How to Lie with Statistics - A book by Darrell Huff

Seminar: How do I lie with statistics?

by

Peter Hügel

Name:	Peter Hügel
Enrollment Number:	3140348
Supervisor:	Prof. Dr. Ullrich Köthe

Contents

1 Introduction 1

2 Simple Ways to Lie 2

2.1 Selection Bias 2

2.2 The Average 3

2.3 Missing Figures 4

2.4 Charts and Pictographs 8

2.4.1 Charts 8

2.4.2 Pictographs 8

2.5 Semi Attached Figures 10

2.6 Correlation and Causation 10

3 Causes for Lies 12

4 Identifying Lies in Statistics 14

Bibliography 15

1 Introduction

The intention of this report, and the corresponding presentation, is to provide an introduction and brief overview to the lies that can be found in statistics. It is heavily based on the book *How to lie with statistics* [2] by Darrell Huff. The book is mostly a collection of examples showcasing the extent to which statistics can be used to convey false information, intentional or not. Statistics are all around us, consumed by us, or produced by us. In our daily lives we make statistical assumptions all the time. We make such assumptions based on a collection of experiences and extrapolate. Statistics we are exposed to may have a lot less or a lot more to them than we think at first glance. While technically being correct, there are many ways they can deceive us. Our goal is to spot and identify these lies. In chapter 2 different methods of lying are explored through a collection of example cases. After covering the possible causes and motivations for these lies, the author suggests a list of questions we should ask ourselves in order to identify lies in statistics.

Figure 1.1: Figure from *How to lie with statistics* [2]

The book is in no way supposed to be a guide on how to successfully lie with statistics. The authors intention is to spread awareness, stating that "the crooks already know these tricks; honest men must learn them in self-defense" - Huff, *How to lie with statistics* [2]. The idea is to help identify lies and avoid producing them accidentally.

2 Simple Ways to Lie

This chapter will cover a collection of different simple ways to lie. Simple, as only the most basic concepts are covered in order to provide an introduction to the topic. Most of these lies can have different causes that lead to the same outcome, technically correct statements that convey false information.

2.1 Selection Bias

"The average Yaleman, Class of 1924 makes $24,111 a year." was stated by *Time magazine*. At first glance one could simply absorb this information. On second glance though, some questions come to mind.

One might ask how this number was derived. Obviously these graduates had to be questioned in some way, at least part of them. It is quite likely that only a sample of the graduates could be contacted. This brings up more questions, who is more likely to be found? It is probably a lot easier to find wealthy and/or famous graduates of that class than the less successful ones. This may, to some degree, introduce a bias to the sample of these Yalemen.

Figure 2.1: Figure from *How to lie with statistics* [2]

This is not the only thing to be considered in this case. Another factor to take into account is the truthfulness to their responses. How likely are they to respond accurately? This may depend on how they are questioned. Are they mailed a form to fill out? Are they questioned over the phone? Are they questioned at the doorstep?

In any case, the graduates responses may have been embellished or downplayed. While these two factors may balance each other out, if they don't, they taint the sample, inducing further bias to the selection. In conclusion, a sample has to either be chosen very carefully, or all biases have to be accounted for in order to avoid a selection bias. "A result of a sampling study is no better than the sample it is based on." [2]

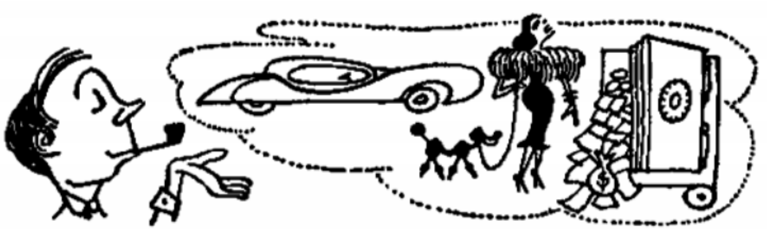

Figure 2.2: Figure from *How to lie with statistics* [2]

2.2 The Average

The average can represent various key figures. This does not have to be a bad thing in itself, different situations require different methods to obtain a meaningful measure representing the average. Depending on the case, some types of averages can make much more sense than others.
In a fictitious factory the salary may be distributed as visible in figure 2.3. Here are some different averages based on this data:

- Mean salary of workers: $\$2,308$
- Mean salary of management: $\$25,000$
- Mean salary: $\$3,309$
- Median: $\$2,400$
- Mode: $\$2,000$

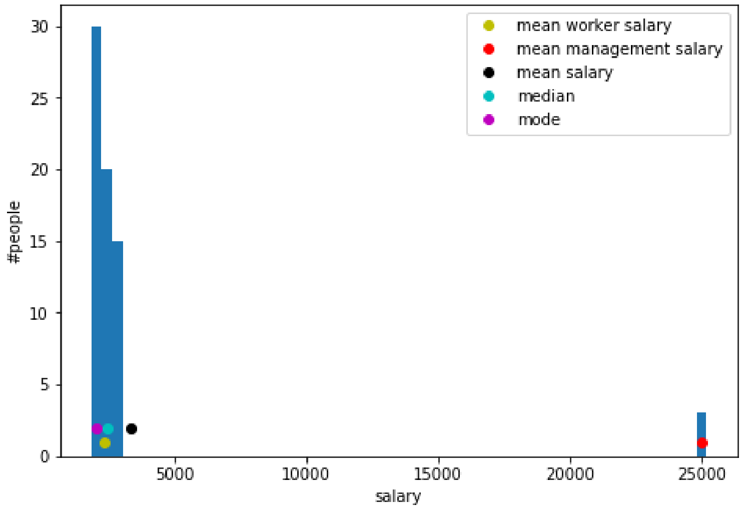

Figure 2.3: Different types of averages are indicated in a bar chart that displays the distribution of salaries in a fictitious factory. Mean, median, and mode may simply be labeled "average".

Different stories can be fabricated from this data. Assuming the Labor Union and the factory Management both publish a report on this matter:

Labor Union	Average salary of employees: $\$2,000$ Average salary of management: $\$25,000$
Management	Average salary payed: $\$3,309$

Special attention should be given to the wording in the above table. In order to paint a worse picture, the mode of all employees, including management, may be concealed with "average salary of employees". This of course does not present reality accurately. The same can be done for the other party: "Average salary payed: $\$3,309$" sounds a lot better. In this case the mean of all employees, including management, was chosen. This of course gives the impression that workers are paid a lot more than they really are. If the average is not explicitly specified, the real picture might be distorted deliberately.

2.3 Missing Figures

The claim that "Users report 23% fewer cavities with toothpaste **X**!" was verified by an independent laboratory and certified by a public accountant. It would seem that it is either a revolutionary new type of toothpaste, or the statistic is lying in some way.

When digging a bit deeper, it turns out that the number of participants was only 12, this is a highly inadequate sample size. In order to get the desired outcome, a study can be repeated until a satisfactory result is obtained. In this case there are three possible outcomes:

- Distinctly more cavities
- Distinctly fewer cavities
- About the same as before

With a small enough sample size it shouldn't take too long to get lucky and obtain the desired result. This process of repeating, cherry-picking, and discarding is called observer selection. So the missing figure in this case is the number of repeats that have been performed, or an indication of how meaningful the statement is.

Analog to the toothpaste example, assume successive coin tosses. We can calculate the probability for getting tails over 75% of the time using the Binomial Coefficient:

$$\binom{N}{k} = \frac{n!}{k! \times (n-k)!} \tag{2.1}$$

The following section will demonstrate how the size of the sample influences the statistical significance of the result.

When throwing a coin 8 times, the probability to get at least 75% tails is $\binom{8}{6}/2^8 = \frac{28}{256} \approx 0.109$. So in about 10.9% of the experiments the result is that the coin ends up with tails 75% or more. We can easily repeat this process a few times to end up with the desired outcome.
When trying to show the same for a series of 128 coin tosses, the probabilities are different. To satisfy the same condition, 75% of throws resulting in tails, tails is required to show up at least 96 times. The probability for the desired result is now $\binom{128}{96}/2^{128} \approx \frac{1.48 \times 10^{30}}{3.4 \times 10^{38}} \approx 4.3 \times 10^{-9}$. This is a considerably lower chance, illustrating the importance of an appropriate sample size. A comparison of the probability density functions is shown in figure 2.4

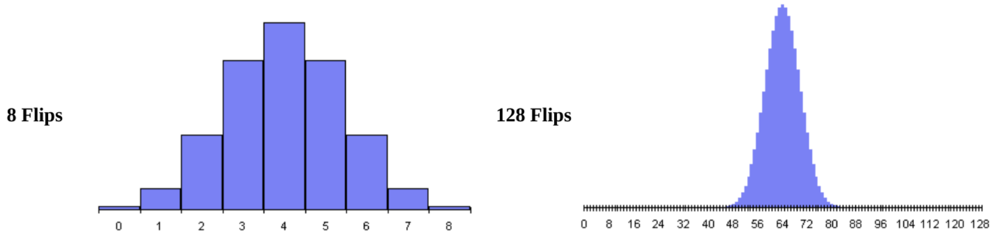

Figure 2.4: As the sample sizes increases, it is more and more likely to end up with an accurate representation instead of an outlier.

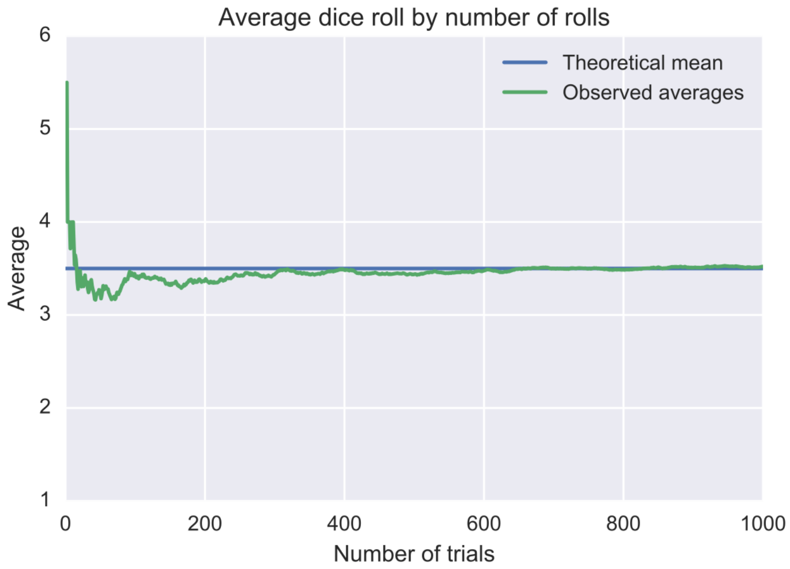

Figure 2.5: With increasing sample size the result converges to the true probability, this is known as the law of large numbers.
Figure from `https://en.wikipedia.org/wiki/Law_of_large_numbers` [3]

Figure 2.5 displays how an increase in sample size leads to more stable results. The amount of statistical significance can be expressed with a number, the p-value. In our coin tossing example, the p-value is calculated for the case of getting tails at least 75% of the time. A visualization of the p-value can be found in figure 2.6.
More generally speaking, the p-value is the chance of getting the observed or a more extreme result while assuming the null hypothesis. The null hypothesis is the default position, assuming that there is nothing new happening. In the case of the coin toss the null hypothesis is to assume equal probabilities for head and tails. For the toothpaste example it is the assumption, that all toothpastes are equally effective.

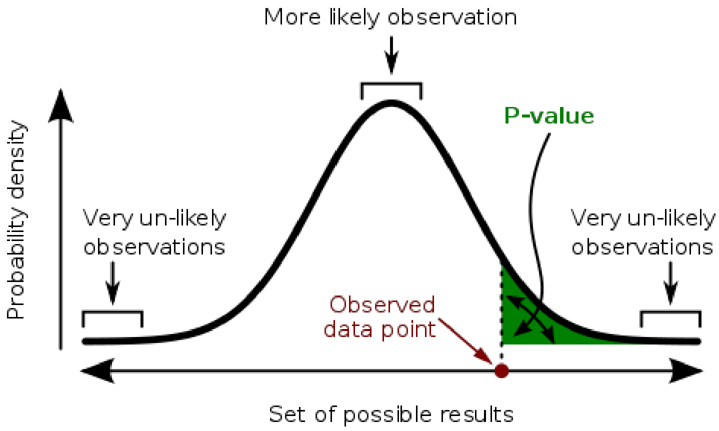

Figure 2.6: A p-value (shaded green area) is the probability of an observed (or more extreme) result, assuming that the null hypothesis is true.
Figure from https://en.wikipedia.org/wiki/P-value [5]

These are the p-values for the coin tossing case:

- 6 out of 8 tails or more: 1.45×10^{-1}
- 96 out of 128 tails or more: 6.42×10^{-9}

Considering the inadequate sample size, the p-value of the toothpaste study would also be quite high, resulting in a low statistical significance.

The editor of *Reader's Digest* published results of a study about the harmful substances in different tobacco brands. As expected there was not much difference between the brands. Because such measurements have an inherent inaccuracy, the actual values varied slightly between the brands. One of the brands had to contain the least harmful substances considering the nature of the study.
After the results were published, the brand that scored the best boasted to have the healthiest cigarettes of them all! While the original publication properly conveyed the fact that the brands are essentially equal, the marketing campaign of the brand conveniently only mentioned who "won". The missing figure in this example is an indication of range.

One indication of range is the standard deviation, a measure of variation or dispersion of a set:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}} \tag{2.2}$$

In the case of the tobacco study, repeatedly testing the same brand would probably yield a standard deviation that is higher than the difference between the brands. So, as reported by *Reader's Digest*, the difference in harmful substances is negligible.
Box & Whisker plots are a visualization method to quickly convey the rough distribution of the samples. The set is split into quartiles, the start and end of each segment is visualized. The whiskers display the bounds of the first and last quartiles, the box contains both center quartiles and is split at the position of the median. Whiskers may be limited in some way, e.g. outliers are drawn as points instead of being included in the whiskers. An example for Box & Whisker plots is shown in figure 2.7.

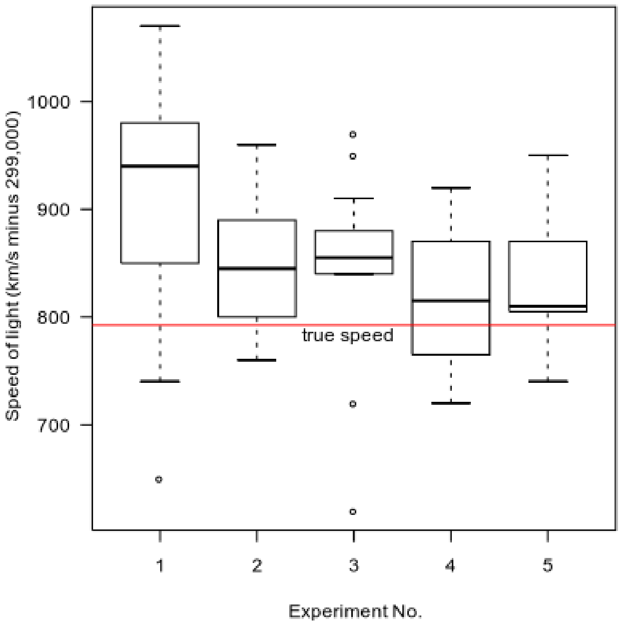

Figure 2.7: Figure from `https://en.wikipedia.org/wiki/Box_plot` [1]

"The deceptive thing about the little figure that is not there is that its absence so often goes unnoticed" Huff,- *How to lie with statistics* [2].

2.4 Charts and Pictographs

As there is a dedicated presentation about charts, this chapter is intentionally kept short and only covers a small sample of the issues that can arise with charts and pictographs.

2.4.1 Charts

One may argue that the first part of figure 2.8 wastes a lot of paper and merely displays empty space. With this line of reasoning the second and third part can seem reasonable. Depending on the circumstances and the realization, reality may be distorted greatly. If an axis does not start at zero, it has to be indicated clearly. Comparing the leftmost and rightmost chart in figure 2.8, two very different narratives can be pushed. While the first accurately depicts a change of about 10%, the latter, on first glance, conveys the presence of a major increase.

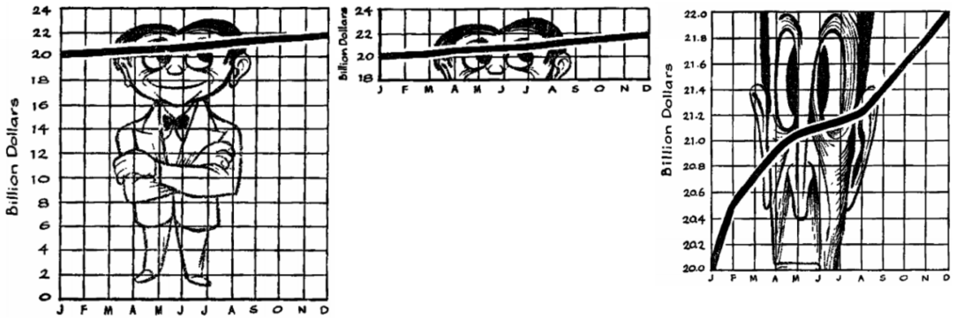

Figure 2.8: Figure from *How to lie with statistics* [2]

2.4.2 Pictographs

In order to make a chart more interesting, pictographs may be used. Figure 2.9 displays two different pictographs that are created from a chart. While the first pictograph properly visualizes the proportions by displaying twice as many bags, the other attempts to do the same by having differently sized bags. The problem with the second pictograph is, while the correct factor of two is indeed present in the figure, that factor does not apply to the relevant property of the bag. The factor two was used in the height of the bag, this of course results in a factor of four for the area. If one thinks of the bag in three dimensions, it could even be interpreted as a difference of factor eight.

The effect can be amplified by adding shadows and other properties to guide the perception of the beholder. This is especially visible in figure 2.10. While the two circles in the image are equal in size, we are deceived by the context. In “Perceived size matters” [7], Sterzer and Rees discuss the flaws in human perception of size.

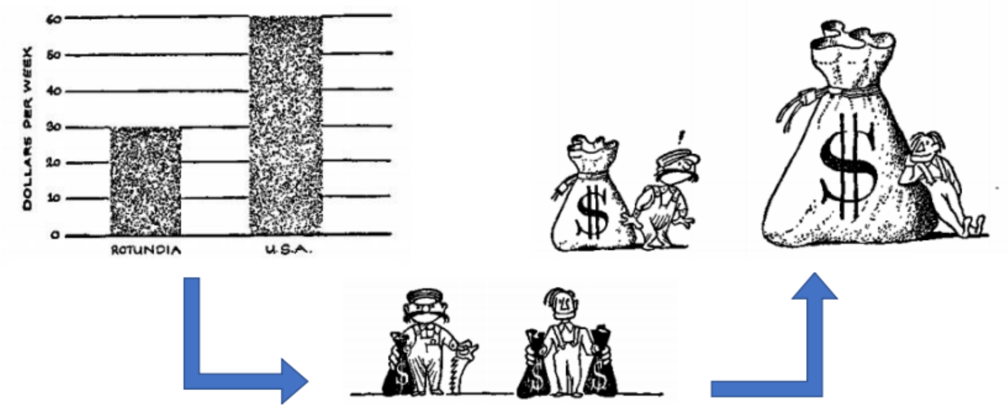

Figure 2.9: Two different realizations of pictographs that are based on the same chart. Figure from *How to lie with statistics* [2]

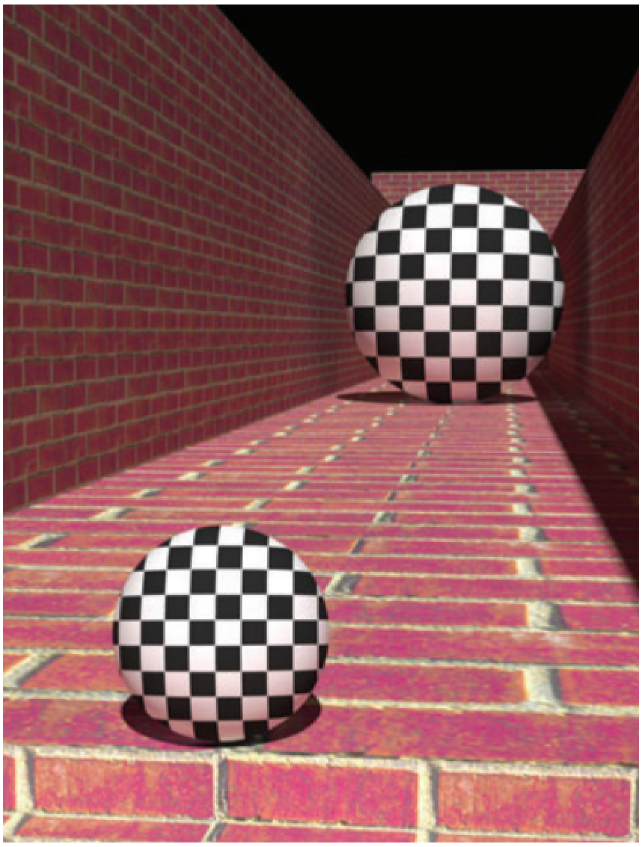

Figure 2.10: While the sphere in the back is perceived to be larger, the area they cover in the image is equal. Figure from “Perceived size matters” [7]

2.5 Semi Attached Figures

A company advertised that their new electric juicer "extracts 26% more juice". When looking for a new juicer, this makes it seem like a good choice. One might ask themselves how this figure was obtained, which juicer was the product compared to? When digging a bit deeper, it turns out, that the 26% come from a comparison to a hand juicer. It is likely that most electric juicers extract more than hand juicers. The advertised 26% are useless and misleading when choosing a good electric juicer to buy.

Another case came up when looking at the death rates of navy personnel and civilians in New York City. While there were 9 deaths per 1000 people in the navy, there were 16 per 1000 for the civilians. Assuming these numbers to be correct, one might claim that it is obviously safer to be in the navy than being a civilian in NYC. After all you are less likely to die according to the statistics.
This is of course not the case, the groups are not comparable. Navy personnel are, on average, younger and healthier than the civilian population. It sounds plausible, that the amount of natural deaths is lower in the navy. Although it may be implied through advertising, just joining the navy doesn't improve your life expectancy.

Things may sound the same at first, but they often times are not. This method is also known as the association fallacy. The following quote sums it up quite nicely:
“If you can't prove what you want to prove, demonstrate something else and pretend that they are the same thing. In the daze that follows the collision of statistics with the human mind, hardly anybody will notice the difference.” - Huff, *How to lie with statistics* [2].

2.6 Correlation and Causation

There is a story about the natives living on the island Vanuatu. According to them, having lice causes good health. Considering the evidence they had, it makes sense how they came to that conclusion. By default, everyone in the tribe had lice. Now if one of them fell ill, there were no lice on them anymore. This might lead some to think, that the reason for the illness was the absence of lice.
Of course there is a better explanation. As it turns out, a temperature increase of a few degrees is fatal for lice. So whenever someone in the tribe became ill and had a fever, their temperature rose slightly, causing the lice to die off.

Figure 2.11: Figure from *How to lie with statistics* [2]

Types of Correlations

Correlation by chance:
Looking back at the toothpaste case from section 2.3, the repeated execution of the same study will eventually come up with a correlation by chance. Here the correlation by chance is that the new toothpaste causes fewer cavities. A list of correlations by chance can be found on the website *Spurious Correlations* [6]. One such case is the apparent correlation between the number of people who drown in swimming pools and the number of films Nicolas Cage appeared in.

Real correlation, but what is the cause and what is the effect:
An example for this type might be the correlation between income and ownership of stock. The more money is made, the more stock is owned. And the more stock is owned, the higher the income. For this type the role of cause and effect may change place or both are cause and effect for the other.

Confounder - real correlation, but a third factor is the cause for both:
An example for this would be the correlation between the number of younger siblings and the presence of down syndrome. One does not cause the other, even though there is a correlation. The third factor here might be high maternal age. It increases the chance of having a larger number of younger siblings and the chance of having down syndrome.

3 Causes for Lies

The author argues that most lies in statistic come from ill intent rather than from incompetence or by chance. His reasoning is, that the mistakes are conveniently one-sided and usually align with some motives of the creator or publisher. As mentioned before, there is usually a narrative that is being pushed. Figure 3.1 pictures two possible stories that are again based on the same data. In this case, within a year, the price of Bread has doubled while the price of milk has halved. If 100% is assigned to the values of the previous year, the arithmetic mean of the current years prices is $\frac{200\%+50\%}{2} = 125\%$. When choosing this years prices as a base, the arithmetic mean of the past years is $\frac{50\%+200\%}{2} = 125\%$. This is a confusion of base. The solution to this issue is the usage of the geometric mean:

$$\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = \sqrt[n]{x_1 x_2 \dots x_n} \tag{3.1}$$

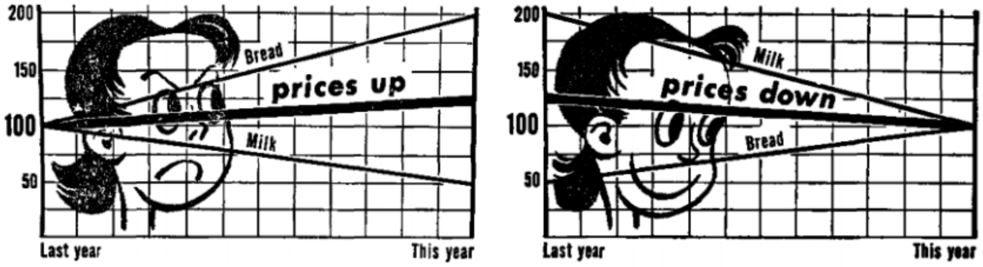

Figure 3.1: Depending on the chosen basis the arithmetic mean conveys contradicting conclusions. The proper type of average to use in this case is the geometric mean.
Figure from *How to lie with statistics* [2]

When applying the geometric mean (equation 3.1) to this case, reality is properly represented:

$$\sqrt[2]{200\% \times 50\%} = 100\%$$

Making a "mistake" here results in one of two stories: "Cost of living up!" or "Cost of living down!". In the past this method was employed when advertising sales, today the following method is illegal and counts as unfair business practice. "Buy your Christmas presents now and save 100%" sounds better than what is actually meant, that the price is halved. The 100% originate from choosing the new price as the base.

Huff points out, that for all these cases, usually the "mistake" that is made benefits the publisher or creator. The distortion and manipulation of statistics is not always the work of professional statisticians. Legitimate work may be cherry-picked, embellished, and used for personal gain. For most "mistakes" or lies in statistics a motive or goal can be found:

- Sensationalize
- Inflate
- Confuse
- Oversimplify

In any case, oftentimes there is a narrative being pushed. In the authors opinion the reader should be able to discern proper statistics and recognize sound and usable data. According to him the following should be ensured:

- Competence & integrity of the statistician
- Competence & integrity of the writer
- Competence of the reader

Instead, maybe statistics can be changed in a way that removes human error from the equation. This topic is explored in topic 9 of this seminar: "How to do better?".

4 Identifying Lies in Statistics

As Huff asks for competence on the part of the reader, he suggests a list of questions to ask in order to discern lies in statistics:

1. Who says so?
    - A reputable name does not imply proper representation of the data.
    - Who is drawing the conclusions?
2. How does he know?
    - How was the data obtained?
3. What's missing?
    - Are averages not specified?
    - What is the quality and size of the sample?
4. Did somebody change the subject?
    - Association fallacy / semi attached figure.
5. Does it make sense?
    - A statistician decided to judge readability solely based on average word-length.

In Rheinische Post there was an article about nitrate levels, sensationally titled *More and more nitrate in groundwater* [4]. It included the following statement: "From 2013 to 2017 the average nitrate concentration in the top 15 polluted regions has increased by 40mg/l".
Regarding the first question, the motive here is clearly to produce a sensational article. How do they know? What is missing? Asking these questions and digging a bit deeper uncovers that there is a combination of missing figures and selection bias at play. As it turns out, the top 15 polluted regions are not only different between the two time periods, but the measurement devices are actively moved from uninteresting / safe regions to regions that are at risk, that have a higher nitrate concentration. The missing figure is what hides behind "top 15 polluted regions", the sampling is obviously heavily biased.
The reader should also be alarmed by the word average, which average was used here? Not only is it not specified, but the average from the year 2013 is a different one than the one from 2017. In 2013 the average was the mean over the whole year. In 2017 the average was made up of maxima over multiple days. Properly comparing the same regions to each other, it turns out that the nitrate levels actually steadily decreased.

Bibliography

1. *Box plot*. URL: `https://en.wikipedia.org/wiki/Box_plot` (visited on 10/15/2019).
2. D. Huff. *How to lie with statistics*. WW Norton & Company, 1993.
3. *Law of large numbers*. URL: `https://en.wikipedia.org/wiki/Law_of_large_numbers` (visited on 10/15/2019).
4. *More and more nitrate in groundwater*. URL: `https://rp-online.de/wirtschaft/immer-mehr-nitrat-im-grundwasser-gefahr-fuer-mensch-und-natur_aid-44825553` (visited on 08/08/2019).
5. *p-value*. URL: `https://en.wikipedia.org/wiki/P-value` (visited on 10/15/2019).
6. *Spurious Correlations*. URL: `http://www.tylervigen.com/spurious-correlations` (visited on 10/15/2019).
7. P. Sterzer and G. Rees. “Perceived size matters”. *Nature Neuroscience* 9:3, 2006, p. 302.