

**CZECH TECHNICAL  
UNIVERSITY  
IN PRAGUE**

**MASARYK  
INSTITUTE  
OF ADVANCED  
STUDIES**



**BACHELOR'S  
THESIS**

**2020**

**JULIE  
VLACHÁ**

# **BACHELOR'S THESIS**

## **Misleading Statistics**

### **STUDY PROGRAMME**

Ekonomika a management

(Economics and Management)

### **FIELD OF STUDY**

Řízení a ekonomika průmyslového podniku

### **THESIS SUPERVISOR**

Mgr. Jana Krajčová, Ph.D., M.A.

**JULIE**

**VLACHÁ**

**2020**

## I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: Vlachá Jméno: Julie Osobní číslo: 478776  
Fakulta/ústav: Masarykův ústav vyšších studií (MÚVS)  
Zadávací katedra/ústav: Oddělení ekonomických studií  
Studijní program: Ekonomika a management  
Studijní obor: Řízení a ekonomika průmyslového podniku

## II. ÚDAJE K BAKALÁŘSKÉ PRÁCI

Název bakalářské práce:

Misleading Statistics (Zavádějící statistika)

Název bakalářské práce anglicky:

Misleading Statistics

Pokyny pro vypracování:

The aim of the thesis is to introduce the topic of misleading statistics that has become prevalent in today's media, and investigate cases in which statistics may be misleading or misinterpreted.

(Cílem této práce je představit téma zavádějící statistika, a zároveň i prozkoumat způsoby a situace, ve kterých se zavádějící statistiky vyskytují a aplikují.)

1.Introduction, 2.Theoretical part including chapters on the topics such as relative and absolute value, purposeful and selective bias, faulty polling, misleading data visualization, Simpson's paradox. 3. Applications, 4.Conclusions.

Seznam doporučené literatury:

HUFF, D. How to Lie with Statistics. W. W. Norton & Company, 1954.

ROSENTHAL, J.S. Struck by Lightning: The Curious World of Probabilities. Jaico Publishing House, 2006.

SWOBODA, H. Moderní statistika. Praha: Nakladatelství Svoboda, 1977.

HINDLS, R., HRONOVÁ, S., SEGER, J. Statistika pro ekonomy. Professional Publishing, 2004.

Jméno a pracoviště vedoucí(ho) bakalářské práce:

Mgr. Jana Krajčová, Ph.D., M.A., MUVS ČVUT v Praze, oddělení ekonomických studií

Jméno a pracoviště konzultanta(ky) bakalářské práce:


\_\_\_\_\_

Datum zadání bakalářské práce: 30.11.2019 Termín odevzdání bakalářské práce: 30.4.2020

Platnost zadání bakalářské práce: 30.9.2021

  
Podpis vedoucí(ho) práce

  
Podpis vedoucí(ho) ústavu/katedry

  
Podpis děkana(ky)

## III. PŘEVZETÍ ZADÁNÍ

27. 03. 2020

Datum převzetí zadání



Podpis studenta(ky)

VLACHÁ, Julie. *Misleading Statistics*. Praha: ČVUT 2020. Bakalářská práce. České vysoké učení technické v Praze, Masarykův ústav vyšších studií.



**MASARYKŮV ÚSTAV  
VYŠŠÍCH STUDIÍ  
ČVUT V PRAZE**

## **Affidavit**

I hereby affirm that this Bachelor's Thesis represents my own written work and that I have used no sources and aids other than those indicated. All passages quoted from publications or paraphrased from these sources are properly cited and attributed.

The thesis was not submitted in the same or in a substantially similar version, not even partially, to another examination board and was not published elsewhere.

Date:

Signature:

## **Acknowledgements**

I would like to thank my supervisor Jana Krajčová for the support and guidance while writing this bachelor's thesis, and for the inspiration to think outside the box and consider different perspectives to come to the best possible outcome.

I would also like to thank my family and friends for having endless discussions with me about this topic, which helped me to understand different outlooks on this topic and eventually inspired many of my ideas.

# **Abstract**

Misleading statistics has become a serious issue in today's age of media. General public reads and misinterprets information published on the Internet which can lead to further misunderstandings. Understanding and distinguishing the difference between low-quality media articles and actual research publications is of key importance nowadays. This thesis summarizes the basic concepts of statistics which general public is often making mistakes in. Further, it demonstrates several real-life cases which media often uses to mislead or shock the reader in order to warn against the outcomes and opinions which may appear due to the misleading information.

## **Key words**

Misleading statistics, media, data, relative and absolute values, biases, correlation and causality, misleading graphs, measures of central tendency, averages

# **Abstrakt**

Zavádějící statistika se stala závažným problémem dnešních médií. Široká veřejnost mnohdy čte a špatně interpretuje informace publikované na internetu, což může vyústit v nedorozumění. Pochopení a rozpoznávání nekvalitních zdrojů od akademického a vědeckého výzkumu je v dnešní době nezbytné. Tato práce shrnuje základní statistické koncepty, ve které široká veřejnost mnohdy chybuje. Dále poskytnu názorné příklady, ve kterých média zavádějí čtenáře, popřípadě se snaží výroky šokovat. Touto prací chci varovat před následky, které mohou nastat při nesprávném používání a interpretování statistiky.

## **Klíčová slova**

Zavádějící statistika, média, data, relativní a absolutní hodnoty, bias, korelace a kauzalita, zavádějící grafy, průměry

# Table of Content

Introduction .....	7
1. Data – How to Select Data, Ask Questions and Avoid Bias .....	12
2. Misleading Representation of Data Distribution .....	19
3. Misleading Data Visualizations.....	22
4. Misinterpreted Correlations.....	30
5. Absolute and Relative Values, Changes and Percentages .....	32
6. Practical Case Studies.....	34
6.1 COVID-19 Case Study – Several Data Representation Issues .....	35
6.2 Mistakes in Visualizations and Graphs Case Study.....	42
6.3 Measures of Central Tendency Case Study.....	48
6.4 The Case of Spurious Correlations.....	54
6.5 Case on Surveys.....	57
Conclusion .....	60
Works Cited.....	63
Appendix A.....	65
Appendix B.....	67
Table of Figures.....	69
Table of Tables.....	71





## **Introduction**

In recent years, misleading or even false statistics have widely spread especially through mass media. While neglected by many people, the accuracy and reliability of information and statistics is crucial. This is because, the use of statistics as logical/rational arguments have the power to sway public opinions. Unfortunately, social media have contributed to the rapid spread of false news, often accompanied by deceitful statistics, as any individual, without any legal or editorial supervision, can post (introduce) and share (circulate) any content. The lack of accountability and the easy accessibility of information have resulted in a dangerous surge of false information and misleading statistics. Information providers are progressively using clickbait titles to encourage users to read and circulate seemingly interesting articles. Titles and content are designed to be novel and shocking – and new and startling information tends to spread wider, deeper and faster, irrespective of its accuracy (Vosoughi, Roy, & Aral, 2018).

Statistics, undoubtedly, has become increasingly decisive and pervasive. Nowadays, every expert field (e.g. political science, agricultural science, medicine, and STEM fields) implements some form of statistical analysis. Consequently, statistics shapes the operations of our political, economic, and social spheres. It has the power to establish statements as facts and refute them as myths. But what exactly is statistics? Statistics means statistically treated data. Raw data are collected, assembled, triangulated, and formulated, oftentimes arbitrarily, and then interpreted as undisputable facts. Statistics, in its true sense, is meant to be apolitical, unbiased, and bona fide. However, the arbitrary treatment of raw data makes the end product of statistics distorted. This distorted version of statistics can then be politicized, commodified, and utilized for personal gains. Nevertheless, statistics, is systematically and increasingly been used and accepted as proxy of rational and unbiased decision-making. While statistics can be a powerful tool to implement evidence-based decision-making, when used discretionarily can result into irreversible societal damages.

For instance, one of the ways statistics aids the medical field is via identifying the determinants of diseases. Currently, one of the major topics discussed in medicine is the causal connection between vaccination and autism. While, many people adhere to the belief that early-childhood vaccination can cause autism (Wakefield, et al., 1998), medical scientists have found no support for this causation (Madsen, et al., 2002; Mrozek-Budzyn,

Kieltyka, & Majewska, 2010). The spread of unfounded beliefs is often supported by false statistics, or statistics presented in purposefully misleading way, that ultimately gets proliferated via the internet (Grinberg, Joseph, Friedland, Swire-Thompson, & Lazer, 2019). More often than not, people without any proper education in medical science comment and decide on topics that have been, for years, studied by scientists and medical professionals. For example, when information based on correlations found between two phenomena that are not related in practice get circulated via the Internet, it can not only trigger public paranoia, but also dangerously empower poorly qualified people to conclude on medical propositions. These unfounded conclusions can have critical ramifications, including loss of lives. Other spheres of socio-economic activity where statistics is widely used include weather forecasting, predicting sport outcomes, gambling, or management of firms, banks, and insurance companies. In politics too, statistics is used as an approach to push and scar political arguments, such as predicting and interpreting results of voting polls.

### *The Information Age as a Blessing and a Curse*

A critical feature of the modern world is the accessibility of information. This accessibility has been instigated by a global boom of easy access to personal computers and internet where a load of information circulates freely. However, while accessibility of information has improved people's lives, it has also made people vulnerable to inaccurate and misleading information. For example, many news providers have shifted their activities to the Internet, such that, their main revenue is gained via advertisements. Consequently, consumers are not required to pay for the information they receive. While the proliferation of providers of free information has successfully ended a monopolistic information market, the escalation of non-verified information has burdened consumers with the task of differentiating between trustworthy and unreliable information sources. Afterall, free information comes with latent costs.

Nevertheless, there are still many news providers who continue to offer information in exchange for paid subscriptions. Paid providers, often, offers higher quality and unbiased content, while allowing the readers to have an ad-free experience and other benefits such as discussion forums and editors' columns. Sooner or later, consumers are tormented with high levels of confusion and misunderstanding regarding the cost and benefits of free versus paid information. The rise of the information economy

has made it imperative for vigilant individuals to check the quality, accuracy, and reliability of the information being provided and received.

Accessibility to information is also redefining the dynamics between professionals and non-professionals. With the rapid advancement of science, technology, and medicine, it is critical for professionals to fact-check and stay updated regarding current and developing innovations. As a result, the internet has become non-expendable. Before the advent of the internet, physical books, journals, newspapers, and archival documents played an important role in the lives of researchers and academic professionals. However, in the era of the internet and digitization, information is stored and accessed through online books and journals, electronic newspapers, and digital data archives. While professionals, such as medical doctors, scientists, statisticians, and professors, because of their expert knowledge can differentiate between reliable and non-reliable information-sources, non-professionals can easily become victims of deceitful and erroneous data. Individuals without the appropriate training, tend to get their information from blogs, non-scientific websites, and public forums. While these platforms are useful to understand public deliberation, they often have hidden political and ideological agendas. For example, websites with covert anti-abortion objectives will likely highlight the negatives aspects of abortion and cite medical studies to support their claim. Similarly, websites with women-rights objectives will focus on the negative psychological ramifications experienced by women who are forced to bear children against their will. While both these websites will cite scientific studies<sup>1</sup>, the information provided will be skewed according to their political affiliations. On the other hand, websites maintained by professional associations, such as the American College of Obstetricians and Gynaecologists, will list both the positives and negatives of abortion and pregnancy-against-will.

However, the easy accessibility and effortless readability provided by public forms and non-scientific websites often persuade individuals to access their information from

---

<sup>1</sup> Even journals such as The Guardian rarely properly reference a study. In the following article, the Guardian mentions Lancet (medical peer-reviewed journal), however, does not specify the study as a source. Lancet also published the falsified study by Wakefield, et. al, 1998, which shows even peer-reviewed journals may publish articles which can be refuted.

Article from The Guardian can be retrieved from:

<https://www.theguardian.com/world/2020/mar/31/tuesday-briefing-covid-19-danger-jumps-from-middle-age>

non-scientific sources. The problem associated with the proliferation of erroneous information and misleading statistics is further exacerbated when non-professionals assume and claim that it is possible to become knowledgeable on a topic by simply reading about it on the internet. Such attitudes precisely are harmful and undermine the relationships between professionals and their clients (e.g. a doctor and a patient), as non-professionals take decision-making into their own hands. Overall, individuals should be aware of the risks associated with free information; content needs to be reviewed and checked by oneself for the quality and reliability of data and information, since, critical life-changing and society-shaping decisions are dependent on the accuracy of information gathered via the Internet.

Finally, although Facebook and other social media were designed to bring people together, in practice, they tend to polarize societies. For example, algorithms, designed by social apps, for suggesting friends, followers and advertisements often create social bubbles. These bubbles initiate social divisions where “public” opinions on socially important matters are frequently formed and strengthened. Consequently, elections may be swayed, vaccination rates may be declining, etc. Users may think of social media as connecting them to the world, however, that might not always be the case. Social media gathers user information which is later used to match similar groups of users. Even though this is mostly done for more targeted marketing purposes, it can have severe social (and other) consequences. For example, users with mutual friends, same hometown, educational institutions, workplaces, and/or complimentary beliefs and values are matched together. While user with dissimilar demographic and/or ideological backgrounds are kept unmatched. This practice of connecting similar people and distancing dissimilar people significantly contributes to the further polarization of societies.

Readers can easily be overwhelmed by the enormous amount of information offered by the internet. Therefore, reader’s caution is a must when assessing the reliability of news and statistics. Readers must be critical and aware of the subject matter, such that they can identify any misuse or misinterpretation of data and statistics. Here it is to be noted that, currently, there are no governmental rules or regulations as to what can and cannot be posted on the internet. Furthermore, social media platforms (such as Facebook, Twitter, and YouTube) and search engines (such as Google, Yahoo) do not fact-check the information they provide. Unlike newspapers or news channels, there are no

editors to verify the accuracy of the content provided and the situation is further complicated with the extensive use of algorithms that are designed to prefer certain types of posts, advertisements, websites, and articles.

In the theoretical part of this thesis, I will introduce the most problematic mistakes, intentional or not, in data presentation and their connection to the misuse of statistics. Finally, I will apply this knowledge to investigate real world cases where statistics tends to be misleading or can get misinterpreted and I will show that choosing different way of data handling and presentation may lead to different conclusion compared to the one originally presented in the media.

# **1. Data – How to Select Data, Ask Questions and Avoid Bias**

Before one starts with any statistics, it is important to obtain the appropriate data. There are many types and sources of data, as well as instruments to collect quantitative and qualitative data. When planning a research study, first, a research idea is necessary. Research idea should be informed by previous studies and ethical considerations. To conduct a research study successfully, a researcher must define the study's measure(s), without any inherent ambiguity, and examine its reliability and validity. Afterwards, it is necessary to consider how samples of research participants will be obtained from the targeted population, and which methods of sampling will be used to assure research validity. (Christensen, Johnson, & Turner, 2014)

There are different approaches to research methods. For example, in experimental research, a researcher systematically manipulates and compares a sample (i.e. treatment group) against another sample (i.e. control group). The most controlled, but the least natural, are laboratory experiments in which the researchers simulate artificially the situations which they want to analyse. Field experiments, popular among economists, implements methods of understanding human behaviour and analysing economic phenomena. By assigning random experimental groups to treatment and to control groups, it is possible to test groups' actions in naturally occurring settings. Of course, with experiments, the need for replication is important to the validity of the study. Therefore, the concept of randomization was introduced to enable equal chance of each experimental unit receiving each treatment. Quasi-experimental research design is alike to the purely experimental, however, there is an absence of one of the characteristics of experiments – randomization. The researcher still can manipulate the variables, however, not to the full extent (Geenstone & Gayer, 2007). Both are now part of the modern-day experimental design (Levitt & List, 2008).

In non-experimental research design (sometimes called observational studies), there is no space for controlling the variables. It allows for a broader range of topics to be studied and lean towards being more flexible. To date, among the most commonly used non-experimental research sampling methods belongs for example census which is a survey that samples the entire population as it attempts to collect data from every single resident no matter whether registered or not; it is perhaps the most accurate as it collects

data from almost the entire population and not only smaller samples which are meant to represent the population (Triola, 2015).

Ensuring accurate, appropriate and representative data collection is the first step for maintaining integrity of any kind of research. University of Illinois lists the following consequences of inadequately collected data in their document *Responsible Conduct in Data Management* on their website<sup>2</sup>:

- *inability to answer research questions accurately*
- *inability to repeat and validate the study*
- *distorted findings resulting in wasted resources*
- *misleading other researchers to pursue fruitless avenues of investigation*
- *compromising decisions for public policy*
- *causing harm to human participants and animal subjects*

The degree of impact may vary by the field of study, however, there is potential to cause excessive harm when the results are used to support decisions which could change public policies and principles<sup>2</sup>.

The remaining part of this chapter summarizes the most common problems connected to poor data collection techniques and points out the most typical consequences.

### *Faulty Polling*

When data is gathered through research questions (e.g. surveys), it is critical for the questions to be worded carefully. The results of a research may be misleading when questions are

1. **loaded** unintentionally by factors such as order, or
2. **leading** intentionally to evoke a desired response (Triola, 2015).

In chapter 1 of *Essentials of Statistics*, Triola points out the differences in response rates of two loaded questions from a poll conducted in Germany:

- *"Would you say that traffic contributes more or less to air pollution than industry?" (45% blamed traffic; 27% blamed industry.)*

---

<sup>2</sup>Northern Illinois University provides online study materials for students. Northern Illinois University (2005). "Data Collection". *Responsible Conduct in Data Management*. [https://ori.hhs.gov/education/products/n\\_illinois\\_u/datamanagement/dctopic.html](https://ori.hhs.gov/education/products/n_illinois_u/datamanagement/dctopic.html)



- *“Would you say that industry contributes more or less to air pollution than industry?” (24% blamed traffic; 57% blamed industry.)*

As an example of intentionally leading questions, Triola gives the actual responses of “yes” for two differently worded questions:

- *97% yes: “Should the President have the line item veto to eliminate waste?”*
- *57% yes: “Should the President have the line item veto, or not?”*

Such faulty polling may lead to inaccurate results of the research.

Asked questions should follow rules to avoid leading the responded on, or influence in another way. For applicability/suitability, it is important for the question to contain existent and accessible data, avoid using hypothetical or fictitious data, or data referring to someone else. The question needs to have clear meaning: ambiguity, vagueness of implicit meaning may lead to misleading results. The key concept has to be understood. In case the format is complicated, or there is difficulty to recall/recognize, deduct (i.e. estimate, guess), or judge, the respondent may deal with high cognitive burden. Technical difficulties such as too many key concepts or clauses, double negation, implicit assumptions may lead to incorrect data collection as well. In some case different ordering of question can also alter the results. Simultaneously, logical flaws should be avoided, while possible motivations and affections should be considered. Social norms such as unbalanced, directive, non- neutral questions, or risk of social desirability may also sway the results (Akkerboom & Dehue, 1997).

### Biases

When addressing the issue of data collection, bias becomes a major factor. “Bias is any trend or deviation from the truth in data collection, data analysis, interpretation and publication which can cause false conclusions” (Šimundić, 2013, p. 12). There are many types of biases and they occur either intentionally or unintentionally (Gardenier & Resnik, 2002). For example, sampling bias, in statistics, is the occurrence of collecting non-representative sample which then makes conclusions inaccurate for the entire population; and which may, as a result, possibly pose a threat in further data-driven decision-making regarding future policies and other societal issues (Berk, 1983).

While conducting research based on statistical analysis, there are several main types of biases which one should pay special attention to. These include:

1. **Sampling Bias:** Population consists of all individuals with a characteristic of interest. Since, due to resource constraints, studying an entire population is often not possible, researchers usually study a population through a representative sample. The underlying assumption is that the findings from the sample can be generalized to the entire population (Gardenier & Resnik, 2002). However, for this to happen, the sample need to be representative of the population. That is, the sample needs to be comprised of all the attributes of the population. However, there are numerous ways, through which sampling bias can occur, resulting in a non-representative sample. For example, during online surveys, individuals with easy access to computers and Internet are more likely to participate than individuals with restrictive access to the Internet. In other words, individuals with easy access to the Internet will be over-represented and individuals with restrictive access to the Internet underrepresented. Therefore, such sample would not be representative of the general population (Gardenier & Resnik, 2002).

**Random sampling** is where each subject in a population has an equal chance of being selected to proportionately represent the whole population. An instance of random sampling is generating random numbers to pick samples or generate random telephone numbers. Simple random sampling is also a variation of random sampling as it takes every  $n$ -th subject in a population, still following the rule that every subject of a population has the same chance of getting selected. Other possible ways to sample is to use stratified random sampling where the population is split into groups of interest and people are randomly selected from these groups so that they are represented appropriately in the overall sample. Sometimes it is too expensive or inconvenient to randomly select individuals, in which case, it is possible to choose cluster sampling; it sorts people by naturally occurring clusters such as schools or cities. Naturally, for it to work, clusters must represent the whole population and all the subjects from a cluster must be chosen (Triola, 2015). To date, census belongs to the largest non-experimental research sampling methods. It is a survey that samples the entire population as it attempts to collect data from every single resident no matter whether registered or not It is perhaps the most accurate sample as it collects data from almost the entire

population and not only smaller samples which are meant to represent the population (Triola, 2015).

1. **Selection Bias:** Ideally, research should be conducted on a random sample that best represents the population that the research is interested in. There should be no system or pattern in selecting respondents, so the sample should be random or diverse, and each respondent should have the same chance of being chosen (Winship & Mare, 1992). For example, people taking surveys on social media such as Facebook are a different group of people from those who are not using social media that frequently or, in some cases, at all. The same applies for customer service surveys where the ones who decide to partake in these surveys are usually those with very negative or, on the contrary, very positive opinions. Another source of bias is *underrepresentation* as it is difficult to get a big enough sample for minority population. In some places certain minority populations might be so small that it is barely noticeable, or possibly non-existent. Therefore, even if a subset of the minority responds, the data is still biased as the sub-set might not represent the entire minority population. Selection bias occurs if the factor(s) that systematically affect selection of people into treatment groups is correlated with factor(s) affecting the outcome of the treatment (e.g. measuring impact of polluted air in areas populated by poor people with little or no access to high quality healthcare can overestimate the total impact of pollution on health; similarly conducting medical research for monetary reward could result in doing the analysis on the poorer and by default less healthy subsample of the population, etc.). Random sampling, when possible, also helps to avoid selection bias. The main problem is that selection bias often occurs in the data, which is a result of uncontrolled, natural, experiments and thus the researcher needs to use other methods (statistical) to mitigate the bias. Or, interpret the data with great caution.
2. **Social Desirability Bias:** Bias is also often associated with questionnaires as certain questions may be biased by simply leading the respondent to give a desirable answer (Fisher, 1993). Respondents answering sensitive issues might feel inclined to answer differently, rather than truthfully, in order to be seen in a better light. This can be the case even when the survey is anonymous. Therefore, it is necessary to ask questions which are neutral to help the

respondent answer with better accuracy. Additionally, the survey should state the reason why the survey is being conducted, who is conducting the research, how the research is being conducted, and exactly how the collected data will be utilized. Moreover, the survey should be devoid of any unrelated and sensitive questions as unrelated questions can make the respondents doubtful about the researcher's intention. Sensitive questions can also make respondents uncomfortable and unwilling to provide answers. Fisher also mentions biased questions are usually not necessarily noticeable, which is why it is critical for the respondents (e.g. customers) to be aware of biased questions in surveys as the people conducting the survey may benefit from certain responses.

3. **Publication Bias:** In professional journals, the publisher may have the tendency to publish positive, or "expected", results as opposed to the negative, or unexpected, ones. On the contrary, regulatory bias focuses more on the negative results. An instance of publication bias is when a study confirming negative impact of pollution on health of people in the neighbourhood is more likely to be accepted for publication than a study not able to confirm it. Reducing these biases can be done through announcing research studies before the results are known (Geenstone & Gayer, 2007).
4. **Non-response Bias:** When respondents decide to not answer the survey questions, non-response bias might be encountered if there is a systematic difference between the people who choose to complete the survey from the ones who do not (Sheikh & Mattingly, 1981).
5. **Attrition Bias:** Due to incomplete data in social experiments (e.g. losses of participants), there are systematic differences between treatment and control groups. In social experiments, individuals are surveyed before the beginning as well as during the experiment. However, as it may last several years, the time factor creates a problem. Attrition may be caused by factors such as subjects not belonging into the population studied, experiments not being worth the payment, commuting or repeated activity, etc. (Levitt & List, 2008)
6. **Confirmation Bias:** It is related to tendency to confirm/publish the results that are generally believed to hold. The above-mentioned example of a study

confirming negative impact of pollution on health can therefore fall into this category as well. It is a type of cognitive bias (Nickerson, 1998).

Understanding and reducing said biases in research is possible through informing oneself about the potential impacts the bias may have on the study, while taking measures to avoid them. Under certain assumptions, employing proper econometric methods can help to mitigate the bias ex post.

## 2. Misleading Representation of Data Distribution

Nowadays, news media provides us with many statistical numbers and statements which are often taken out of context. Central measures are often one of those numbers. Since these measures are meant to help with representing a distribution, it is not recommended to use them as individual values when they can lead to leaving out important information and thus contribute to misinterpretation of the data.

The following is theory is a summary from chapter 3 of Triola's Essentials of Statistics as before talking about interpretation issues, it is important to understand the meaning of individual, most commonly used measures.

Measures of central tendency, often classed as summary statistics, offer a single value which attempts to represent a set of data as a middle or centre of its distribution. There are several measures of central tendency, each of which describes the centre of a distribution under different conditions. The most familiar is most likely the mean or average, however, other such as the median and the mode often have more appropriate uses.

### Mean

The mean (average) is the most popular and known measure of central tendency. Arithmetic average is equal to the sum of the values of each observation in a dataset divided by the number of observations:

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}$$
$$\bar{x} = \frac{\sum x}{n}$$

Where the  $x$  refers to the values of observation,  $n$  is the number of observations, and  $\bar{x}$  is the sample mean (smaller number of observations is  $n - 1$ ), which is important for statistics as it differentiates between samples and population, both of which have very different meanings, despite being calculated in the same way. In case of calculating the population mean, the Greek letter  $\mu$  is used instead of the  $\bar{x}$ :

$$\mu = \frac{\sum x}{n}$$

The advantages of the mean are that it can be used for continuous and discrete numeric data, it is impossible to calculate it for categorical data as the values cannot be

summed. As it takes all the values of a set, it can be affected by outliers and skewed distributions. The median is efficient in case that all data and the distribution are normal, otherwise the mean does not necessarily reflect the centre measure as it may be shifted from the centre by possible outliers.

### Median

Another measure of central tendency which is used often is the median which refers to the middle value of a distribution that is arranged in order of magnitude; it divides the distribution in two halves. In case of an odd number of observations, it is the value between two middle values. The median is less influenced than the mean by extremes and skewed data, which is why it is often preferred to mean in case the distribution is not symmetrical.

### Mode

The mode is the value with the highest frequency in a data set. While it is not used much with numerical data, it can measure data at the nominal level of measurement.

A data set can have:

1. No mode: no data value is repeated;
2. One mode: one highest value;
3. Two modes (bimodal): two data values with the same greatest frequency;
4. Several modes (multimodal): several data values with the same greatest frequency.

### Midrange

Midrange is another measure of central tendency which measures the centre between the minimum and maximum in a data set. It is defined as:

$$M = \frac{\max x + \min x}{2}$$

Where  $M$  is the midrange, and  $x$  is the data set.

Midrange is very sensitive to extreme values as it uses only the minimum and maximum. It is very rarely used; however, the calculation is very easy, and gives more insight into the different ways to define centre. It is sometimes used incorrectly as it is in practice often confused for median.

In a perfectly symmetrical distribution, all of the measures of central tendencies would be in the same place in the middle of the distribution. However, with any asymmetry, the measures move apart from each other. It is also important to choose the right sampling method for each measure as it is illogical to numerically calculate data at the nominal level of measurement.

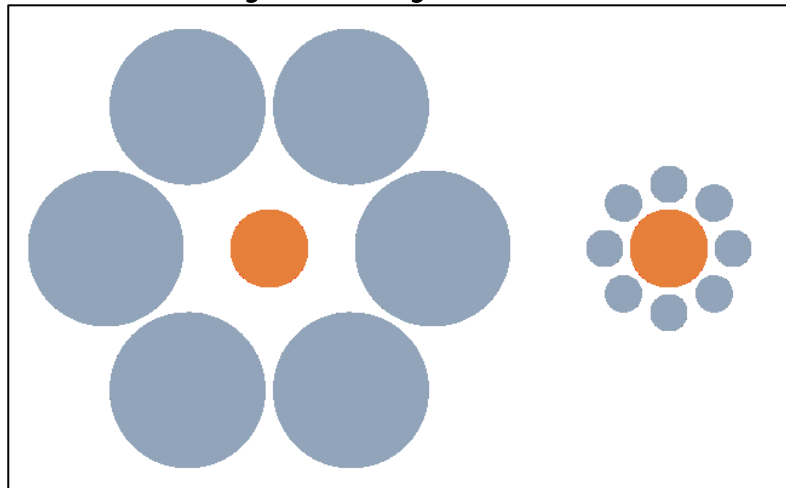
Another problem with using single-value representation of the data set is that we might lose important information. In many cases the variability of values can describe important aspects of the analysed phenomenon. Using single value is thus, in general, more easily used to lead the reader in the desired direction, especially a reader who is untrained in statistics.



### 3. Misleading Data Visualizations

Presenting data is another part of statistics which is especially popular as visualization is quick and gives the reader an instant idea of the whole picture. This is especially important nowadays with so called "snack culture." The term came from the South Korean trend<sup>3</sup>; essentially, digital consumers spend only brief periods of time on media due to short attention spans. While this trend originated in South Korea, it can apply to the majority digital consumers worldwide, which opens many opportunities for the media economically (Moura, 2011).

Figure 1: Ebbinghaus Illusion



Source: Bach, Michael<sup>4</sup>

<sup>3</sup> Hanyang Journal is an English university newspaper at the Hanyang University. Kim Geon-pyo (2015). "In Snack Culture, Possibilities of New Contents Grow". Hanyang Journal. Retrieved March 8, 2020 at <http://www.hanyangian.com/news/articleView.html?idxno=487>

<sup>4</sup> <https://michaelbach.de/ot/cog-Ebbinghaus>

**Figure 2: Simultaneous Contrast Illusion**



**Source: Carbon, Claus-Christian. (2014). Understanding human perception by human-made illusions. *Frontiers in human neuroscience*.<sup>5</sup>**

The two optical illusions (see Figures 1 and 2) prove that the situation and surroundings matter as the human mind puts information into a certain order, context, within the first few moments. Figure 1 shows the Ebbinghaus Illusion which belongs to as size-contrast illusion; the orange circles are the same size despite one being seemingly smaller than the other. Figure 2 displays simultaneous contrast optical illusion which makes one believe that the middle rectangle is the changing colour, while it is only the gradient background which creates the illusion of the middle rectangle changing colours.

Seeing how one's eyes can perceive optical illusions, which are essentially graphical visualizations as well, it can be assumed that first perception of graphs, charts and other visualizations may appear different from reality. However, through the needed knowledge and experience, the reader can learn to find any misleading data and claims much faster than readers without experience in recognizing and analysing misleading information (UNECE, 2009). That said, it is very easy to manipulate the whole picture through data visualizations as these figures may often be taken out of context and become quite misleading.

---

<sup>5</sup>[https://www.researchgate.net/publication/264866231\\_Understanding\\_human\\_perception\\_by\\_human-made\\_illusions](https://www.researchgate.net/publication/264866231_Understanding_human_perception_by_human-made_illusions)

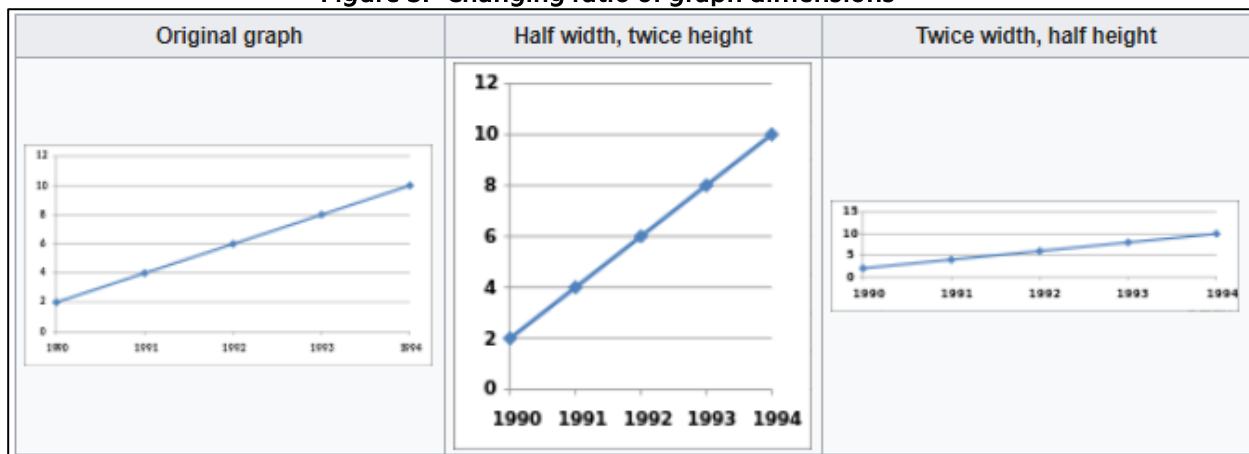
Typically occurring examples of misleading data visualizations may include:

- truncated or manipulated y-axis
- omitting zero-baseline axis
- omitting data
- wrongly chosen type of graph
- design (size, spacing, thickness, colour)
- going against conventions
- presenting too much data
- choosing a reference group that is non-representative and leads to specific conclusions

### Truncated and Manipulated Y-axis

Manipulating with the y-axis creates a disproportionately scaled graph; usually, applies to graphs with trends or timelines. By expanding the axis and fitting more numbers on it, making the data seem more condensed, and the difference smaller. The same applies for a compressed graph where the change in data may seem more significant than it is. Similarly, omitting the zero-baseline axis creates a truncated axis where starting at higher numbers distorts the picture and presents us with much more significant changes.

**Figure 3: "Changing ratio of graph dimensions"**



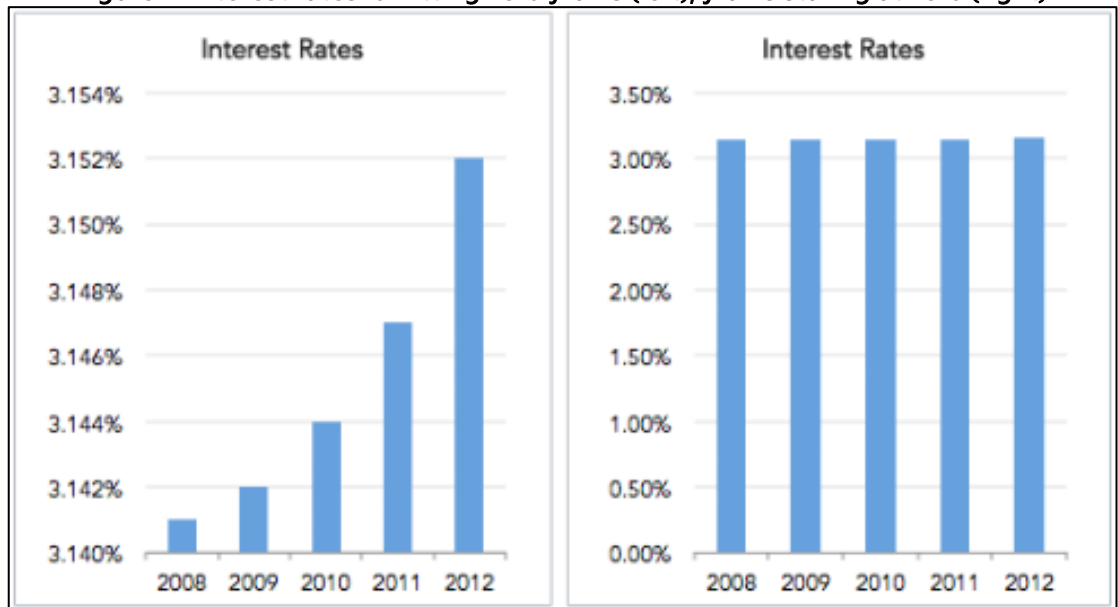
Source: Wikipedia

From Figure 3 it is visible that the spacing and density of numbers have obvious influence on the slope of the function. This way, it is possible to change the reader's view on the growth of the function.

Omission of Zero Y-axis

Figure 4 shows the impact on omitting the zero y-axis in a graph, and therefore showing distorted data which can show even the most minimal changes in a timeline. This gives the author of the distorted graph to mislead readers into thinking there is an extreme change in the interest rates over a period of time.

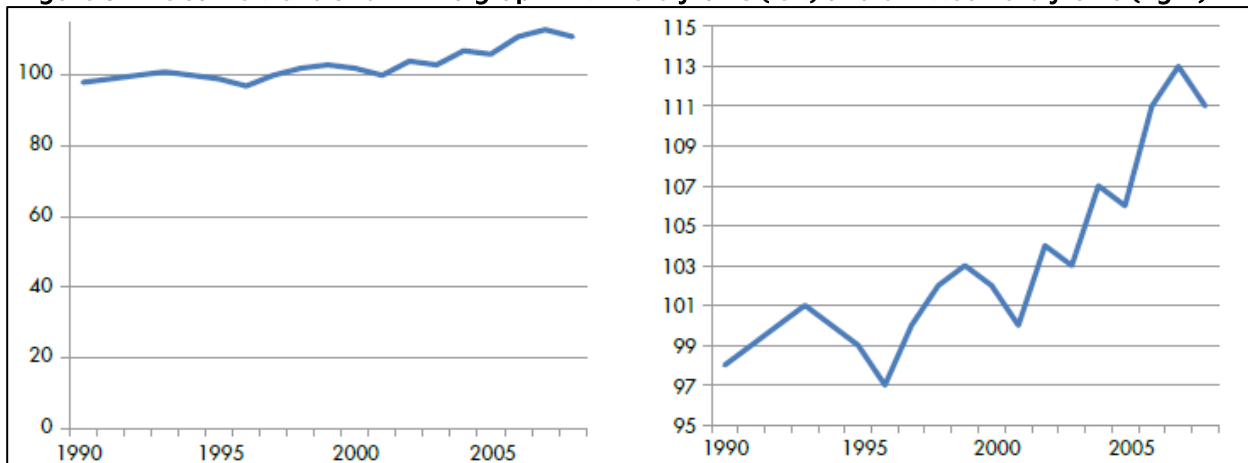
**Figure 4: Interest rates: omitting zero y-axis (left), y-axis starting at zero (right)**



Source: Heap.io

In Figure 5 the y-axis is also omitted. This time, the distorted graph on the right shows a more dramatic image of the same data compared to the graph on the left which is not distorted and shows minimal changes throughout the years.

**Figure 5: The same trend shown in a graph with zero y-axis (left) and omitted zero y-axis (right)**



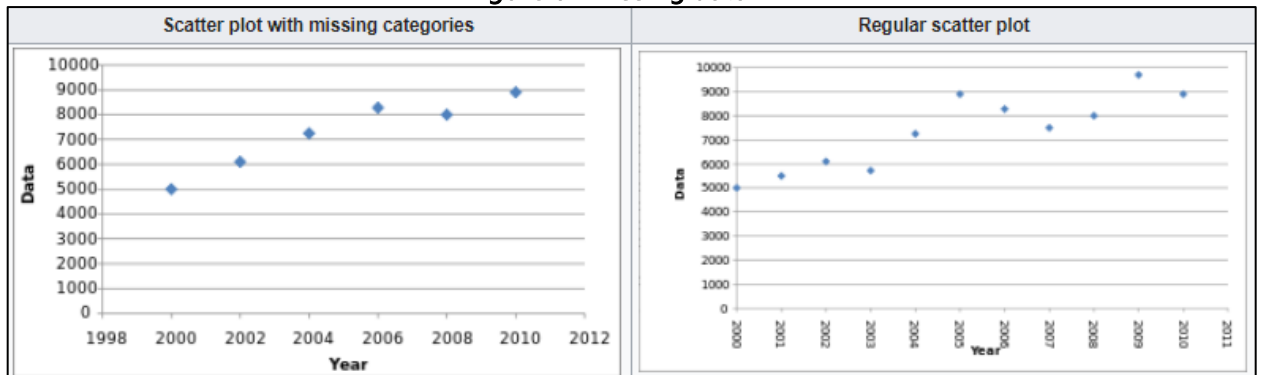
Source: UNECE

These mistakes often occur in media where the author tries to present the audience with a shocking piece of information. Graphs which are distorted this way can be useful when a researcher is trying to focus on the details, for example, whether certain values are negligible. In publications, the author provides the reader (who is often also knowledgeable in the field) with context and the whole research purpose and approach (method) with the whole process. However, it is not appropriate for media/news audience which would only get incomplete and out of context information.

### Omission of Data

Simultaneously, omitting data in a graph may result in a completely different curve of a trend. It is especially the case in seasonal data; in this case, it is important that all the months are shown. The same applies to weeks, years, and other time periods, where every period in a time must be displayed, otherwise the graph will be showing different results. For instance, a graph should show every single year in a time period instead of every two years, which might appear smoother as values might ascend or descent without fluctuation as compared to the case when all years are included.

Figure 6: "Missing data"



Source: Wikipedia

Types of Graphs Depending on Type of Data

It is important to know what type of data to work with as the type of data is necessary to pick the most efficient graph for that data category. There are several types of data which one can work with: Quantitative (numerical) data or quantities which are numbers that have order as well as consistent spacing, and qualitative (categorical) data which does not have order, nor does it have consistent spacing (Triola, 2015).

Types of data described by Triola are summarized in the following table:

Table 1: Level of Measurement

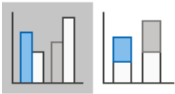





Level of Measurement	Brief Description	Example
Ratio	There is a natural zero starting point and ratios are meaningful.	Heights, lengths, distances, volumes
Interval	Differences are meaningful, but there is no natural zero starting point and ratios are meaningless.	Body temperatures in degrees Fahrenheit or Celsius
Ordinal	Data can be arranged in order, but differences either can't be found or are meaningless.	Ranks of colleges in U.S. News & World Report
Nominal	Categories only. Data cannot be arranged in order.	Eye colours.

Source: (Triola, 2015, p. 38)

Depending on the data type, it is possible to use different graphs, charts and visualizations.

With so many options regarding graphs and the ways data can be visualized nowadays, it is not surprising that some graphs might be used wrongly or are not chosen based on the important points one wants to convey.

**Table 2: Types of Graphs and Charts**

Type of graph/chart	Type of data/description	Example
Bar graph 	Categorical/qualitative data on horizontal scale, vertical scale represents frequencies, (possible to compare more data sets)	Comparison of male/female income in time
Pareto chart 	Bar graph where the bars are arranged in a descending order and show cumulative values	Measuring what contributes to happiness
Pie chart 	Categorical data as slices of circle, slices proportional to the frequency count of category, one variable	Same as Pareto chart, only depicted differently
Line graph 	Dependent data, particularly trends	Sales figures over time
Histogram/frequency polygon 	Histogram displays continuous sample data, represents a distribution. Frequency polygon is similar to histogram (type of bar graph), uses lines to show segments instead of bars	IQ scores
Pictogram	Categorical data, similar to bar graph, however, the bar is replaced by one or multiple icons/symbols	Number of people in each age category
Scatterplot 	Shows relationship between two numerical variables	Correlation
Choropleth Map	Classed and unclassified data, often popular in media, yet misused (More examples below)	Density of population in different regions

Source: Own summarization of chapter 2 from Triola's Essentials of Statistics

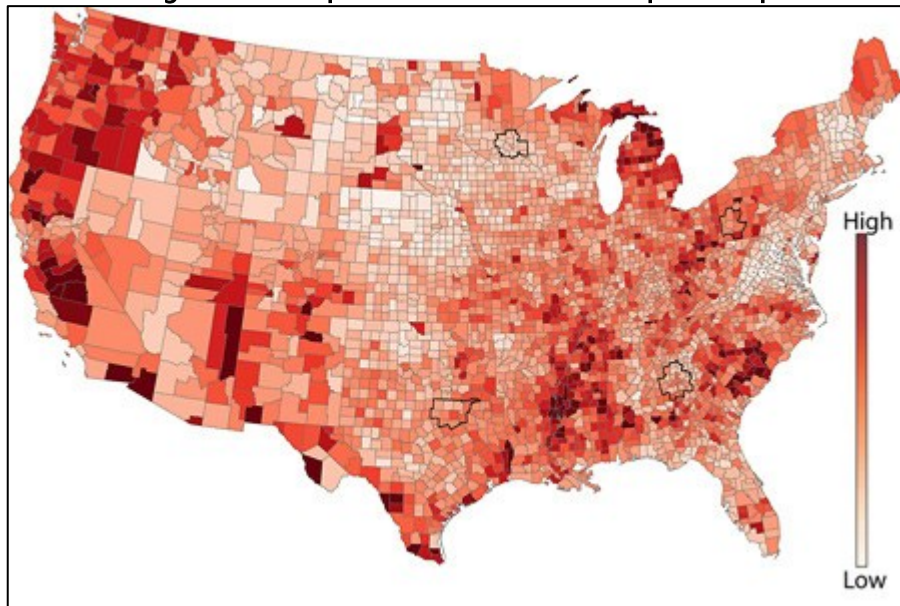
Choropleth Map

Choropleth map which represents spatial variations of quantity. It can represent density of regions or geographic areas using colour shades based on magnitude (the

darker the higher). These types of maps are usually used for relative data as comparison could be difficult with regions of different sized populations. This type of map can be used for classed and unclassed data. With classed data, the map shows more of a filtered look which shows specific data while an unclassed version of the map can be used where there is no apparent classification scheme.<sup>6</sup>

Figure 7 shows an example of a choropleth map, specifically with unclassified data.

**Figure 7: Example of an unclassified choropleth map**



Source: [axismaps.com](http://axismaps.com)

With visualizations, and graphs, it is important to define the target group, what one wants to convey, and what type of data one works with. Information should be clear, accurate and simple to understand. One should not use too many variables which can make a lot of information go unnoticed. Generally, one should not go against conventions or use distracting colours as it may only cause more confusion. (UNECE, 2009)

---

<sup>6</sup> <https://www.axismaps.com/guide/univariate/choropleth/>



## 4. Misinterpreted Correlations

*Correlation does not imply causation.* While this sentence is often spread around, many without statistical knowledge do not understand its meaning. However, it is important for individuals with no understanding of statistics and its terminology to be educated on at least the basic theory to be able to make their personal judgement on whether certain claims (e.g. in media) are misleading or not.

The following theory is a summarization of chapter 10 in Essentials of Statistics by Triola.

While correlation provides us with valuable information, it only shows the association between two variables. Causality can be stated in case there is a physical evidence which justifies the causation statement. Misleading conclusions tend to happen when there is no physical evidence to support the claim that one variable causes the other.

The strength of correlation is measured with correlation coefficient. The following formula calculates the linear correlation coefficient  $r$  for sample data:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}}$$
$$r = \frac{\sum(z_x z_y)}{n - 1}$$

Where  $x$  and  $y$  are the two variables,  $n$  refers to the number of pairs of sample data, and  $z$  is  $z$ -score for individual sample values  $x$  and  $y$ . In case  $r$  is exchanged for  $\rho$  the linear correlation coefficient is calculated for pairs of population data.

Although it is always possible to compute the lineal correlation coefficient with any collection of sample-paired quantitative data, there are requirements which should be fulfilled when using the sample data to conclude correlation in the population:

1. Sample of paired data has to be a simple random sample, which has been collected appropriately;
2. Points should follow approximate straight-line pattern which can be confirmed by a visual examination of a scatterplot.
3. Outliers may need to be possibly removed in case they are known errors.

Second and third point are attempts at simply analyse whether the pairs of data have a bivariate normal distribution. This means the distribution is made up of two independent normally distributed random variables.

The result of linear correlation coefficient can be interpreted by using the critical values of the Pearson correlation coefficient  $r$ . If the absolute value of  $r$  is higher than the critical value of the Pearson correlation coefficient, it concludes that there is a statistically significant correlation between the two variables. Otherwise, it is not possible to support the conclusion of linear correlation.

Another possibility is to use software for interpreting  $r$ . In this case, the computer usually computes the  $P$ -value from  $r$ . If the  $P$ -value (i.e. probability value) is less or equal to the significance level, the claim of linear correlation can be supported. In the other case, there is not enough statistical evidence.

Since linear correlation coefficient measures only the strength of a linear relationship, there can be another association/correlation which is not linear and therefore will not necessarily be detected with enough confidence by the above described procedure.

Another issue is, as mentioned before, that correlation does not imply causality. Causality is usually dependent on other possible variables. In some cases, lurking variables may influence the interpretation results involving correlation. Lurking variables are such which affect the variables being studied. It does not necessarily have to be included in the study or considered which may cause misleading or wrong results.

Another possible error may arise when correlation is analysed on averages. Since data based on averages can be suppressed in its individual variation, it may inflate the correlation coefficient. An example will be given in the practical case study.

## 5. Absolute and Relative Values, Changes and Percentages

The following chapter will explain the differences between absolute and relative values, changes and percentages. While these numbers are often used, many readers (especially of media) are not aware of the differences between them, which can then cause difficulties when interpreting an article or a simple argument. It is also important to decide which one of these is the best to use as it may be misleading to use percentages with very small samples, or absolute numbers when dealing with large amounts of data.

Absolute and relative values/numbers are both important types of information. Absolute value is a value that does not change when compared to another number. In other words, they are precise numbers. For example, an absolute change refers to an *amount* of increase or of decrease. Relative value is one that is compared/dependent on other numbers, they are related to absolute numbers. A relative number is one such as 2 in 10, which can also be written as a percentage (20 %) or a fraction ( $\frac{2}{10}$ ). Relative change is an amount of change divided by the original amount (both are absolute numbers), which is then usually expressed as a percent change by multiplying it by hundred.

Percentages, which are relative numbers, are useful numbers/ratios, which are a fraction of 100. While in some cases using percentages rather than absolute numbers (e.g. number of COVID-19 cases in the US vs. in the Czech Republic) can prove useful, in other cases it can be misleading. They are often used to impress a reader or considered as a simple way to argument. For example, if there are 7 respondents and out of them one replies "no", it is already 14.28 % out of all respondents. Using percentages often forgives the lack of the calculation base, or the sample size. Just as 1 out of 7 equals 14.28 %, the same result can be found by 500 out of 3,500, or 1,998 out of 13,990. In percentages, all of those have the same result as a percentage is calculated as a fraction of 100, and it can provide more insight with both absolute and relative values. Whatever the writer's intentions are when choosing percentages in case of small sample sizes, such as only seven respondents, it is recommended to use absolute numbers instead of percentages. With small numbers, the difference is much more recognizable, and a reader can make their own opinion on how significant a response of one person out of seven may be (Swoboda, 1997).

Percentage point change is another term which is often confused for percent change. As mentioned previously, percent change is the amount of change divided by

the new value multiplied by hundred. On the other side, percentage point change is simply the arithmetic difference between two percentages.

For example, when the central bank increases interest rates from 0.5% to 0.6%, it is a change by 0.1 percentage points, relatively it is an increase by 20%, which is a much more significant number. Again, a trained reader will have no problems understanding and interpreting the numbers properly. But readers with little or no statistical backgrounds can be easily be manipulated or simply misled as they might their ability to distinguish and properly interpret these numbers may be limited.

## 6. Practical Case Studies

In this chapter, I will be discussing and analysing issues that readers might, consciously or unconsciously, encounter while interpreting statistics presented via avenues that are not academic (i.e. not peer-reviewed academic journals). Examples of non-academic avenues are news reports, and political, social, and religious advertisements or personal blogs, that use statistical repertoire and visual imagery (e.g. infographics), without clearly describing how the data was statistically treated and for which purposes. As a result, readers/consumers of statistics that are found in the domain of our everyday lives, need to be aware of the quality of the content being presented to them. There exists an inexhaustible list of ways in how statistics is and could be misrepresented and misinterpreted. Here, I have considered several examples of the most misleading forms of statistics that currently circulate among the general public.

As this thesis was written during the 2020 COVID-19 pandemic, I have focused on several graphs and figures that contain information that are supposedly aimed at making the public aware of the current and future trends of the COVID-19 pandemic. These graphs and figures are recurrently shared on multiple media platforms, and as a result quickly disseminated through public discourse. The data presented in these graphs and figures include information which could be misleading or, if nothing else ambiguous, due to several reasons. For example, while making claims regarding different demographic populations, they systematically exclude providing relative numbers/percentages and instead produce absolute numbers. Simultaneously, they contain significant sampling and computing issues.

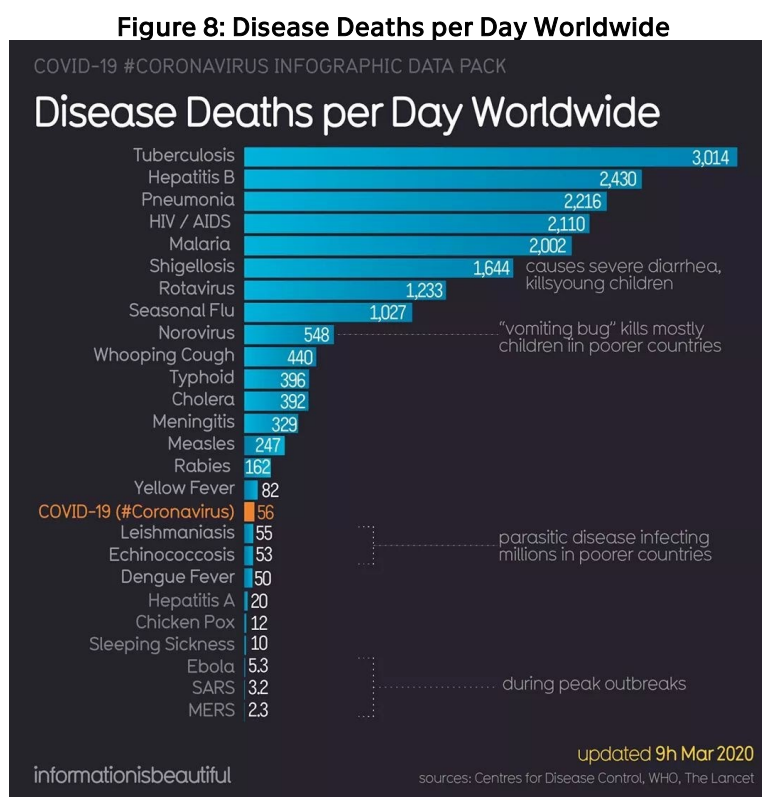
In the following subsections, I have provided several empirical evidences of statistical graphs that can exacerbate misreading of statistical information. I have discussed how easily readers can become victims of convoluted and problematic demonstration of statistics, unless they analytically scrutinize the information being given. Specifically, I have expanded upon the logical gap that lies between what the graphs actually present and what they claim to present. For example, I discuss the problems of measurement error by arguing that mean and median, although measures of central tendency, refer to two different mechanisms of measurement and therefore needs to be applied appropriately. I describe cases of correlations and demonstrate the perils of the widespread use of correlations and causations as interchangeable, because

they are two different concepts and require different statistical analysis. I also provide a real-life example of a survey instrument and establish how certain types of questions can make the survey inherently biased and, consequently, its findings unreliable.

## 6.1 COVID-19 Case Study – Several Data Representation Issues

The following graphs were posted on March 9, 2020 at 11:50 PM (figures 7 to 11) through a Facebook page titled *Information is Beautiful*<sup>7</sup>, owned by a company of the same name. The company provides visual information in the form of graphics and diagrams. In addition to Facebook, their content is available on such social media platforms such as Twitter and Instagram, and accessible through their website<sup>8</sup> and books authored by company's founder, David McCandless.

### Case I: Comparing Death Causes Per Day Worldwide (Use of Absolute Numbers, Double Counting)



<sup>7</sup> <https://www.facebook.com/informationisbeautiful/>

<sup>8</sup> <https://informationisbeautiful.net>

**Analysing Figure 8:** Pneumonia is mentioned as one of the diseases with the highest death count? per day worldwide, however, some COVID-19 patients experience pneumonia as a symptom of the disease<sup>9</sup>. Double counting can change the number of the total number of deaths per day. For example, if a COVID-19 patient had pneumonia during the time of death, it can be unclear whether the death is being categorized as been caused by COVID-19 or pneumonia. Ideally, such cases should be registered as deaths caused by multiple diseases, or at least indicate the primary disease that resulted in death. However, cautions need to be taken to ensure that one death is not counted twice or thrice, as it will change future statistical analysis to an even greater extent, especially if other factors such as wrong sampling are also present.

If trying to provide support to the hierarchy of deadly diseases, then the rate of deaths is a more appropriate parameter than absolute numbers. For example, Ebola has a much higher fatality rate (i.e. up to 90%<sup>10</sup>) than COVID-19, which is a little higher than the regular flu (current rates are estimates). However, in this particular graph, COVID-19 is ranked higher than Ebola because the number of people who got Ebola is significantly lower than the number of people who got infected by COVID-19. Another issue with this graph and comparison of data is in the difference of infection curves of each disease. In addition, Ebola was mainly spread in Africa where the healthcare system is less advanced than that of developed nations. These circumstances can contribute to the rates. In case of the infection curves, each disease has a different latency periods and if a new disease emerges, it naturally takes time before the full extent of the disease is known, which is making the comparison inaccurate.

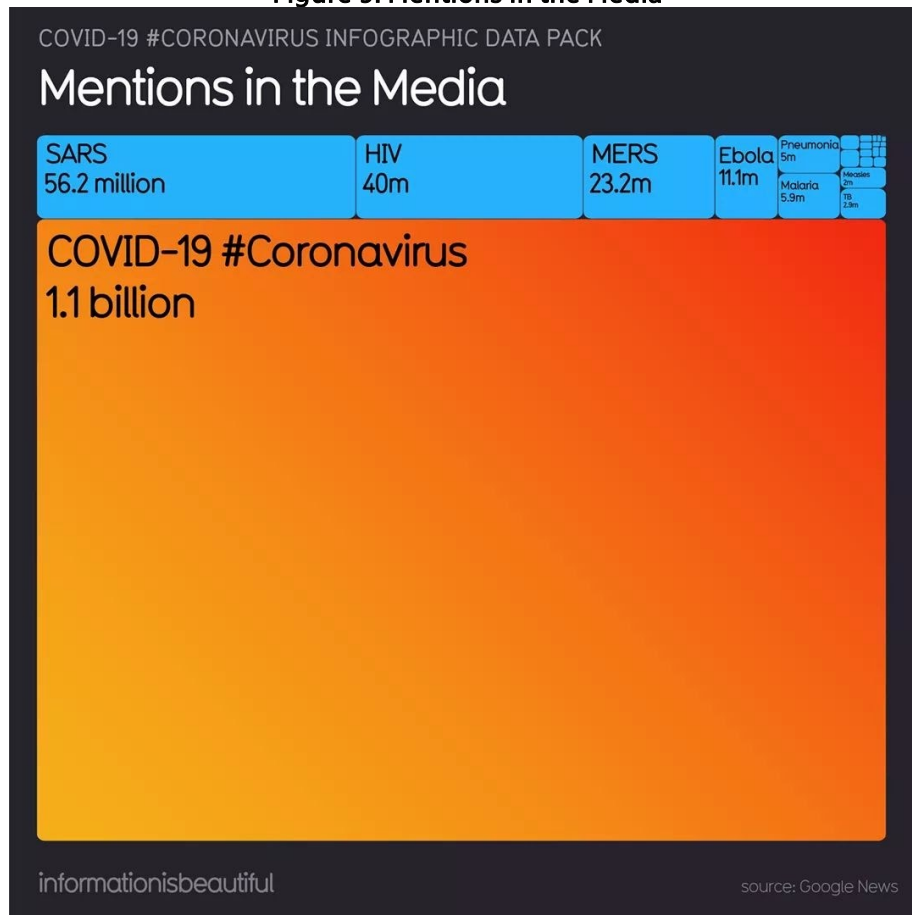
---

<sup>9</sup> <https://www.health.gov.au/news/health-alerts/novel-coronavirus-2019-ncov-health-alert/what-you-need-to-know-about-coronavirus-covid-19>

<sup>10</sup> <https://www.who.int/health-topics/ebola/>

## Case II: Mentions of COVID-19 in the Media (Unclear Sampling)

Figure 9: Mentions in the Media

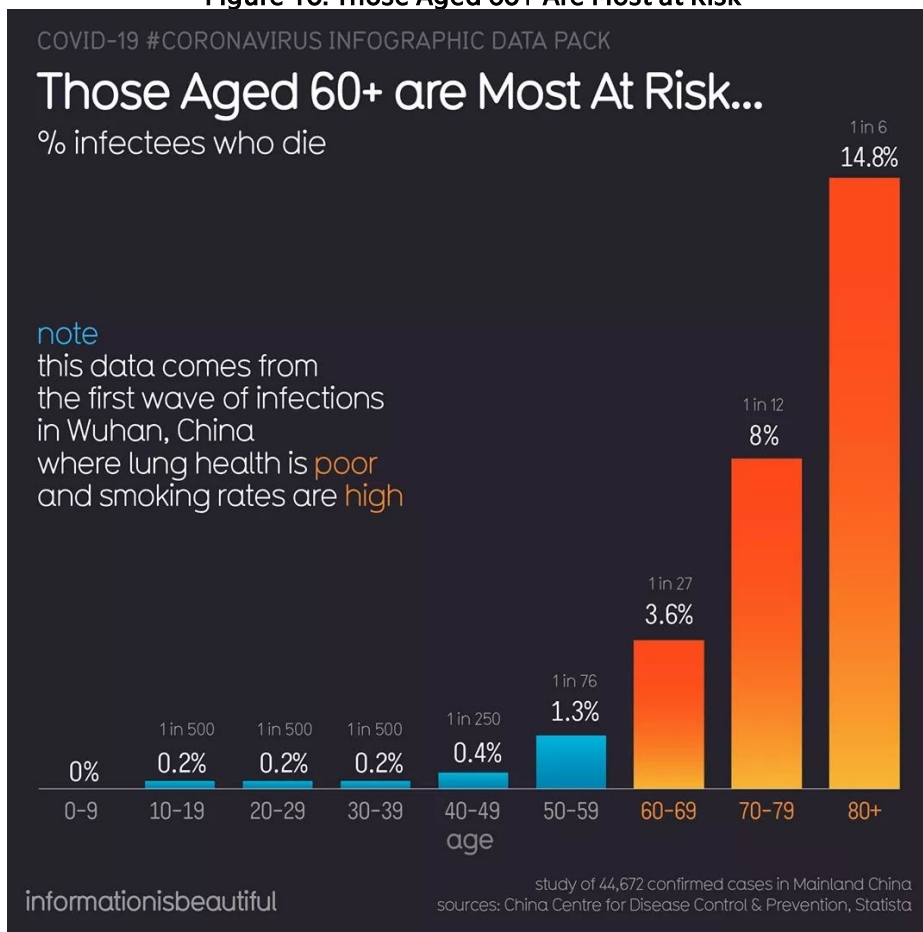


**Analysing Figure 9:** While the source is stated to be Google News, it is unclear whether by “media” is meant the number of search results on Google as the term can refer to communication means in many areas. Thus, the sampling method, despite being the key step of any data analysis, is very unclear here. With large amounts of data like these where the source was most likely word/phrase search through Google. In colloquial vocabulary, COVID-19 and coronavirus are used as synonyms, however, pathologically, SARS and MERS and a number of other less known diseases also belong to the family of coronaviruses. Therefore, simply combining the terms COVID-19 and coronavirus (as well as other terms representing the disease) together, will create serious quantification issues.



### Case III: Elders Are at Higher Risk (Non-Random Sampling, Selection Bias)

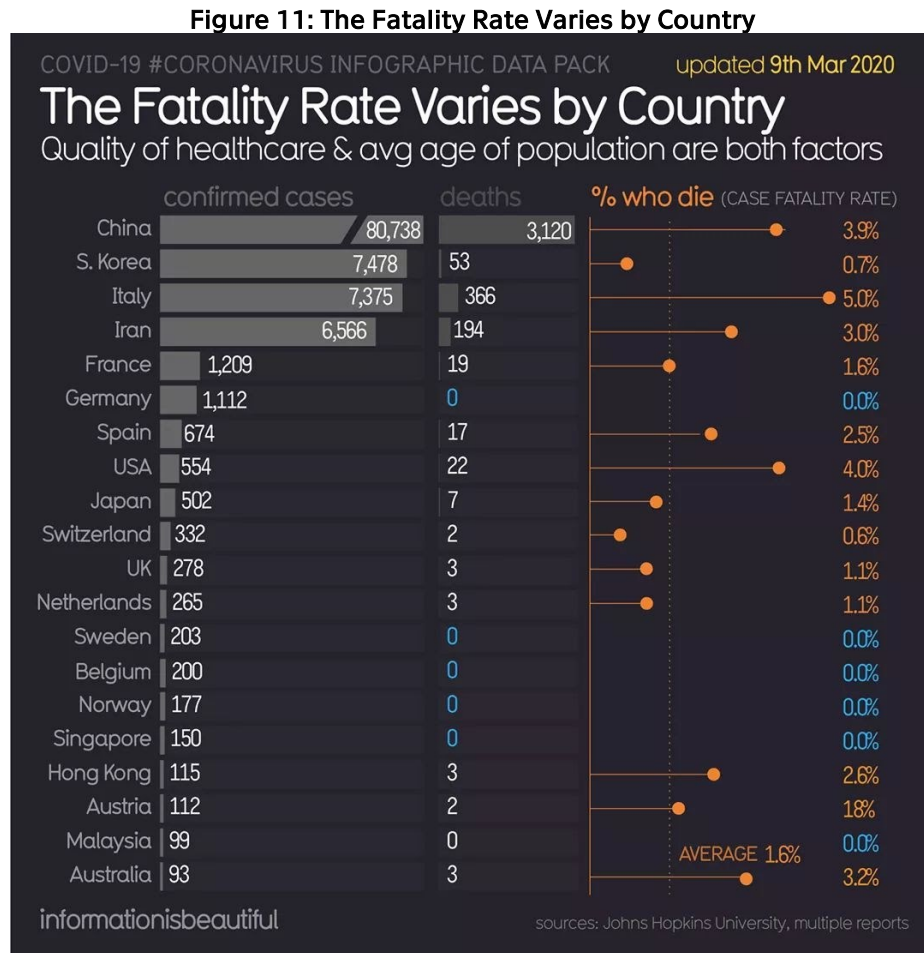
Figure 10: Those Aged 60+ Are Most at Risk



**Analysing Figure 10:** The fatality rate is given as: number of people who died divided by the number of people who got infected. Since, the number of people who got infected was computed through voluntary testing, the actual number of people who got infected could be much higher. This is a consequence of selection bias as subjects are selecting themselves for the testing. If, for example, people who have come in contact with infected people, or who are aware of not following all the precautional recommendations fully, are more likely to participate in the testing then the tested sample will over-represent infected people. And the results will underestimate the real fatality rate. Unless consistently relying on random sampling of participants, the percentages represented in this graph will differ from the real-life numbers. Plenty of factors can play a role here, which then hinders our ability to understand the disease properly. For example, young people are more reluctant to get tested than old people. Similarly, individuals in villages

or less connected areas might face difficulties in getting tested. The percentage of fatalities for the age sub-groups will most likely differ when we control for factors related to willingness to get tested and access to testing.

**Case IV: Fatality Rate by Country (Axis Break, Comparison of Different Samples and Populations in Time)**



**Analysing Figure 11:** Also, this figure suffers several problems that can lead to faulty interpretation of the data. First, Figure 11 faces a formatting issue as China has much higher number in confirmed cases than other countries, but there is an indication of an axis break, meaning the graph does not reflect the exact difference. The deaths by country then appear unproportionate to the graph of confirmed cases.

Second, the outbreak in China started in December 2019, while the outbreak in countries beyond China started around February 2020. Therefore, on March 9th, 2020 (i.e.

when the numbers presented got computed), China was almost at the final stages of the outbreak, whereas other countries were at the initial stages of the outbreak. Therefore, the timeline (infection curves) for the countries represented do not match. The graph would make more sense if the fatality rate was presented according to stages of the outbreak process.

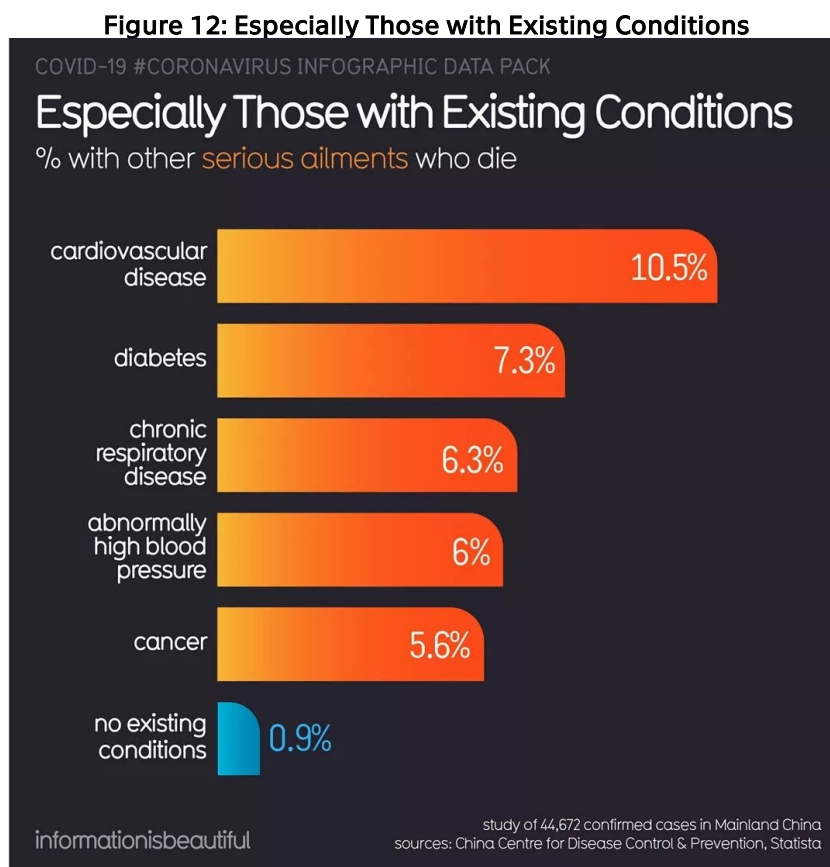
Third, the totals per country may depend on availability of testing in different countries. For example, countries which do not test much may experience higher fatality rates as the confirmed cases may be only the most serious cases. Patients in countries where healthcare and health facilities are not easily accessible might have only a few confirmed cases, or none at all. In this instance, it is possible that only wealthy people can afford the testing or treatment which changes the overall numbers in the equation. The age structure of each country is also important as elderly may experience symptoms more intensely, possibly influencing the fatality rate yet again (e.g. Italy). On the contrary, countries which have higher number of people tested such as people who have come into contact with the disease or experience symptoms, however, do not necessarily need to carry the disease, can have a better representation in the data. Random testing is not always possible (e.g. a nationwide lack of COVID-19 tests), in which case it is necessary to approach a different system of selection so that each available test is used as efficiently as possible in order to get a precise estimate of infection rate in the country.

Fourth, comparing countries of different sizes with different population density could pose a problem as the spread of the virus could vary depending on the number of first sources of the virus in each country. Other influential factors could possibly include cultural habits such as the size of social circles or frequency of attending gatherings and meetings, which could also influence the spread.

Even for the least sophisticated but informative comparison, it is crucial to consider that in case of different sized countries with a different population density, one should consider turning absolute numbers into relative. Of course, it is possible that two countries with the same population density but different sized areas may show different values, however, only after considering further aforementioned factors such as density, cultural habits, and healthcare, it is possible to reach answers to how fast the disease is spreading in different environments, to what extent it is deadly in a controlled and uncontrolled setting (e.g. healthcare, government recommendations, public awareness), etc. Simple comparison of absolute numbers of cases provides very little and likely

misleading information about how the severity of pandemics in countries with very diverse cultural, legal and demographic conditions, differently equipped healthcare systems, or very different approaches to sampling and testing procedures.

### Case V: Risk for Those with Existing Conditions (Double Counting)



**Analysing Figure 12:** The serious ailments sub-groups mentioned in the graph tend to be related. For example, patients with cardiovascular disease tend to also have chronic respiratory disease. Similarly, patients with diabetes are likely to have high blood pressure. The graph does not mention how the sample sizes for the individual sub-groups were counted. For example, if a patient had more than one of the diseases, was he/she counted multiple times? Ideally, the graph should have subgroups showing cases of single disease and subgroups depicting cases with multiple diseases. For example, cardiovascular and chronic respiratory disease, cancer and high blood pressure, and so on. These “risk factors” can have different impact for different age groups as well.

While spreading panic through the public (e.g. in the form of misleading statistics) is not ideal due to the economy being negatively affected (e.g. by changing consumer decisions), awareness is still needed to stimulate responsible behaviour.

## 6.2 Mistakes in Visualizations and Graphs Case Study

### Case VI: Confirmed COVID-19 Cases Reported by *Novinky.cz* (Base Neglected, Colouring Conventions Ignored)

For the first case in visualizations, I have chosen two choropleth maps from the same article on the news website *Novinky.com* where the following visualizations were updated each day.

Figure 13: Confirmed Cases of COVID-19 in Czech Republic, March 12, 2020

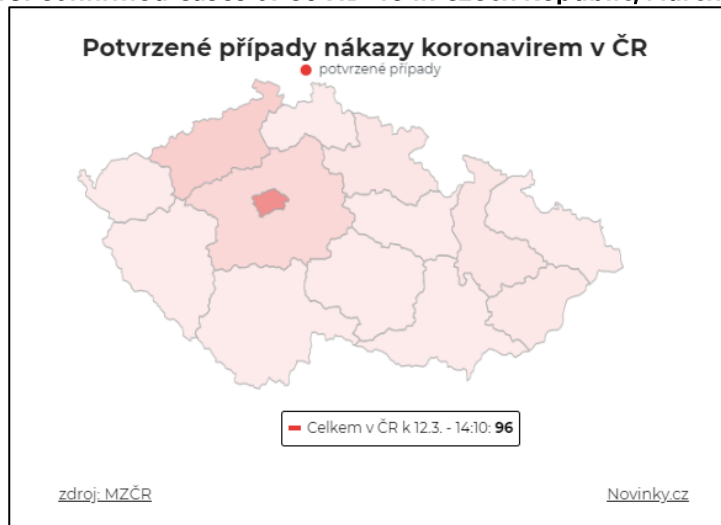
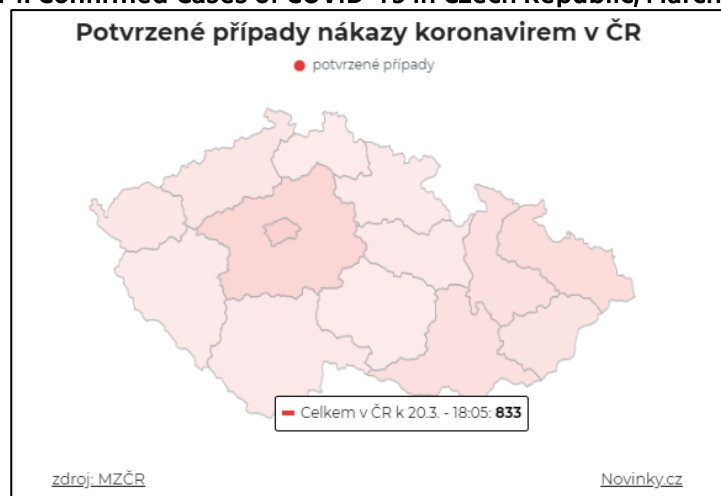


Figure 14: Confirmed Cases of COVID-19 in Czech Republic, March 20, 2020



**Analysing Figures 13 and 14:** The visualizations<sup>11</sup> do not follow conventions - the most affected region is usually indicated as bright red. Moreover, there is no description of minimum and maximum values of each shade categories. While the map on the website, shows number of cases in each region when hovering over the region with the cursor, it is not apparent at first glance. Considering that graphs, charts and visualizations should mainly give us majority of information within one look, it is not really possible to compare each region. Despite the visualizations being from the same article, only updated, if they should be compared the each other to see how each region is affected, it would appear that Prague has managed to lower the cases of infected people. To make the graph less misleading, the region with most cases should be bright red (dark red, depending on the colour scale chosen), while the regions with no cases would be coloured white. Whether the graph is meant to show classed or unclassed data, it should also include a key to describe the colour scheme for specific data.

This particular type of visualization might be even more problematic if each region has different population density (which is the case here). For example, having one confirmed case among 500 people is of course better than having one confirmed case per a hundred. Therefore, calculating relative values for each region to later apply for the shades would give the reader more insight into the severity in each region.

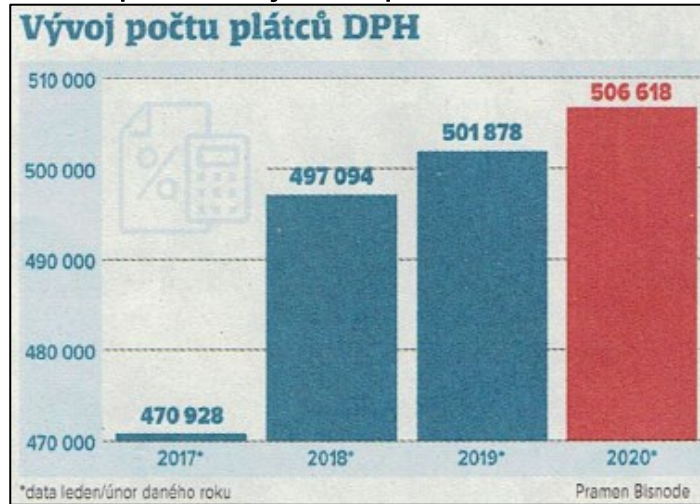
---

<sup>11</sup> Retrieved March 12, and March 20, 2020 from <https://www.novinky.cz/zahranicni/koronavirus/clanek/hamacek-vlada-by-mela-rozhodnout-o-vyhlaseni-nouzoveho-stavu-40316389>

### Case VII: E15's Graph of the Day (Omitted Zero Y-axis)

For the second case of misleading visualizations, I have been analysing a graph of the day from a daily newspaper E15, where each day a graph is presented.

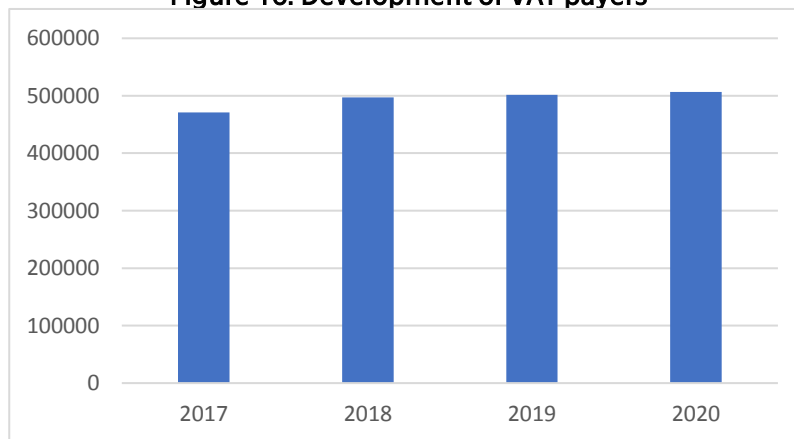
Figure 15: Graph of the day: Development in Number of VAT Payers



Source: Deník E15, Tuesday February 18, 2020, issue 3030

**Analysing Figure 15:** The graph displays development in number of payers of VAT, however, the y-axis is not starting in zero which is misleading as it causes a distortion as the graph is showing a trend. From the data given in the graph, I have redone the graph to start at the zero y-axis. From the newly made graph (Figure 16), it can be seen that it is not ascending steeply as the previous graph (Figure 15) makes it seem.

Figure 16: Development of VAT payers

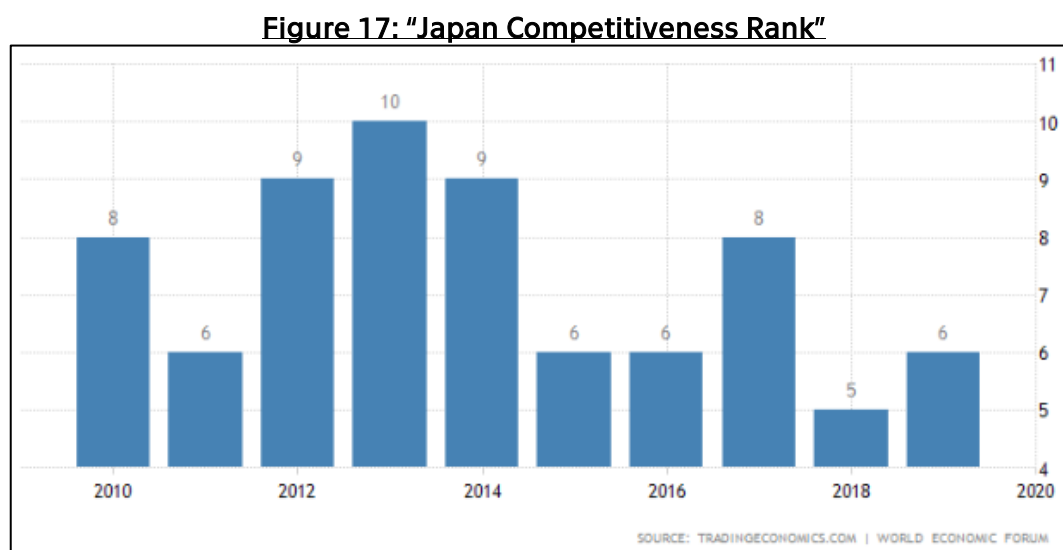


In Figure 15, the writer's aim, whether intentional or not, could have been to shock the reader with the sudden jump in VAT-payers between the years 2017 and 2018, which

slowed down in the following years. On the contrary, the reader's first reaction could be seeing the bars in the graph. One does not necessarily have to notice the high numbers either. Even then it is difficult for the reader to compute the proportions of those numbers on the actual scale which can be seen in Figure 16.

**Case VIII: Trading Economics' Competitiveness Ranks and Indexes (Omitted Zero Y-axis, Wrongly Chosen Graph, Missing Information)**

The next graphs taken from the website Trading Economics<sup>12</sup> (figures 17 and 18) show the importance of choosing the right graph and how important it is that the data used are converted into a form that can be comparable.



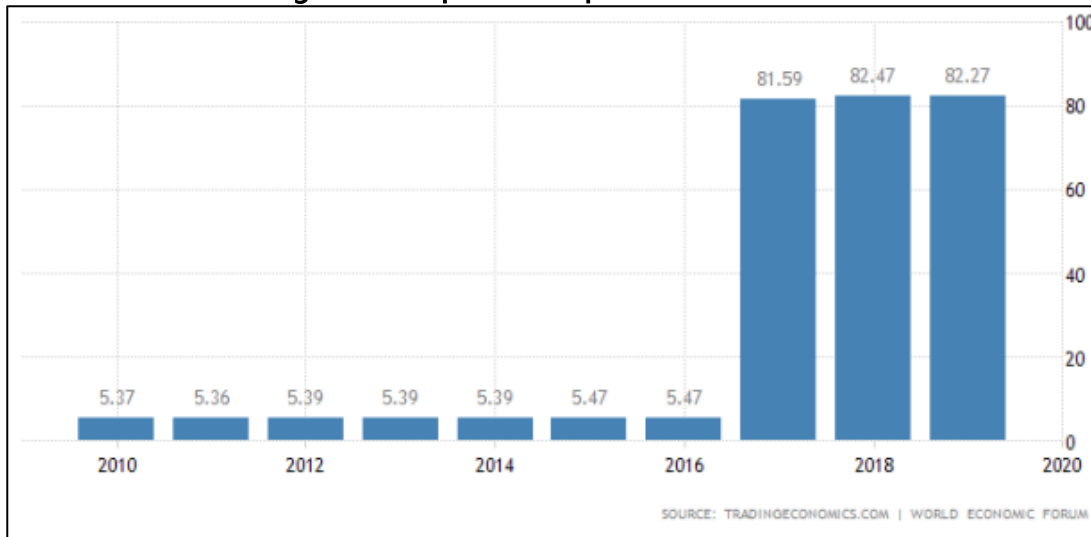
**Analysing Figure 17:** The graph is supposed to show the rank of a country through a bar graph. The type of graph chosen is not describing the type of data used accurately. The rank would be better displayed by simply writing the number. If the author would want to compare the competitiveness index that the rank is based on, it would be much better to use the index in a bar graph or to show a relative international position of Japan among other countries. Another mistake found in this graph is that these data described in time do not have a zero y-axis – that is, however, unimportant as the graph that has been chosen is not accurate.

---

<sup>12</sup> <https://tradingeconomics.com>



**Figure 18: Japan's Competitiveness Index**



**Analysing Figure 18:** Figure 18 would be the ideal for comparing how a country (in this case Japan) did in competitiveness (measuring macroeconomic and micro/business aspects) over time. However, a new index system was introduced in the year 2018, changing the scale from 1-7 to 1-100. This was not considered in Figure 18, which is a mistake which could lead to confusion from the reader's side if one did not know about the changes in the scales. The change was not mentioned on the same page, although it was mentioned on the page which contained Figure 17. Not every reader does necessarily search through other pages on the same website, which is why this mistake could lead to confusion on the reader's side. The information about a change in scales should be included under the graph, or ideally in the graph itself.

Simultaneously, to an untrained reader interested in economics, a website such as Trading Economics may seem as an accurate and professional source. That itself may be problematic, however when one searches for competitiveness rank or index on Google search, Trading Economics is one of the first sources found. While this is not necessarily either side's fault, there should be more caution on the side of the writer as well the side of the reader. Since the graphs have most likely been generated by a computer, as there are many mistakes in them, some which can be detected easily just by noticing the values. The reader should pay attention to these. However, the writer/editor of the website should be aware of mistakes like these and attempt correcting them. The reader often also does not realize that there can be lack of editors for some websites, and therefore, lack of accurate and checked information.

A particular aspect relates to the way social media (such as Google or Facebook) rank sources. It often leads to readers primarily finding sources which, rather than containing professional analysis, feature expected or attractive outcomes. Moreover, the Internet is full of advice how to increase “the visibility of your webpage” in Google or other search engines.<sup>13</sup> This means there are ways, unrelated to accuracy or relevancy of the contents, how a skilled and devoted web-site administrator can move the web site up on the list of search results.

---

<sup>13</sup> Various websites provide advice, for example <https://www.stylefactoryproductions.com/blog/six-simple-ways-to-make-your-site-more-visible-in-google-search-results> or <https://auxiliumtechnology.com/10-ways-improve-website-visibility-search-engines/>.

### 6.3 Measures of Central Tendency Case Study

#### Case IX: Regional Differences in Wages (Use of Averages)

In this sub-section, I will be inspecting an article from *Seznam.cz*, which is one of the most popular search engines and news providers in Czech Republic. The article is titled "Average Czech earns 33,697 crowns. An improvement by two thousand within a year"<sup>14</sup>. The article makes several claims, which after further inspection, are at best misleading or simply give the reader an incorrect representation of the social reality.

Figure 19: Gross Average Wage in Regions for 3Q 2019, Annual Growth



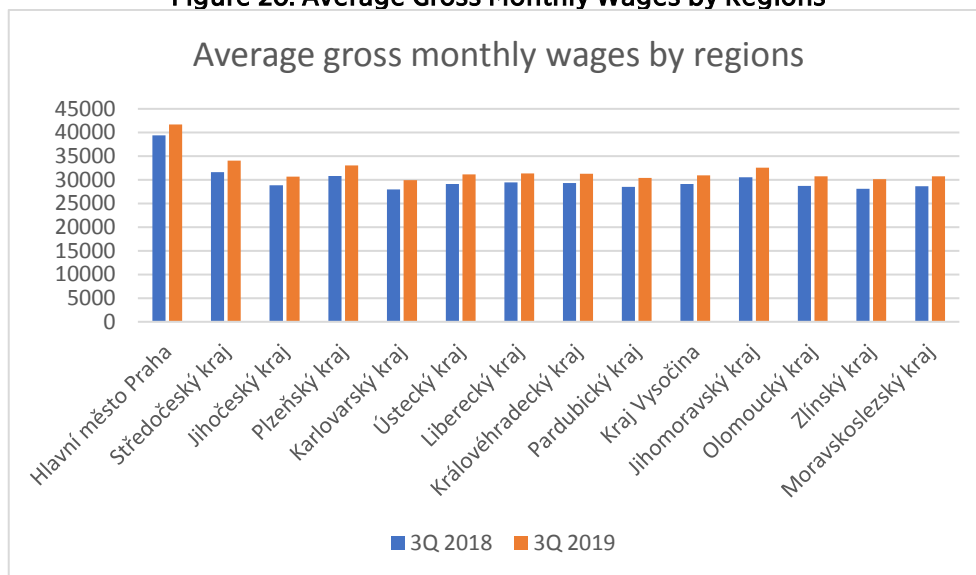
**Analysing Figure 18:** At the beginning of the article, a graph of gross monthly average wages by region in 3rd quarter of 2019 (see Figure 18 below) is shown; it is apparent, the purpose is to catch attention. Rather than focusing on comparing all the regions, the graphs emphasizes the highest and the lowest wage regions by highlighting them in different bright colours (in this case, red and green). The reader will most likely be only interested in these values, disregarding the rest of information.

Extracting the data from Figure 18, I sorted the region according to the standardized order and calculated the values for year 2018. I recalculated the average, median and other values, and found that the results do not match the average stated in

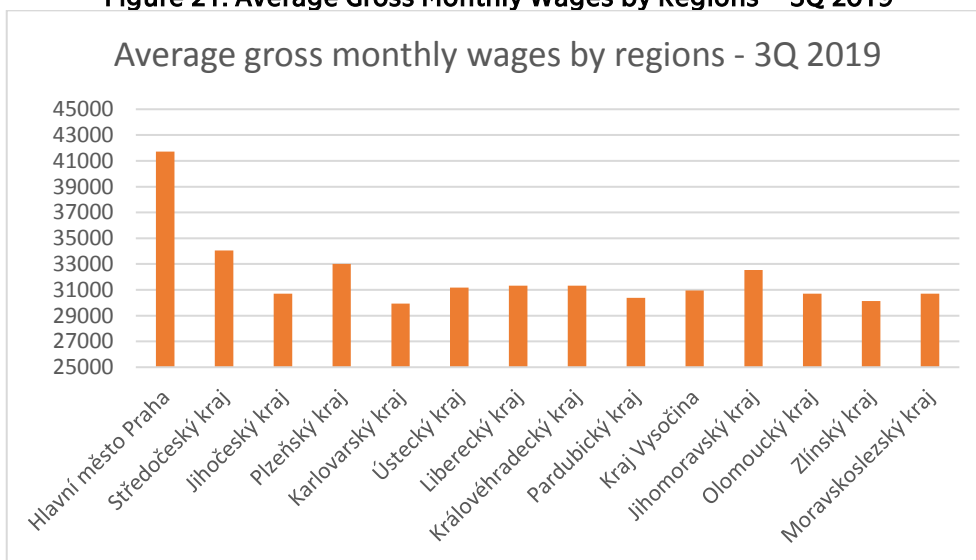
<sup>14</sup> See Appendix A

the title as my calculations were CZK32,048, a different value by approximately CZK1,500. After further search, I discovered another source which indicated that it is the nominal average which was not necessarily clear in the beginning of the analysed article. This demonstrates how important it is to distinguish between terms in statistics, and how certain values may vary depending on the used method of calculations. Consequently, I have created two graphs (Figures 19 and 20) which show the average gross monthly wages by regions.

**Figure 20: Average Gross Monthly Wages by Regions**



**Figure 21: Average Gross Monthly Wages by Regions – 3Q 2019**



It is quite evident that the Prague region stands out with its values, however, the variability of values between the other regions is much smaller. Even the Karlovy Vary region (i.e. Karlovarský kraj), does not stand out as much as can be inferred from the map in figure 19. This could be possibly because it is the only one below the 30 thousand CZK mark, which again gives the reader an illusion that it is much poorer than even the other regions. While the numbers in the discussed article may have been calculated by rules of statistics, it is important to remember that the interpretation may vary greatly and, in themselves, the averages have little meaning. Therefore, I will try to provide some context.

### Labour Structure

Table 3 shows that managers and professionals earn considerably higher wages. As a capital city, Prague accommodates a higher number of economics subjects in the area of banking and insurance, and professional, scientific and technical activities<sup>15</sup>. Therefore, it could be reasonable to assume that Prague will have more higher paid workers. Accordingly, the average wages will be higher. Another factor could be living expenses which are considerably higher than in other regions.<sup>16</sup> It is therefore likely that if we statistically clean the data of important fundamental regional differences, the resulting wage differentials would diminish.<sup>17</sup>

---

<sup>15</sup> <https://www.ispv.cz/cz/Vysledky-setreni/Aktualni.aspx>

<sup>16</sup> <https://www.czso.cz/csu/xa/zivotni-podminky-prazskych-domacnosti-v-roce-2018>

<sup>17</sup> The extent of such exercise goes beyond the illustrative purposes of this example and exceeds the goals of this thesis.

**Table 3: Average Wages by Occupation by CZ-ISCO-08 Major Group**

2018 <sup>1)</sup>			Occupation
Total	Males	Females	
<b>33,684</b>	<b>37,008</b>	<b>29,627</b>	<b>Total</b>
69,044	75,100	55,529	Managers
47,382	54,688	40,586	Professionals
36,938	40,627	33,161	Technicians and associate professionals
27,793	32,499	26,331	Clerical support workers
23,069	24,971	21,940	Service and sales workers
23,131	24,799	21,701	Skilled agricultural, forestry and fishery workers
28,743	29,533	23,620	Craft and related trades workers
28,037	29,245	25,063	Plant and machine operators and assemblers
19,841	22,325	17,950	Elementary occupations
39,428	39,180	41,423	Armed forces occupations

in CZK

<sup>1)</sup> Preliminary data.

**Source: Ministry of Labour and Social Affairs**

### Employment Pull Factor

In the *Seznam* article, it is also stated that “differences between regions through the whole Czech Republic are high. An average employee in Prague earns CZK11,779 higher wage than an average employee in the Karlové Vary region. This situation leads to the migration of Czechs from the poorer regions to Prague.”

Stating that the situation is leading to migration from poorer regions to Prague could be possibly a misleading statement. While one could easily assume that money is the main motivator, there are other variables that need to be considered. Those can be job opportunities in the desired field, living costs, etc.

Prague has the lowest unemployment rate (1.9% in 2019<sup>18</sup>) out of all the regions. As a result, employers could be compelled to compete for skilled workers which would ultimately drive up the wages. The ratio of skilled workers to specialized jobs available is possibly reversed in regions outside Prague. For example, in some of the poorer regions, there might be higher numbers of skilled workers than there are jobs. Therefore, along with high wages, the availability of employment for skilled workers could drive the migration to Prague (or other big cities). Then I would argue that workers in the glasswork

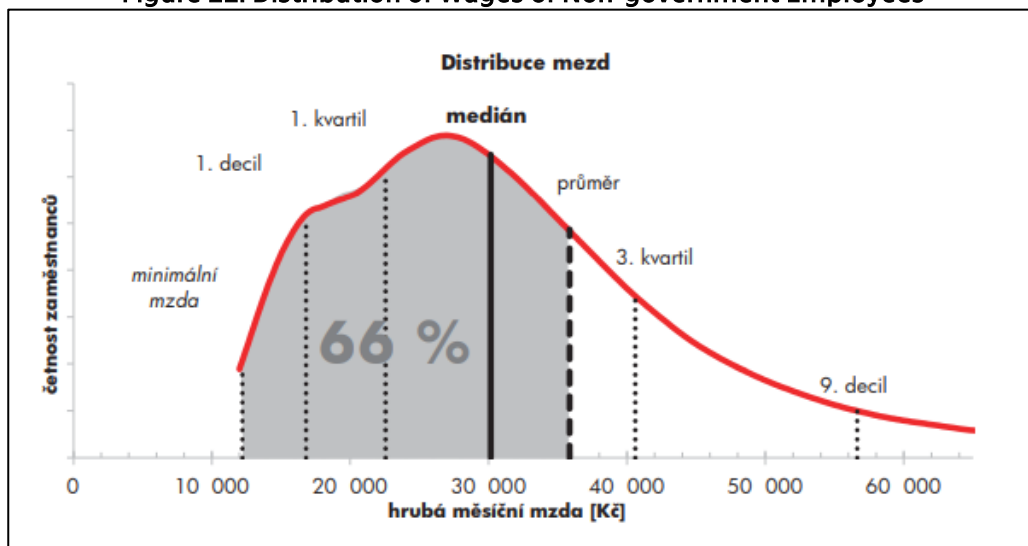
<sup>18</sup> Data from Ministry of Labour and Social Affairs

industry, which is focused mainly in Ústecký kraj and Liberecký kraj<sup>19</sup>, would have higher chances finding a job in those regions, possibly, earning higher wages as well. Since Prague and some other regions lack employment opportunities for glass artistry, glass workers are more likely to migrate to the aforementioned regions rather than Prague.

### **Case X: Demonstrating Mean and Median in Distributions**

The graphs (see Figures 22 and 23) show the current distribution of wages of non-government and government employees. The graphs show the amount of symmetry in each distribution as well as how it influences the mean and median. In the distribution of government employees where the distribution is more symmetrical, most likely due to the system the wages are divided. In case of Figure 22, the distribution is asymmetrical, partially due to the minimal wage (much more than in Figure 23).

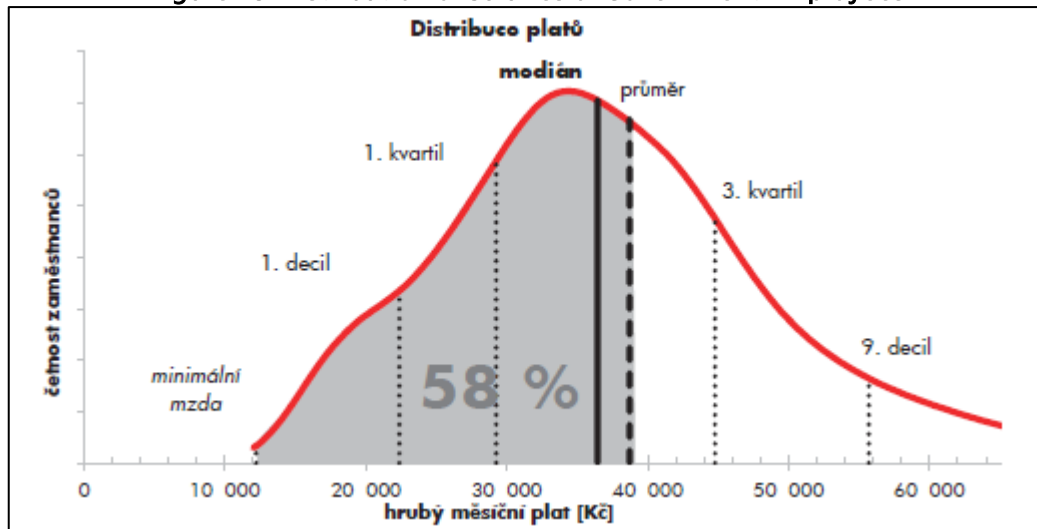
**Figure 22: Distribution of Wages of Non-government Employees**



Source: ISPV, sample size: 3,071,000 employees

<sup>19</sup> <https://askpccr.cz/o-skle/cesky-sklarsky-prumysl>

Figure 23: Distribution of Salaries of Government Employees



Source: ISPV, sample size: 647,200 employees

These graphs also show the difference between mean and median in distributions, and how median is much closer to the highest frequency than the mean. While one is not necessarily more important than the other, a reader has to know in what context these values are being used.

In particular, non-professionals are normally not aware of the difference between symmetrical and skewed distributions and the reasons for the skew (e.g. wages not being negative or the impact of minimum wage legislation). An increase in the minimum wage thus increases the difference between median and mean, which most people incorrectly consider to be a signal of growing wage inequality.

Another example to demonstrating the difficulties with working with averages are moving averages. A company produces two types of products A and B. Its production can be reported on a daily basis, or as a 4-day moving average ( $A_{ma}$ ,  $B_{ma}$ ). From the following table (Table 3), I have calculated the correlation coefficient for the two variables A and B, and the variables  $A_{ma}$  and  $B_{ma}$ .



**Table 4: Daily production and 4-day moving average**

A	A <sub>ma</sub>	B	B <sub>ma</sub>
5	-	9	-
8	-	7	-
6	-	10	-
9	7	8	8.5
7	7.5	11	9
10	8	9	9.5
8	8.5	12	10
11	9	10	10.5

The correlation coefficients for each set of variables was as follows:

$$\text{for } A, B: r = -0.089087081$$

$$\text{for } A_{ma}, B_{ma}: r = 1$$

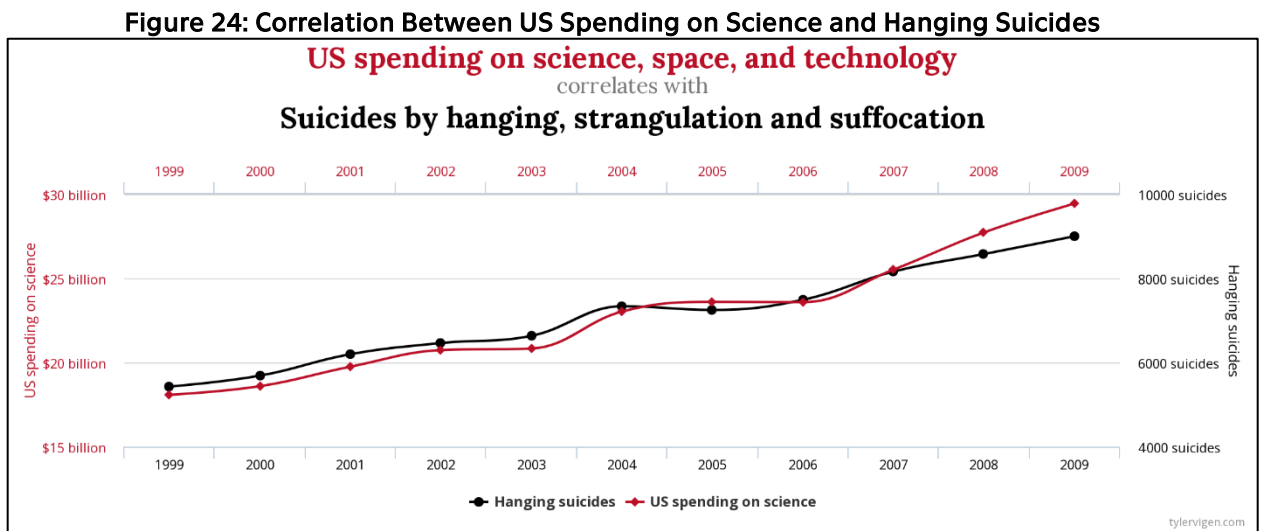
The low coefficient in case of variables A and B means that for product B, the leftover capacity from lower production of product A will be used (or vice versa) as there is no correlation. If production is described with moving averages (for an evident trend), the correlation coefficient shows a strong correlation.

## **6.4 The Case of Spurious Correlations**

This sub-section will focus on the relation, not the lack of relation, between correlation and causation. As mentioned before in chapter 5, correlation does not equal to causation. Whenever discussing if one variable causes the other, it is important, yet not enough, to establish correlation. Statistically, correlation can be established through the Pearson product-moment correlation coefficient, among others. However, to establish causation, further analysis which goes beyond statistics is required. Additionally, the correlation should be logical and intuitive so that one variable can be considered a cause of the other.

### Case I: Spurious Correlations by Tyler Vigen

As an example, for this case, I have chosen Spurious Correlations by Tyler Vigen who presents several correlations on his website<sup>20</sup>. The aim of the website is to show how a selection of unrelated data sets which correlate can deceive. Merriam-Webster dictionary defines spurious as *"of deceitful nature or quality"*. One who does not understand the difference between correlation and causation may easily misinterpret these graphs, thinking there could be causation due to their lack of understanding of statistics.



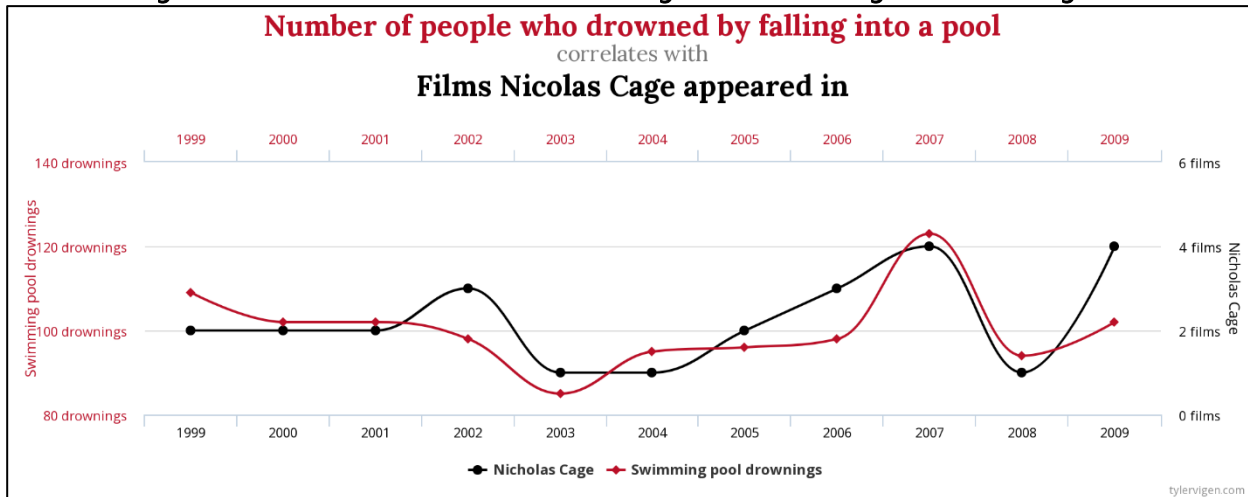
Source: Tyler Vigen

**Analysing Figure 23:** Figure 23 has a strong positive relationship ( $r = 0,99789126^{18}$ ). With a correlation such as this one, it is necessary to consider many factors such as who were the people who suicided. It could be possible that with the growth and development of the field/industry, work related stress increased which ultimately propelled people to commit suicide. If this is indeed the underlying causation, then it is quite possible that the overall number of suicides have increased, -- not just the suicides committed by hanging, strangulation and suffocation. This present graph does not provide any information on suicides committed by other means, such as drowning or taking sleeping pills. The likely reason for omitting them is that for other individual causes, or for all causes combined, the correlation disappears. Thus, we are facing here a case of carefully selected two variables that most likely correlate only by coincidence. Overall, this graph

<sup>20</sup> <https://www.tylervigen.com/spurious-correlations>

provides no explanation for how the rise of US spending on science, space, and technology could contribute towards the rise of suicide rates. However, the general public may not be aware of the difference between correlation and causation. Also, the graph presents “correlates with” in fading colours and small fonts, while the two correlating variables are presenting in bright, bold, and bigger fonts. One can only imagine the purpose behind this varying treatment of the fonts and formatting.

**Figure 25: Correlation Between Nicolas Cage and Swimming Pool Drownings**



Source: Tyler Vigen

**Analysing Figure 24:** In figure 24, there is a moderate correlation ( $r=0.666004$ ). For deciding on causation, a stronger correlation is required. Yet the curves in the graph are surprisingly close to each other. Still a logical intuitive reason behind the relationship suggested here is lacking. If films with Nicolas Cage had dangerous stunts involving water, it is possible there would be people who would want to try it and it would not always result in success, and cause deaths. Again, a theoretical explanation and more information about the victims and circumstances is needed to make any meaningful conclusions.

Tyler Vigen displays many other examples of correlations on his website. All of these correlations, similar to the ones discussed in this thesis, merely represent a correlation, and provides no evidence to conclude causation. Simultaneous variations within two variables, does not necessarily determine causality. In other words, these graphs cannot be used to conclude whether one variable is causing the other, or vice versa.

In both cases, other factors may play role. It could also be a way of scaling the data or of pure coincidence. When two variables move as closely, the statistics will detect the dependency, however, it is the researcher's responsibility to distinguish the real causal relationship from cases of data mishandling (data mining), strange coincidences, data visualisation errors, etc.

### **Case XI: Dinking Filter Coffee**

While the previous case showcased mainly correlations that are unlikely to have causal relationship, there are often correlations where causality is much more difficult to determine. As an example, one can contemplate the case of vaccination and autism, which I have described in the introduction. Another talked topic is coffee and its impact on health.

There are plenty articles in the media claiming coffee is good for the consumer. I have chosen an article which claims filtered coffee is good for the drinker even in larger quantities (up to four cups a day). The article was published in *MailOnline* (online version of *Daily Mail*)<sup>21</sup>.

This article in particular states drinkers of filtered coffee are less likely to die, however, does not specify the control group. Thus, it is unclear compared to which treatment group, the drinkers of filtered coffee are less likely to die. It also does not include the information of how much higher the risk is compared to the other groups. Furthermore, the *OnlineMail* article states that "filtered brew was linked to a 15 per cent reduced risk of death from any cause." This phrasing is difficult to understand and interpret properly. Does the author of the *OnlineMail* article believe that coffee is a good cancer prevention, or that it can even prevent car accidents as a cause of death? Probably not. To get a better understanding of the actual results of the study, one really needs to look it up and read. Death from any cause in the study can be in fact any that of a participant that occurred in the timeline of the study. However, these deaths do not necessarily need to relate to coffee consumption. Also, the article does not mention other details such as all the demographic and other background factors that the study controlled (gender, age) and are relevant for understanding how drinking coffee (and which type of coffee or in which quantities) can affect the reader's health. The study, for

---

<sup>21</sup> See Appendix B

example, also finds differences between men and women, which the *OnlineMail* article fails to even mention. The authors of the study also consider that the country the study was conducted in (i.e. Norway) could have affected the results but this aspect is completely neglected by the *OnlineMail*. Essentially everything important and interesting is left out for the reader to find out by himself, from the original, academic study. But how many readers go over the effort? And how many just pick up the information from the headline?

The ongoing case about whether coffee is healthy or not, has been a long existing issue. Many researchers have been and still are trying to study the impacts of coffee on the human body, and the results are often differing from each other. The media, however, tends to take only certain information out of those studies. Oftentimes, it is combined with lack of linked studies, while some of those studies are not the whole study (only an abstract, studies with small sample, or irrelevant studies). Media articles of this nature<sup>22</sup> are easily accessible as well, which requires raised awareness when reading these articles. Especially topics with different interpretations should be read with care as these topics can be used to further polarize society.

Yet it is important to realize, that untrained readers are not likely to proceed as far as to opening and studying more closely the original study. Additionally, without statistical training, they still might not be able to understand and interpret the results properly. All their understanding therefore hinges on the ability of the author of the newspaper article to provide expert and unbiased summary of the most relevant facts discovered by the scientific study. Moreover, as I have already pointed out numerous newspapers or online articles do not provide the reference to source study and thus it is impossible even for the trained readers to check the accuracy of the information.

## **6.5 Case on Surveys**

### **Case XII: Self-reported Data and Its Impact on a Study (Social Desirability Bias, Discrepancies in Self-reported and Measured Data)**

To demonstrate the difficulty with survey results, I decided to use the article "*Do the Obese Know They Are Obese?*" which was published through HHS Public Access by

---

<sup>22</sup> Article without a linked or cited study: <https://time.com/4768860/is-coffee-good-for-you/>  
Article with incomplete studies or irrelevant sources in many cases:  
<https://www.healthline.com/nutrition/coffee-good-or-bad>

Kimberly P. Truesdale and June Stevens<sup>23</sup>. Participants of this study were asked to report their BMI and weight status categories based on their self-reported height and weight. Those results were compared to actual measurements done by a physician. The results of the study have shown that the participants have generally underestimated their weight and overestimated their height, however, were often classifying their weight status inaccurately.

This case shows how self-reported data can be different from the actual data measured. While this study examined the accuracy of self-reported height and weight as well as perceived status category, and it was consistent with other studies, it also admitted participant's possible reluctance to report their weight status to a health care researcher.

As obesity has a social association as negative bias, weight stigma and discrimination (Puhl, Moss-Racusin, M.B., & K.D., 2008), it is possible that respondents of such studies will not be completely honest, especially if there is a lack of measurements being done. This kind of studies raise questions such as which social topics are too sensitive for the respondent to experience a social desirability bias and to what extent these responses can change the results of the survey/study and how to minimize any discrepancies.

Social desirability bias such as this could easily be a topic for surveys in marketing or customer service. In these, it could be influenced by the way the questions are asked as marketing and customer service surveys are often done to not only inform the management about its performance among customers and public. These types of data can often be used for ads, commercials and other marketing platforms to attract new customers and keep the existing ones. The urge to attract and keep customers often poses an issue such as the lack of clear and easily understandable questions or create biases in surveys. Often, the questions are leading or loaded. Albeit useful for attracting customers, said data may be not useful for the management. While it is possible to create two different surveys, it saves money to prioritize a well-made survey to serve both purposes.

---

<sup>23</sup> <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234679/>

## **Conclusion**

I chose the topic of misleading statistics as statistics, no doubt, has infiltrated and shape many aspects of our public and private lives. Although, data-driven policies and decision-making have many positive impacts on socio-political and economical lives, statistics is, more often than not, heavily misunderstood and sometimes even overused by people who lack proper understanding of statistics. Currently, there are numerous online platforms that, intentionally or unintentionally, spread misleading information. At times, these information/claims are supported by misinterpreted or misrepresented statistics; while other times, the online platforms simply do not provide information on how the data was collected, processed, analysed, and interpreted.

Consequently, in this bachelor's thesis, I have discussed and examined the common statistical terms that are used, rather misused, by the general public and media. I have provided examples of multiple real-life practices and their associated problems to describe where and how reader could become victims of false statistics. I have examined topics such as sampling, relative and absolute values, visualizations, and correlations and causations.

Given the recent socio-political and economic ramifications of the COVID-19 pandemic, I have examined the statistical representations of the COVID-19 as well. This allowed me to provide strong of examples of real-life instances where statistics is abused. Simultaneously, it provided me with an insight into how people react to news and media statements that are supported by a statistical claim, while also making me realize just how fast false information can potentially spread. The spread is proliferated as false information travels, without any regulation, through search engines and algorithms. The more shocking the information is the faster it spreads, irrespective of its accuracy.

I consider it important to make the public aware about the use and misuse statistics, especially because, nowadays, statistics forms the basis for critical arguments and decision-making. In politics, statistical claims which are misleading or downright false are used for political campaigns. Eventually, these faulty statistics can potentially end up shaping democratic voting. In healthcare systems, patients' lives depend on statistics as statistical findings are used to determine the effectiveness of new drugs and medical treatments. Critical management and marketing practices rely on the usage of

statistical claims. Companies use statistics in their marketing to attract new and retain old customers. Some companies are often misleading with their claims about their products. In the short term, these claims can cause serious damages to customers, for example by risking their health. In the long term, when these claims do not translate into reality, they can significantly hurt the company's reputation.

This thesis is mainly aimed at the general public which often does not realize the dangers of misleading or false information. Media often creates shock or even panic due to the spread of misleading information. The tendency of the Internet (and media generally) to spread misinformation faster than promoting accurate and reliable sources demands for each consumer of media to consider learning more about the way the media works and critically read media published articles. Understanding the basics of statistics could possibly help the readers to distinguish between the good and the bad sources. Furthermore, statistics can help the reader expand on their critical thinking skills as basic mathematical arguments can be utilized in every day arguments based on logic.

In conclusion, while many rules in statistics are strictly set and must be followed for proper interpretation of data, there is also enough space for the writer's own discretion and creativity, especially in case of visualizations. While simple graphs are easy to follow and often best at conveying the message, many infographics can be fundamentally misleading. It is, however, critical to first understand the practice of statistics and the theory behind it, before using it to make life altering decisions. Or, at minimum, to be aware of great variability of quality of freely available information and lack of editorial supervision. Now more than ever, the readers need to assume the responsibility for the information they take-in, and often share and thereby help to spread. Some basic steps such as checking for availability of reference to the original scientific study and quick look at the quality or reputation of its publisher or skimming for the most common visualization tricks such as those presented in this thesis, can be very helpful especially when making important decisions.





## Works Cited

- Akkerboom, H., & Dehue, H. (1997). The Dutch Model of Data Collection Development of Official Surveys. *International Journal of Public Opinion Research*, 126-145.
- Berk, R. A. (1983). An introduction to sample selection bias in sociological data. *American sociological review*, 386-398.
- Christensen, L., Johnson, R., & Turner, L. (2014). *Research Methods, Design, And Analysis*. Boston: Pearson.
- Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of consumer research*, 303-315.
- Gardenier, J., & Resnik, D. (2002). The misuse of statistics: concepts, tools, and a research agenda. *Accountability in Research: Policies and Quality Assurance*, 65-74.
- Geenstone, M., & Gayer, T. (2007). Quasi-Experimental and Experimental Approaches to Environmental Economics.
- Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 US presidential election. *Science*, 374-378.
- Levitt, S. D., & List, J. A. (2008). Field experiments in economics: The past, the present and the future. *European Economic Review*, 1-18.
- Madsen, K. M., Hviid, A., Vestergaard, M., Schendel, D., Wohlfahrt, J., Thorsen, P., . . . Melbye, M. (2002). A population-based study of measles, mumps, and rubella vaccination and autism. *New England Journal of Medicine*, 1477-1482.
- Moura, H. (2011). Sharing Bites on Global Screens: The emergence of Snack Culture. In D. Y. Jin, *Global Media Convergence and Cultural Transformation: Emerging Social Patterns and Characteristics* (p. ch. 3). Hershey.
- Mrozek-Budzyn, D., Kieltyka, A., & Majewska, R. (2010). Lack of association between measles-mumps-rubella vaccination and autism in children: a case-control study. *The Pediatric infectious disease journal*, 397-400.
- Nickerson, R. S. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology*, 175-220.

- Puhl, R., Moss-Racusin, C., M.B., S., & K.D., B. (2008). Weight stigmatization and bias reduction: perspectives of overweight and obese adults. *Health Educ Res*, 347-358.
- Sheikh, K., & Mattingly, S. (1981). Investigating non-response bias in mail surveys. *Journal of Epidemiology & Community Health*, 293-296.
- Šimundić, A. M. (2013). Bias in research. *Biochemia medica*, 12-15.
- Swoboda, H. (1997). *Moderní statistika*. Praha: Nakladatelství Svoboda.
- Triola, M. F. (2015). *Essentials of Statistics*. Boston: Pearson.
- UNECE, U. N. (2009). Making Data Meaningful. *Part 2: A guide to presenting statistics*. United Nations.
- Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, 1146-1151.
- Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., . . . Walker-Smith, J. A. (1998). *Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children*. *The Lancet*.
- Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual review of sociology*, 327-350.

## Appendix A

2019 12. 04. 10:06 Redakce Seznam

<https://www.seznamzpravy.cz/clanek/prumerny-cech-bere-33-697-korun-za-rok-si-polepsil-o-dva-tisice-83977>

### Průměrný Čech bere 33 697 korun. Za rok si polepšil o dva tisíce

Jaká je průměrná hrubá mzda? (Video: Jan Marek, ČSÚ)

Průměrná mzda v Česku vzrostla o 6,9 procenta, ale ukously z ní rostoucí ceny v obchodech. Lidé si tak ve skutečnosti polepšili o čtyři procenta.

Průměrná mzda v Česku se v letošním třetím čtvrtletí meziročně zvýšila o 6,9 procenta na 33 697 korun. Zaměstnanci tak dostávali v hrubém průměrně o 2 163 korun více než před rokem. Reálně po zohlednění růstu spotřebitelských cen měsíční výdělek vzrostl o čtyři procenta. Údaje dnes zveřejnil Český statistický úřad. Obecně platí, že dvě třetiny zaměstnanců na průměrnou mzdu nedosáhnou. Analytici očekávali, že reálný růst mezd zpomalí pod čtyři procenta.



Průměrná mzda v Česku vzrostla o 6,9 procenta, ale ukously z ní rostoucí ceny v obchodech. Lidé si tak ve skutečnosti polepšili o čtyři procenta.

Průměrná mzda v Česku se v letošním třetím čtvrtletí meziročně zvýšila o 6,9 procenta na 33 697 korun. Zaměstnanci tak dostávali v hrubém průměrně o 2 163 korun více než před rokem. Reálně po zohlednění růstu spotřebitelských cen měsíční výdělek vzrostl o čtyři

procenta. Údaje dnes zveřejnil Český statistický úřad. Obecně platí, že dvě třetiny zaměstnanců na průměrnou mzdu nedosáhnou. Analytici očekávali, že reálný růst mezd zpomalí pod čtyři procenta.

„Růst nominální průměrné mzdy o 6,9 procenta byl výsledkem kompromisu mezi finančními možnostmi podniků a jejich přetrvávající vysokou poptávkou po pracovní síle. Dvě pětiny mzdového nárůstu pohltila inflace, a reálně tak mzdy vzrostly o čtyři procenta,“ uvedl ředitel odboru statistiky trhu práce a rovných příležitostí ČSÚ Dalibor Holý.

## Appendix B

**Filter coffee is GOOD for you: Study shows people who drink up to four cups a day are 15 per cent less likely to die from a heart attack**

<https://www.dailymail.co.uk/sciencetech/article-8245781/Filter-coffee-drinkers-15-cent-likely-die-heart-attack-study-shows.html>

By [IAN RANDALL FOR MAILONLINE](#)

PUBLISHED: 00:15 BST, 23 April 2020 | UPDATED: 01:00 BST, 23 April 2020

- Experts studied the coffee habits of more than 500,000 adults over 20 years
- They found that filter is coffee best for healthy hearts and lowering cholesterol
- In fact, the team report that filter coffee is better for you than no coffee at all
- The filter helps by removing the oily components that help to raise cholesterol

Filter coffee can cut risk of heart attack, with its drinkers being 15 per cent less likely to die from the condition, a study has concluded.

Preparing your morning brew using a filter-based approach is best to avoid heart problems and lower cholesterol — and is better for you than not drinking any coffee.

When coffee is poured through a filter, the oily components that can raise cholesterol and lead to health complications are removed.

Previous studies have shown coffee consumption to be linked to so-called 'bad' cholesterol — and be potentially detrimental to the health of one's heart.

Further research revealed the substances in coffee responsible for this effect can be removed simply by using a filter.

In fact, a cup of filtered coffee contains around a 30 times lower concentration of fatty lipid-raising substances in comparison with unfiltered coffee.

In the new study, European researchers looked at the relationship between different methods of brewing coffee and risk of heart attack and death.

More than 500,000 healthy men and women aged 20–79 recorded the amount and type of coffee they drank, for an average of 20 years.

They also recorded other factors that could influence heart health — such as smoking, physical activity, blood pressure and cholesterol.

The team's analysis revealed that drinking coffee is not a dangerous habit in and of itself — but drinking filtered coffee was safer than no coffee at all.

Filtered brew was linked to a 15 per cent reduced risk of death from any cause — regardless of age, gender or any lifestyle choices.

The risk of death from heart disease, specifically, was lowered by 20 per cent in women and 12 per cent in men who drank filtered coffee.

Furthermore, people who drank between one and four cups of filtered coffee a day had the lowest mortality levels.

Unfiltered coffee did not raise the risk of death compared to no coffee at all — with the exception of with men aged 60 and over, where it was linked to an increased chance of cardiovascular death.

'Our study provides strong and convincing evidence of a link between coffee brewing methods, heart attacks and longevity,' said paper author and epidemiologist Dag Thelle, of the University of Gothenburg in Sweden.

Professor Thelle recommends switching to filtered coffee, especially if you are concerned about high cholesterol.

'For people who know they have high cholesterol levels and want to do something about it, stay away from unfiltered brew,' he said — noting that this would include coffee made with a cafetière, or French press.

'For everyone else, drink your coffee with a clear conscience and go for filtered.'

The full findings of the study were published in the [European Journal of Preventive Cardiology](#)<sup>24</sup>.

---

<sup>24</sup> <https://journals.sagepub.com/doi/10.1177/2047487320914443>

## Table of Figures

Figure 1: Ebbinghaus Illusion.....	21
Figure 2: Simultaneous Contrast Illusion.....	22
Figure 3: "Changing ratio of graph dimensions" .....	23
Figure 4: Interest rates: omitting zero y-axis (left), y-axis starting at zero (right) .....	24
Figure 5: The same trend shown in a graph with zero y-axis (left) and omitted zero y-axis (right).....	25
Figure 6: "Missing data".....	26
Figure 7: Example of an unclassed choropleth map.....	28
Figure 8: Disease Deaths per Day Worldwide.....	35
Figure 9: Mentions in the Media .....	37
Figure 10: Those Aged 60+ Are Most at Risk .....	38
Figure 11: The Fatality Rate Varies by Country.....	39
Figure 12: Especially Those with Existing Conditions.....	41
Figure 13: Confirmed Cases of COVID-19 in Czech Republic, March 12, 2020 .....	42
Figure 14: Confirmed Cases of COVID-19 in Czech Republic, March 20, 2020 .....	42
Figure 15: Graph of the day: Development in Number of VAT Payers .....	44
Figure 16: Development of VAT payers.....	44
Figure 17: "Japan Competitiveness Rank" .....	45
Figure 18: Japan's Competitiveness Index.....	46
Figure 19: Gross Average Wage in Regions for 3Q 2019, Annual Growth.....	48
Figure 20: Average Gross Monthly Wages by Regions.....	49
Figure 21: Average Gross Monthly Wages by Regions – 3Q 2019.....	49



**Figure 22: Distribution of Wages of Non-government Employees ..... 52**

**Figure 23: Distribution of Salaries of Government Employees ..... 52**

**Figure 24: Correlation Between US Spending on Science and Hanging Suicides ..... 53**

**Figure 25: Correlation Between Nicolas Cage and Swimming Pool Drownings ..... 54**

## **Table of Tables**

Table 1: Level of Measurement.....	27
Table 2: Types of Graphs and Charts .....	28
Table 3: Average Wages by Occupation by CZ-ISCO-08 Major Group.....	51
Table 4: Daily production and 4-day moving average.....	54