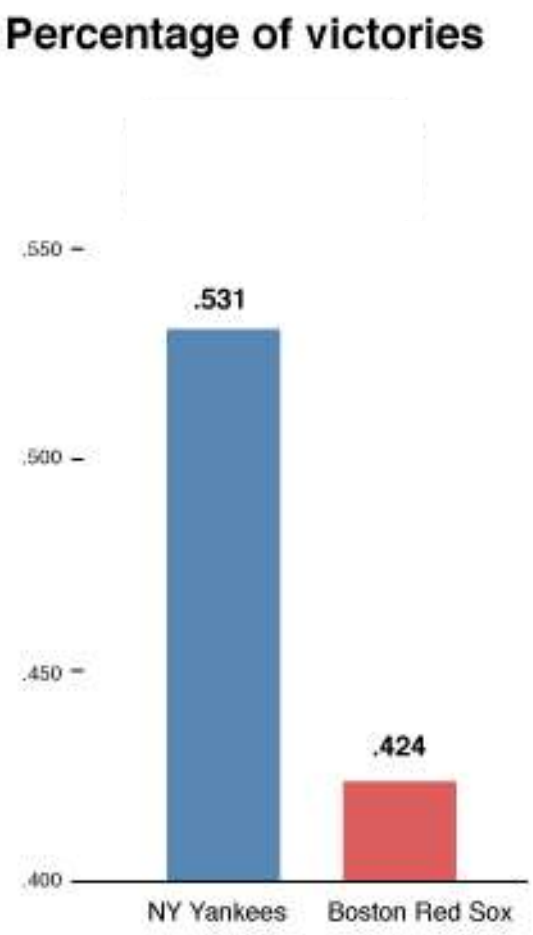
- birth control pill (1/7000 -> 2/7000)

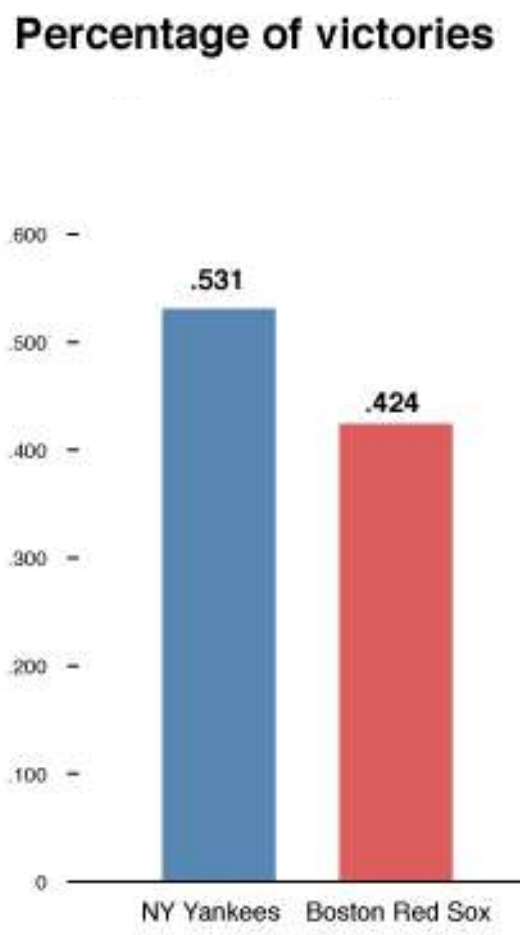
-> increase by 100% ?

“67 % of the British people think that Trump is…“

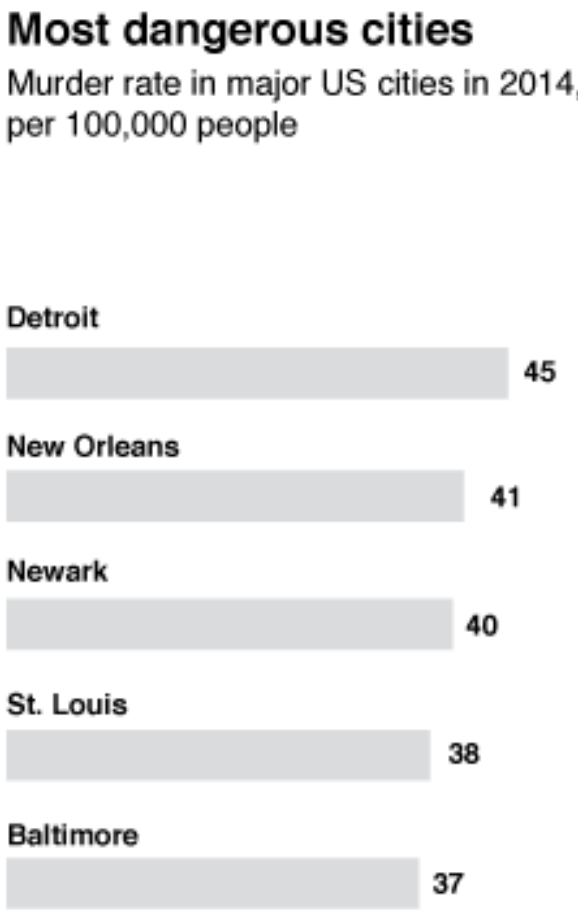
Graphs and diagrams

- when reading an article etc., be especially skeptical of unlabeled graphs -> note graph’s labels along the axes
- graphs are often used to create an impression or mislead people

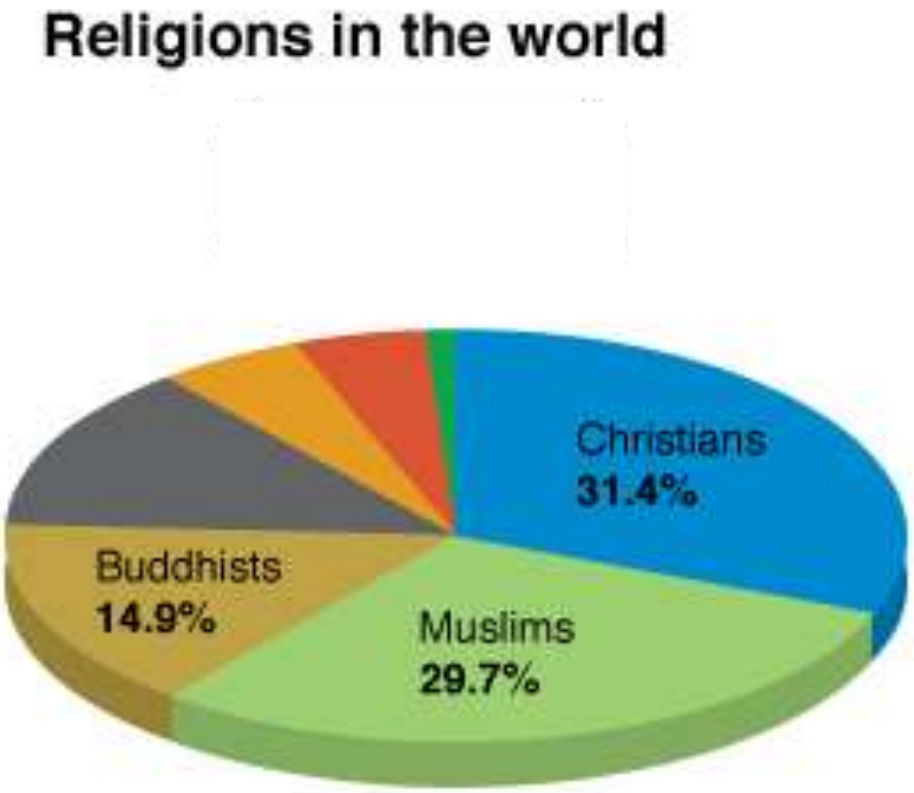


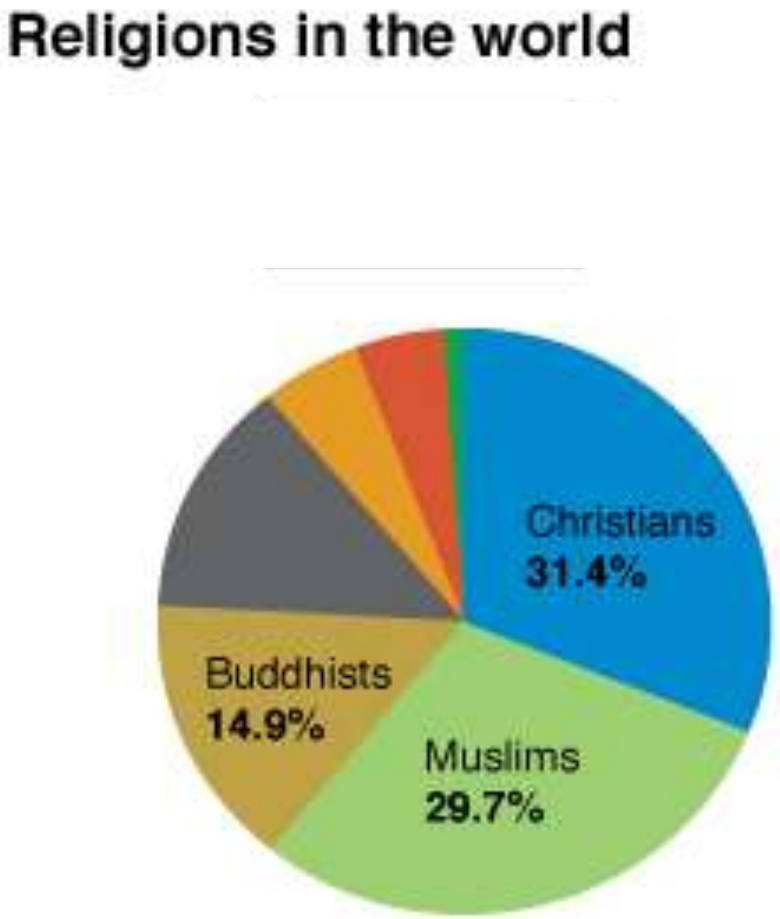


BROKEN SCALES SHOW DRAMA WHERE IT DOESN'T EXIST.



IGNORING POPULATION SIZE MAKES RATES IMPOSSIBLE TO COMPARE.





DECORATION CAN BE DECEIVING.

TECHNISCHE UNIVERSITÄT CHEMNITZ

5. Bibliography

Abrami, P. C. (2001). Improving Judgments About Teaching Effectiveness: How to Lie Without Statistics. *New Directions for Institutional Research*, *2001*(109), 97 - 102.

Abt, K. (1996). How statistics can 'lie'. *European Journal of Anaesthesiology, 13*(5), 427 - 431.

Best, J. (2005). Lies, Calculations and Constructions: Beyond "How to Lie with Statistics". *Statistical Science, 20(*3), 210 - 214.

Best, J. (2002). Thicker Than Blood: How Racial Statistics Lie. *Contemporary Sociology, 31*(5), 529.

Dekking, M. (2005). *A Modern Introduction To Probability And Statistics: Understandig Why And How.* London: Springer.

De Veaux, R. D. (2005). How to Lie with Bad Data. *Statistical Science, 20*(3), 231 - 238.

Field, A. (2012). *Discovering Statistics Using R*. Los Angeles: SAGE.

Fleming, P. J. (1986). How not to lie with statistics: the correct way to summarize benchmark results. *Communications of the ACM*, *29*(3), 218 - 221.

Huff, D. (1973). *How to Lie with Statistics*. Harmondsworth: Penguin Books.

Huff, D. (2010). *How to Lie with Statistics*. New York: W. W. Norton & Company.

TECHNISCHE UNIVERSITÄT CHEMNITZ

Bibliography

King, G. (1986). How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*, *30*(3), 666.

Starr, N. (2006). Special Section: How to Lie with Statistics Turns Fifty. *The College Mathematics Journal*, *37*(3), 244.

Steele, J. M. (2005). Darrell Huff and Fifty Years of How to Lie with Statistics. *Statistical Science*, *20*(3), 205-209.

Weiers, R. M. (2008). *Introduction To Business Statistics* (6. ed., internat. student ed.). Mason, Ohio: Thomson South-Western.

TECHNISCHE UNIVERSITÄT CHEMNITZ

“Don‘t believe statistics you didn’t fake yourself!”

TECHNISCHE UNIVERSITÄT CHEMNITZ

Sampling and Balance

- - explain sampling, balance, representativeness in studies for term papers, theses

- This course is an introduction to the use of corpora in the study of language. In modern samp**represent a sample of a particular variety of use of language(s)**. In a more general sense, the term refers to any collection of **authentic** and **naturally** occurring texts in an electronic form.

- **Sampling**: How the text chunks for each genre are selected
- **Balance**: The range of genres included in a corpus and their proportion
- A balanced corpus covers a wide range of text categories which are supposed to be representative of the language (variety) under consideration
- The proportions of different kinds of text it contains should correspond with informed and intuitive judgements
- There is no scientific measure for balance – just best estimation
- **The acceptable balance is determined by the intended use – your research questions**
- sampled (bits of text taken from multiple sources)

TECHNISCHE UNIVERSITÄT CHEMNITZ

What is representativeness?

- “A corpus is thought to be representative of the language variety it is supposed to represent if the findings based on its contents can be generalized to the said language variety.” (Leech 1991: 27)
- Representativeness refers to the extent to which a sample includes the full range of variability in a population. (cf. Biber 1993)

TECHNISCHE UNIVERSITÄT CHEMNITZ

What is representativeness?

Representativeness is a fluid concept closely related to your research questions:

- for a corpus representative of general English, a corpus of newspapers will not be enough
- for a corpus representative of newspapers, a corpus representative of *The Times* will not be enough

TECHNISCHE UNIVERSITÄT CHEMNITZ

Why should we care about representativeness?

- Reader of corpus-based studies (assessment)
  - To interpret the results of corpus research with caution
  - To considering whether the corpus data and the method used in the study was appropriate
- Corpus user (assessment)
  - Important to "know your corpus"
  - To decide whether a given corpus is appropriate for their specific research question
  - To make appropriate claims on the basis of such a corpus
- Corpus creator
  - To make your corpus as representative as possible of a language (variety)
  - To document design criteria explicitly and make the documentation available to corpus users

TECHNISCHE UNIVERSITÄT CHEMNITZ

Corpus Statistics: Terminology Frequency comparison

- In order to compare data obtained from corpora (or parts of a corpus) of different sizes, you must normalize them to a common base.
- **descriptive statistics:** statistics which do not seek to test for significance, but simply describe the data in some way (through percentages or relative frequencies)
- **raw/total frequency:** the arithmetic count of the number of a linguistic feature (a word, a structure etc.) found in a corpus/ provided by a corpus tool
- **normalized/relative frequency:** shows how often a words etc. occurs per *x* words (usually per thousand words or per million words)
- **significance tests:** Chi-square test, log-likelihood test, Fisher's exact test are used to test the statistical significance and to see how high the chances are that a result is *not* a coincidence
- **null hypothesis:** a hypothesis that you try to reject in your study
- **p-value:** is the value of probability that the null hypothesis holds, it tells you how probable it is that a and b are not correlated