UPPSALA
UNIVERSITET

An explorative literature review on misinterpretations of confidence intervals

Åsa Håkansson

Bachelor's thesis in Statistics

Advisor

Ronnie Pingel

2021

Abstract

Confidence intervals are presented in various scientific fields and are used to justify claims, although in several studies it has been shown that various groups of people, for example researchers, teachers and students incorrectly interpret the confidence interval. To get an overview over articles which studies how confidence intervals are misinterpreted we have performed an explorative literature review. The search for articles to include was conducted in a semi-structured way, and the explorative literature review exist of 36 articles. The results are presented in tables, where the articles have been placed in the table of the relevant misinterpretation. In this study there are five different tables which are probability fallacies, precision fallacies, likelihood fallacies, overlaps fallacies and miscellaneous fallacies. This paper state that confidence intervals are commonly misinterpreted accordingly to our first four categories. In the last category, the less common misinterpretations are presented, at least less common based on the articles included in this paper.

Keywords: Confidence interval, misinterpretation, explorative literature review

Table of contents

1. Introduction

1.1 Background

A central part in the scientific fields when it comes to justify claims is statistical inference. When answering research questions quantitatively, published results are typically supported by statistical outputs such as p-values and/or confidence intervals. In social sciences, null hypothesis significance testing (NHST) is the most used inferential technique to justify claims (Hoekstra et al., 2014). This technique has received criticism for a long time whilst confidence intervals got a growing advocacy (see e.g. Rothman, 1978). Rothman, the founder and the previous editor for the journal Epidemiology, strongly discouraged the use of p-values (Lang et al., 1998). Although Lang, Rothman and Cann (1998) wrote that a ban of reporting p-values in the journal would be too dogmatic, but they still pointed out and argue that the p-value is confounded information.

Compared to NHST, confidence intervals are often claimed to be an alternative that is better and more useful (Hoekstra et al., 2014; Harlow, 1997). Confidence intervals are also recommended by many proponents of statistical reform in social and behavioural sciences. The American Psychological Association (APA) Publication Manual, and various APA and other peer reviewed journals have also recommended confidence intervals. Confidence intervals compared to p-values offer more information and still remain in the same frequentist framework. This reason is why confidence intervals are recommended instead of other statistical techniques (Kalinowski, 2010). The advertising of confidence intervals is presented in such a way that it is superior compared to NHST (e.g., Cumming & Finch, 2001; Fidler & Loftus, 2009; Schmidt, 1996; Schmidt & Hunter, 1997).

As we will see, various arguments of why confidence intervals should be used instead exists, but as mentioned in the article *Robust misinterpretation of confidence intervals* by Hoekstra et al. (2014) this rest on the idea that confidence intervals can be properly interpreted by researchers. Because even though the above articles argue for the use of confidence intervals, the interpretation is not straightforward.

That confidence intervals and p-values are often *incorrectly* inferred is known and has been studied thoroughly (e.g., Naimi & Whitcomb, 2020; Hoekstra et al., 2014; Greenland et al., 2016; Biau et al., 2010). As mentioned earlier, these statistics are the ground for scientific claims, and it is therefore essential that researchers interpret them appropriately. One of the main goals in science is to justify knowledge, and this can be undermined if the statistics are not appropriately interpreted (Hoekstra et al., 2014). The confidence intervals are introduced in almost every introductory statistics text, and they are also by methodological guidelines of many prominent journals recommended or required (Wilkinson & the Task Force on Statistical Inference, 1999); and in the methodological reformers' proposed programs confidence intervals are the foundation (Cumming, 2014; Loftus, 1996).

In this study we will perform an explorative literature review concerning how confidence intervals are misinterpreted in various studies. The decision to make an explorative literature review is because no previous explorative literature review about this has been made. A critical point of deciding how science is going to be done in the future, not to speak of the current atmosphere of methodological reform, is what kind of inferences confidence interval theory allows or not (Hoekstra et al., 2014). It is therefore relevant to conduct an explorative literature review and get an overview over how confidence intervals are inferred.

1.2 Aim

The aim of the paper is to perform an explorative literature review that consists of articles, which studies how confidence intervals are misinterpreted. To meet the purpose, one research question has been formed, which is presented below.

- How are confidence intervals misinterpreted?

2. Theory

2.1 Confidence intervals

An important component of statistical analyses is interval estimation, which allows taking sampling uncertainty into account when estimation of parameters are made. The different approaches of interval estimation differ regarding their philosophical foundation and computation. By giving the parameter a range of values, instead of a single one, interval estimates are supposed to form an accountability for sampling uncertainty or measurement. The most popular interval estimate is the confidence interval (Hoekstra et al., 2014) which is a basal tool if one wants to be able to quantify uncertainty due to sampling associated with a point estimate (Naimi & Whitcomb, 2020).

In 1937 Neyman laid the formal foundation for confidence intervals, which he defined as followed:

"*An X% confidence interval for a parameter θ is an interval (L, U) generated by a procedure that in repeated sampling has an X% probability of containing the true value of θ, for all possible values of θ*".

To exemplify, we will present a confidence interval for the parameter mean in a normal distribution with known variance σ^2 below. We have an *unknown* mean, this parameter is called μ and our unbiased estimator for our parameter $\bar{X}$. Let $X_1, X_2, \dots, X_n$ be an iid sample from N(μ,σ^2) and let the sample mean, $\bar{X}$, be the point estimate for μ. To construct a two-sided confidence interval with confidence level 1-alpha, we find a number $z_{\alpha/2}$ from a table such that

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Because $\frac{\bar{X},}{\frac{\sigma}{\sqrt{n}}}$ follows a standard normal distribution, we easily get values for $z_{\alpha/2}$.

$$P\left(-z_{\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha,$$

$$P\left(-z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \bar{X}-\mu \leq z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha,$$

$$P\left(-\bar{X}-z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \leq -\mu \leq -\bar{X}+z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha,$$

$$P\left(\bar{X}+z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \geq \mu \geq \bar{X}-z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha.$$

The probability of the first of these is $1-\alpha$ which means that the probability of the last one also must be $1-\alpha$. The latter is true if and only if the former is true. This means that we have

$$\mathrm{P}\left(\bar{X}-z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \bar{X}+z_{\frac{\alpha}{2}}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha.$$

So, the formula above is telling us that a random interval has the probability $1-\alpha$ to include μ. We can also see that μ is centered between the point estimate and the quantity $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
Once we have an observed sample and a calculated sample mean we have a known interval

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Before a sample has been drawn, the probability to get a random interval that covers μ is equal to $1-\alpha$. The computed interval $\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ is called a 100(1- α)% confidence interval for the unknown μ. So, in our case for example a 95% confidence interval for μ is equal to $\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$ because z_{α}= 1.96. The confidence coefficient is equal to the number $1-\alpha$. So, a 95% confidence interval has 0.95 as confidence coefficient. We have the same confidence coefficient $1-\alpha$ independently of an increase in n, the increase only results in a decrease in $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ which means we get a shorter confidence interval. The length of a confidence interval

can, by decreasing the confidence coefficient, also be shortened with a fixed sample size n but this results in a shorter confidence interval with loss in confidence.

In the confidence interval, the confidence coefficients are derived from the *procedure* that generated it. To differentiate a procedure from a confidence *interval* is hence helpful. The X % confidence procedure is a procedure that is generating intervals that covers θ in X % of repeated samples, and a confidence interval can be explained as a specific interval generated by that kind of process. In other words, a confidence interval is observed and fixed, while a confidence procedure is a random process (Morey et al., 2016).

There were three steps that Neyman (1937) suggested researchers to follow:

- *Perform an experiment, collecting the relevant data.*
- *Compute two numbers – the smaller of which we can call L, the greater U – forming an interval (L, U) according to a specified procedure.*
- *State that $L<\theta<U$ – that is, that θ is in the interval.*

This suggestion is justified by, in the second step, choosing a procedure so that the researchers claim, in the third step, will be correct on an average of X % of the time in the long term. By using such a procedure, with any interval, you will have computed a confidence interval.

The last step is simply a dichotomous statement, that in the long run is supposed to have a specified probability of being true. It is not a conviction, conclusion, or reasoning from the data. It's neither about an uncertainty about whether θ is in the interval or not.

The problem that confidence interval theory was formed to solve, was a very constrained one which was the following

"How can one construct a procedure that produces intervals containing the true parameter a fixed proportion of the time" (Morey et al., 2016).

A confidence interval can be described as a numerical interval set up around the estimate of a parameter. It is not for a statement about the parameter that the confidence interval is supposed to be used, as it relates to the specific sample at hand. Instead, it provides a statement about the conception of the procedure of drawing such intervals repeatedly (Hoekstra et al., 2014). All a given confidence interval that is calculated by hand from the data can tell us, is that the parameter either lies in the interval or not (Briggs, 2012).

2.2 Confidence intervals vs. Credible intervals

In contrast to confidence intervals, we will now look at credible intervals. Confidence intervals are used in the frequentist approach while credible intervals are analogously used in Bayesian inference (Hespanhol et al., 2019). These intervals should not be treated interchangeably (Hoekstra et al., 2014). The frequentist approach is the most known and widely used approach for statistical inference, where the population parameters of interest are treated as *fixed* values. Unlike the frequentist approach, the parameters of interest in the Bayesian approach are treated as *random* variables and can therefore be described by probability distributions. The aim in the Bayesian approach is to estimate a certain parameter from the population distribution, given the observed (collected) data (Hespanhol et al., 2019).

One distinguished characteristic of the Bayesian approach is the comparison between prior evidence and the observed data. They are both represented with probability distributions, in a Bayesian terminology, and described as prior respectively likelihood distributions. By combining the prior distribution with the likelihood distribution, we are able to update the previous knowledge which results in the posterior distribution (Hespanhol et al., 2019).

To exemplify, we will present a credible interval for the parameter mean in a normal distribution with known variance σ^2. Let $X_1, X_2, \dots, X_n$ be an iid sample from N(μ,σ^2) and let the sample mean, $\bar{X}$ be equal to Y. Also, $g(y|\mu)$ is $N(\mu, \frac{\sigma^2}{n})$. Furthermore, we can suppose that we can assign prior weights to θ through a prior probability mass distribution function (p.d.f.) $h(\mu)$, which is $N(\mu_0, \sigma_0^2)$. The weights in the prior $h(\mu)$ are assigned to the different possible values of μ. The $g(y|\mu)$ can be thought of as the conditional p.d.f. of Y, given the μ.

We then get

$$\mathrm{k}(\mu|\mathrm{y}) \propto \frac{1}{\sqrt{2\pi}\frac{\sigma^2}{\sqrt{\mathrm{n}}}} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left[-\frac{(y-\mu)^2}{2\left(\frac{\sigma^2}{n}\right)} - \frac{\left(\mu-\mu_0\right)^2}{2\sigma_0^2}\right].$$

If all constant factors are eliminated, we get

$$\mathrm{k}(\mu|\mathrm{y}) \propto \exp\left[-\frac{\left(\sigma_0^2+\frac{\sigma^2}{n}\right)\mu^2 - 2\left(y\sigma_0^2+\frac{\mu_0\sigma^2}{n}\right)\theta}{2\left(\frac{\sigma^2}{n}\right)\sigma_0^2}\right].$$

We can then simplify the formula above, so we get

$$\mathrm{k}(\mu|\mathrm{y}) \propto \exp\left\{-\frac{\left[\mu-(y\sigma_0^2+\frac{\mu_0\sigma^2}{n}/\,(\frac{\sigma_0^2+\sigma^2}{n})\right]^2}{\left[2(\frac{\sigma^2}{n})\sigma_0^2\right]/[\sigma_0^2/n]}\right\}.$$

This is, that the posterior p.d.f. of the parameter is normal with mean

$$\frac{y\sigma_0^2+\theta_0\sigma^2/n}{\sigma_0^2+\sigma^2/n} = \left(\frac{\sigma_0^2}{\sigma_0^2+\sigma^2/n}\right)y + \left(\frac{\sigma^2/n}{\sigma_0^2+\sigma^2/n}\right)\mu_0$$

and variance $\left(\frac{\sigma^2}{n}\right)\sigma_0^2/(\sigma_0^2+\frac{\sigma^2}{n})$.

The posterior p.d.f. $k(\mu|y)$ in Bayesian statistics contain all the information. And to get an interval estimate of μ, we would find two functions of y, for example u(y) and v(y), which would get us the following interval

$$\int_{u(y)}^{v(y)} k(\mu|y)d\theta = 1 - \alpha$$

An observed interval would lead to an interval estimate for the parameter in the way that the posterior probability of the parameter in the interval is $1 - \alpha$. In a case where the posterior p.d.f. of the *parameter is normal* we get the following interval

$$\frac{y\sigma_0^2 + \frac{\mu_0\sigma_0^2}{n}}{\sigma_0^2 + \frac{\sigma^2}{n}} \pm 1.96\sqrt{\frac{(\frac{\sigma^2}{n})\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}}}$$

The interval presented above serves as an interval estimate for μ, which has the posterior probability of 0.95. We suppose that Y is equal to the mean $\bar{X}$ of a random sample that follows a normal distribution. As we can see in this example, the Bayesian solution, which is the credible interval, approach a confidence interval as n increases. This also means that the prior gets less and less important when the n increases, and reason for that is while we have more information.

Contrasting the confidence and credible intervals is useful. First, the credible interval actually provides the researcher with an answer that researcher want, i.e., Pr(μ|data). However, this relies on assumptions regarding the prior. Although the Pr(data| μ) is less useful for the resercher, it does not rely on a choice of prior. There exist different procedures for assessing parameters with prior probabilities. Priors reflect beliefs about what value the parameter have in advance before we have observed the data. If the parameter is a constant, i.e. it has a uniform distribution, it is called noninformative prior. These should be avoided if any

information about the knowledge of the parameter exists beforehand (Hogg & Tanis, 2005). This information about the parameter may come from an experimentation or something similar. Since the posterior distribution consists of both the likelihood and the prior distribution this means that our choice of prior has an effect on the posterior distribution.

3. Method

3.1 Systematic review

In this paper an explorative literature review was conducted which is sort of systematic review. A systematic review is a form of literature review that with a specific methodology compiles all existing research in a restricted subject (Karolinska institutet, 2021). A relevant step in a systematic review is to make a comprehensive literature research that covers all the studies in the subject that is being examined, which usually is made in a very specific structured and organized way. The search in this paper will, because of time constraints, be made in a less time-consuming way. This means the searching is conducted in a semi-structured way. After the articles were found, the extraction of results from them began. In the following sections this will be explained, motivated, and discussed.

3.2 Searching for articles

The purpose of the study has been in mind while analyzing and picking out articles that could be used in the explorative literature review, as they examine how confidence intervals are interpreted. Both theoretical and empirical articles have been used in this paper. To present the working process as clearly as possible and make it easy for the writer to extract results to use in the explorative literature review, the articles were categorized and accordingly formed a data set.

For the search of articles Google Scholar was used as a database. In Table 1 *date* represents the time of when the search of the articles were conducted, the *search-words* represents the words used to find the articles and *quantity* is, how many articles that were chosen from the respective search-word presented.

Table 1. Search-words

Date	Search-words	Quantity
2021-11-20	Confidence intervals and statistical inference	3
2021-11-20	Confidence Intervals and misinterpretation	10
2021-11-20	Confidence Intervals and misunderstanding	2
2021-11-20	Interpretation of confidence intervals study	2
2021-11-20	Understanding confidence intervals study	3
2021-11-20	Students misconceptions of confidence intervals	6
2021-11-21	Researchers misconceptions about CIs	9
2021-11-21	Clinicians understanding of confidence intervals	1
2021-11-21	Teachers understanding of confidence intervals	2
2021-11-22	Confidence Intervals and misunderstanding -> Citation: Robust misinterpretation of confidence intervals	1
2021-11-22	Confidence interval interpret	1

Three days were set aside for researching articles. Eleven different search-words were used, and 40 articles were chosen to be used in the explorative literature review. Both the search-words and which articles to include were decided by the writer of this paper, a more detailed discussion about this will be presented later. The search-words were supposed to cover the purpose of the paper and to get articles that were relevant for the explorative literature review. Articles irrelevant for the purpose of the study were screened out. The decision of which articles to include or not will also be covered later on in greater detail.

3.3 Comments concerning the method

A review can be made in several alternative ways. The time restraint in this paper has foremost been the prime reason of the choices of some easier conducted solutions, instead of more time consuming alternatives. But, as for the most choices, it may affect the result in the study in different ways. We will below present some different alternatives that was *not* made in this study that may affect the results.

- A way to decompose a research question is to use PICO (**P**opulation, **I**ntervention, **C**omparison, **O**utcome) or PEO (**P**opulation, **E**xposure, **O**utcome). For quantitative issues PICO is most used and for qualitative ones PEO is (Karolinska institutet, 2021).
- At least three different databases are appropriate to use (Karolinska institutet, 2021). This may have affected the results, since we only used Google scholar and therefore can have missed some relevant articles for our study. This one was chosen since it is relatively comprehensive, but also we didn't have any specific research area and therefore were not in a position where we had to use a specific database.
- To get a structure in the final search for articles, the search-words can be organized in so called search blocks (Karolinska institutet, 2021). In this paper no search block was used, because of time restraints. The search-words used in the paper were chosen to cover articles that are relevant for our review.
- All articles should go through a quality review (Karolinska institutet, 2021). This can be used to identify flaws in the articles, that will be included in the review. In our paper this was not made, once again due to time restraints.

Some further comments are concerning the articles. The articles used in this study were chosen subjectively but have been treated objectively. The search-words to collect these articles were as mentioned before chosen by the writer. The search-words used were thought to cover articles that were relevant for the study's purpose. Of course, a lot of other search-words than those chosen in this study might have been equally valid, but because of time limits the mentioned search-words were chosen. Different search-words could probably cover other studies than those in this study, which could lead to different results. Furthermore, we get several articles by using the different search-words, whereas most of these articles are not included in the explorative literature review. Only the articles that seemed relevant were included, which mainly was based on each abstract and/or title. Just as the choice of search-words, this decision affected which articles were included in the explorative literature review and might affect the result of the study.

To make this study as transparent as possible, despite the comments above, we present the steps and working process for the reader.

3.4 Comments about the categories used to present the results

The misinterpretations from the articles have been divided into five different categories, and will be presented in five different tables, meaning each table is a category. This means that the articles that have similar misinterpretations about confidence intervals will be in a mutual table, and an article that include several misconceptions that belongs in different categories will be included in multiple categories. The choice regarding which categories to use was inspired by Morey et. al. (2016) where three of the five categories were taken. These three categories are *probability fallacies, precision fallacies* and *likelihood fallacies*. While going through the results we included two new categories after the articles were examined, these are *overlaps fallacies* and *miscellaneous fallacies*. The *overlaps fallacies* category was included since it was a relatively common misinterpretation category of confidence intervals. The rest of the articles, that couldn't be placed in any of the existing categories, were placed in the *miscellaneous fallacies*. This category got the role of including the misconceptions that didn't belong in any of the previous four categories as there were too many misinterpretations to make a new category to each one of them.

4. Results

As mentioned earlier, the results will be presented in five different tables, where each table consists of articles with misinterpretations in the same category. We will first present a short summary of *all the articles* in bar charts, before each table is presented. The summary will present the frequency of the *Misinterpretations* in each category, whether the article is *Empirical vs. theoretical*, the *Origin*, the R*esearch design* and *Target population*. In total we included 40 articles in this study, but 4 of them were excluded since they didn't present any actual misinterpretations about confidence intervals.

Figure 1.1 Misinterpretations

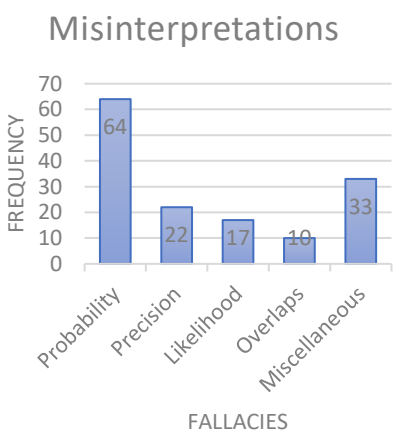


Figure 1.2 Empirical vs. theoretical

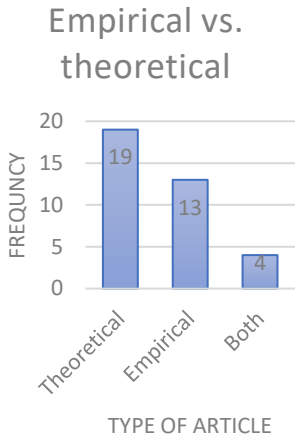


Figure 1.3 Origin

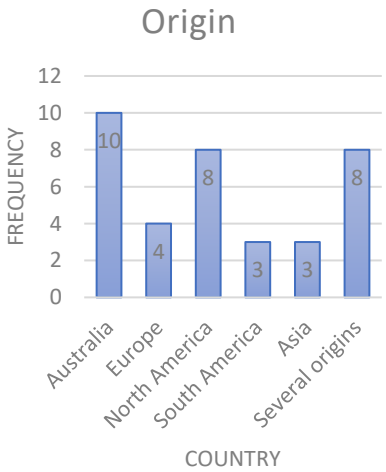


Figure 1.4 Research design

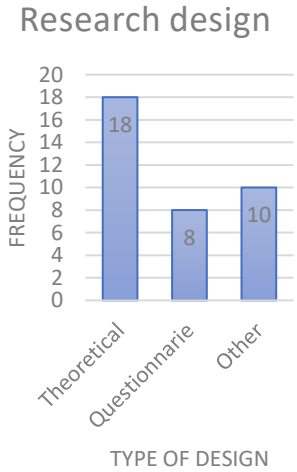
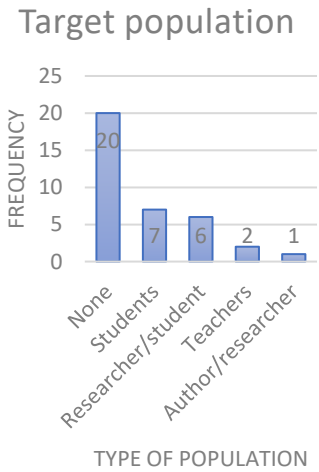


Figure 1.5 Target population



In Figure 1.1 above we present the frequency of misinterpretations in each individual category. Some articles have mentioned several misinterpretations of confidence interval, which means that in this study the same article may appear several times with different misinterpretations. Different articles may also mention the same misinterpretations, for example some articles use the same questionnaire as another article etc. In total we got 146 misinterpretations in this study, including repetition of misinterpretations. The category with most misinterpretations is category 1, which is *probability fallacies* where we have 64 misinterpretations from 29 articles (44 %). Then we have category 5 which is *miscellaneous fallacies* in which we have 33 misinterpretations from 18 articles (23 %). The third largest category is *precision fallacies* category 2, which have 22 misinterpretations from 8 articles (15 %). Category 3 is *likelihood fallacies and* have 17 misinterpretations from 8 articles (11 %), and the smallest one i.e., with least misinterpretations is *overlaps fallacies* i.e., category 4 with 10 misinterpretations from 8 articles (7 %).

In Figure 1.2 we present whether the article is theoretical, empirical or both. We have 19 articles that are theoretical, which is the largest group (53%), then we have 13 articles that are empirical which is the second largest group (36%) and finally the articles that are both empirical and theoretical, the smallest group, which consist of four articles (11%).

Figure 1.3 present the origin of all the articles. While there are relatively many origins to take in consideration from the articles, we have decided to present them as continents. Although, when we present the individual table, we will write out the specific country. We have six different categories; the largest category consists of ten articles from Australia (28%). Second, we have North America which consist of eight articles (22%). We have made one cateogry called *Several origins*, and as the name indicate it is articles that have authors from different origins. We have two articles that have Asia and Europe, three with Europe and North America, one with Europe and South America, one New Zealand and Germany and the last one of Australia, Europe and North America. In total this category consists of 8 articles (22%). Then we have Europe which consist of four articles (11%), and at last, with categories equally large we have South America and Asia with three articles each (17%).

The Figure 1.4 present the research design for all articles we have divided the categories in three. The largest category is theoretical which advocates that theoretical articles don't have a research design and consist of 18 articles (50%). The second largest category is *Other* and consist of 10 articles (28%). We do not write out all the different research designs since it would be too complex, but all the different kinds is included in the individual tables that will be presented later on in the text. The smallest category it called *Questionnaire* and consists of eight articles (22%).

The last Figure 1.5 consist of five different categories concerning target population. The largest category is the chart which don't have any specific target population, which consists of of 20 articles (55%). The reason why it is the largest one is because almost none of the theoretical articles have a target population, and the largest group in this study is theoretical articles. The next largest is the category with *students* consisting of seven articles (19%). The specific kind of students with different subjects and education levels are presented in the individual tables, but it would have been too complex to specify each type of students in the category above. The category called *researchers and students* has six articles (17 %). Same as for *students* we don't present information about what kind of researchers and students. The second smallest category with two articles (6%) is the one that have *teachers* as target population, what kind of teachers are presented in the individual tables. At last, the smallest category is *authors/researchers* and include only one article (3%) and just as the previous category, specific information will be in the individual tables.

4.1 Probability fallacies

In the individual tables the same articles can be presented several times and an article can occur in more than one table; this is because some articles have several kinds of misinterpretations. Each table that is presented have seven columns, which will be stated in chronologic order - the authors of the article; the origin of the article; a concise description of the purpose of the study; whether the article is theoretical and/or empirical; the research design (how the data was collected); target population; the misinterpretation of the confidence interval.

The *first table* presented in this study include the articles in which the misinterpretation of the confidence interval is related to probability. This misinterpretation seems to be the most common misinterpretation of confidence intervals in the articles about probability fallacies. This problem is a severe, widespread and clear misinterpretation. The interpretation is incorrect since one according to the frequentist approach cannot interpret a confidence interval this way. The probability regarding the interval to have or not have the value in percentage cannot be explained because as mentioned in the theory, the interval either contain the value or not, meaning it is either 0 % or 100 %. There are also some other misinterpretations that are not directly connected to this problem but that has to do with probability. The Table 1.6 is called *probability fallacies* and as we mentioned earlier this table include 29 articles with different misinterpretations and from these articles, 64 misconceptions could be extracted.

Table 1.6. Probability fallacies

Authors	Origin	Purpose	Theoretical and/or empirical	Research design	Target population	Misinterpretation of confidence interval
Morey, Hoekstra, Rouder, Lee and Wagenmakers (2016)	Europe and USA	Bring up examples of confidence intervals with different properties aswell as showing why these are incorrect	Theoretical	Theoretical	Folk (no specific target population)	If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observed interval contains the true value is also X%; or, alternatively, we can have X% confidence that the observed interval contains the true value.
Greenland, Senn, Rothman, Carlin, Poole, Goodman, and Altman (2016)	USA, Luxembourg, Australia, and UK	To form a discussion of the basic statistics and through this provide definitions. Also, present misinterpretations of p-values, CIs and power	Theoretical	Theoretical	NA	The specific 95 % confidence interval presented by a study has a 95 % chance of containing the true effect size
Reaburn (2014)	Australia	To learn of students' difficulties of understanding CIs and beliefs about them. This is then used to develop improved teaching programs	Empirical	Action research, where the researcher was the lecturer of the unit	Students from one-semester introductory statistics	The mean is within the confidence interval 95% of the time
Foster (2014)	UK	Examine the problem of the frequent misinterpretation of the nature of confidence intervals, as an aspect of the learning of mathematical definitions. This while noticing the tension between parroting mathematically rigorous (although uninternalized) statements and expressing imperfect but developing understandings	Empirical	Questionnaire	High school and college teachers of mathematics	The probability that the true population mean is within the confidence interval is 95%
Miller and Ulrich (2016)	New Zealand and Germany	To question Hoekstra et al's study (2014) where they got the result that first year & master students and researchers "have no reliable knowledge about the correct interpretation of CIs"	Theoretical	Theoretical	First year, master students & researchers	There is a 95 % probability that the true mean lies between 0.1 and 0.4
Canal and Gutiérrez (2010)	Colombia	They take notion from a former study where the authors tries to see what a sample of experts and university students actually understand by CIs	Empirical	Questionnaire	Engineering and Business Administration undergraduate students at several universities in Cali and experts in the same subjects	The probability that the interval includes the sample mean is 95%
Lyu, Xu, Zhao, Zuo and Hu (2020)	China and Europa	To fill the research gap, since existing surveys on CI understanding mostly fall under psychology and biomedical research, while data from other disciplines are rare. They also inspect that the confidence of researchers while constructing judgments remains fairly unclear	Empirical	Questionnaire	Respondents from different backgrounds (respondents' academic background was based on the degree they awarded in China)	In a 95% CI, there is a 95% probability that the true mean lies within the interval
Canal and Gutiérrez (2010)	Colombia	They take notion from a former study where the authors tries to see what a sample of experts and university students actually understand by CIs	Empirical	Questionnarie	Engineering and Business Administration undergraduate students of several universities in Cali and experts in the same subjects	The probability that the interval includes the population mean is 95%
Crooks, Bartel and Alibali (2019)	USA	To assess conceptual knowledge of CIs in undergraduate as well as graduate psychology students	Empirical	Lab setting	Psychology graduate students and undergraduate students	A 95% confidence interval is the interval for which you are 95% confident that the population mean falls within it.
Miller and Ulrich (2016)	New Zealand and Germany	To question Hoekstra et al's study (2014) where they got the result that first year & master students and researchers "have no reliable knowledge about the correct interpretation of CIs"	Theoretical	Theoretical	First year, master students & researchers	We can be 95 % confident that the true mean lies between 0.1 and 0.4

Gutiérrez and Yáñez (2009)	Colombia	Research what a sample of experts and university students actually understand about CIs	Empirical	Questionnaire	Experts (professionals devoted to statistics, or its teaching) and senior statistics students	95% of weights are between 42 and 48 pounds
Gutiérrez and Yáñez (2009)	Colombia	Research what a sample of experts and university students actually understand about CIs	Empirical	Questionnaire	Experts (professionals devoted to statistics, or its teaching) and senior statistics students	The probability that the interval includes the sample mean is 95%
Tan and Tan (2010)	Singapore	Discussing the correct interpretation of CIs, and highlighting some common misunderstandings	Theoretical	Theoretical	NA	A 95% CI (A to B), there is a 95% probability that the true population mean lies between A and B
Finch and Cumming (2009)	Australia	To discuss the meaning and interpretation of CIs in single studies, to review and integrate	Theoretical	Theoretical	NA	There is a 95% chance that the actual CI includes the true parameter value
Naimi and Whitcomb (2020)	USA	To bring up and demonstrate key properties of frequentist CIs, to elucidate interpretations and explain common misunderstandings	Theoretical	Theoretical	NA	95% confident the true values lies between 1.02 and 1.31 (the interval).
Miller and Ulrich (2016)	New Zealand and Germany	To question Hoekstra et al's study (2014) where they got the result that first year & master students and researchers "have no reliable knowledge about the correct interpretation of CIs"	Theoretical	Theoretical	First year, master students & researchers	If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4
Hespanhol, Vallio, Costa and Saragiotto (2019)	São Paul and Amsterdam	To discuss CIs around effect estimates, understand CIs estimation in both frequentist and Bayesian approaches, and to interpret uncertainty measures	Both	Masterclass	Clinicians	There is a 95% probability that the true (unknown) effect estimate lies within the 95% CI.
O'Brien and Yi (2016)	Canada	To bring up and describe some basic principles of CIs and their interpretation	Theoretical	Theoretical	NA	95% CI mean that 95% of the population data falls within the CI
Callaert (2007)	Belgium	Address statistical reasoning at a first course in statistics level, in higher education	Theoretical	Theoretical	Students	The population mean is between 20 and 30 90% of the time.
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of CIs	Empirical	Questionnaire	Researchers and students—all in the field of psychology	There is a 95 % probability that the true mean lies between 0.1 and 0.4.
Lyu, Peng and Hu (2018)	China and Germany	To introduce raw data that is available for anyone interested in examining how students aswell as researchers misinterpret inter alia CIs, and how NHST and CIs affect the interpretation of study- or research results	Empirical	Questionnaire	Psychological researchers in different positions in related fields of psychology.	There is a 95 % probability that the true mean lies between 0.1 and 0.4.
Karlen (2002)	Canada	To clarify the issue of when confidence intervals are empty or rule out parameter values for which the experiment is insensitive. This is made through exploring the relation between CIs and credible intervals	Theoretical	Theoretical	NA	Misinterpreting classical confidence intervals as credible interval
Whitener (1990)	USA	To investigate the interpreting problem of CIs (and credibility intervals) for meta analyses, and suggest how to construct and interpret these.	Theoretical	Theoretical	NA	Interpret confidence intervals as credibility intervals
Greenland, Senn, Rothman, Carlin, Poole, Goodman and Altman (2016)	USA, Luxembourg, Australia, and UK	To form a discussion of the basic statistics and through this provide definitions. Also, present misinterpretations of p-values, CIs and power	Theoretical	Theoretical	NA	An observed 95 % confidence interval predicts that 95 % of the estimates from future studies will fall inside the observed interval.
Lyu, Xu, Zhao, Zuo and Hu (2020)	China and Europa	To fill the research gap, since existing surveys on CI understanding mostly fall under psychology and biomedical research, while data from other disciplines are rare. They also inspect that the confidence of researchers	Empirical	Questionnaire	Respondents from different backgrounds (respondents' academic background was based on the degree	If we were to repeat an experiment over and over and make a 95% CI, then 95% of the time the true mean falls within the interval

		while constructing judgments remains fairly unclear			they awarded in China)	
García-Pérez and Alcalá-Quintana (2016)	Spain	To re-analyze Hoekstra et al's study of interpretation of CIs, and discuss the two interpretations of CIs and why misinterpretation can't be inferred from endorsement of some of the items	Both	(replication) questionnaire	First year and master students	There is a 95 % probability that the true mean lies between 0.1 and 0.4
García-Pérez and Alcalá-Quintana (2016)	Spain	To re-analyze Hoekstra et al's study of interpretation of CIs, and discuss the two interpretations of CIs and why misinterpretation can't be inferred from endorsement of some of the items	Both	(replication) questionnaire	First year and master students	We can be 95 % confident that the true mean lies between 0.1 and 0.4
Crooks, Bartel and Alibali (2019)	USA	To assess conceptual knowledge of CIs in undergraduate as well as graduate psychology students	Empirical	Lab setting	Psychology graduate students and undergraduate students	If you were to conduct an infinite number of experiments exactly like the original experiment, a 95% confidence interval would contain 95% of the sample means from these experiments.
Morey, Hoekstra, Rouder et al. (2016)	UK, Netherlands, USA	To answer Miller and Ulrich's questioning of their previous study from 2014. Among other questionings, Morey, Hoekstra et al. mean that Miller and Ulrich alternative interpretations, even though correct, can't be deemed acceptable renderings of the questions in the survey due to the reference class problem	Theoretical	Theoretical	Researchers and students—all in the field of psychology	There is a 95 % probability that the true mean lies between 0.1 and 0.4.
Morey, Hoekstra, Rouder et al. (2016)	UK, Netherlands, USA	To answer Miller and Ulrich's questioning of their previous study from 2014. Among other questionings, Morey, Hoekstra et al. mean that Miller and Ulrich alternative interpretations, even though correct, can't be deemed acceptable renderings of the questions in the survey due to the reference class problem	Theoretical	Theoretical	Researchers and students—all in the field of psychology	We can be 95 % confident that the true mean lies between 0.1 and 0.4.
Morey, Hoekstra, Rouder et al. (2016)	UK, Netherlands, USA	To answer Miller and Ulrich's questioning of their previous study from 2014. Among other questionings, Morey, Hoekstra et al. mean that Miller and Ulrich alternative interpretations, even though correct, can't be deemed acceptable renderings of the questions in the survey due to the reference class problem	Theoretical	Theoretical	Researchers and students—all in the field of psychology	If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4
Kalinowski, Lai and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	The Confidence Level Misconception
García-Pérez and Alcalá-Quintana (2016)	Spain	To re-analyze Hoekstra et al's study of interpretation of CIs, and discuss the two interpretations of CIs and why misinterpretation can't be inferred from endorsement of some of the items	Both	(replication) questionnaire	First year and master students	If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4
Cumming and Maillardet (2006)	Australia	To shed light upon the fact that the percentage of future replication means captured by a particular CI varies distinctly, depending on where in relation to the population mean that CI falls. Also investigate the distribution of the problem percentage for sigma known and unknown,	Theoretical	Theoretical	NA	about C% of future replication means will fall within an original C% CI. (ex a 95% CI will include about 95% of future replication means)

		for various sample sizes, and for robust CIs				
Fidler and Cumming (2005)	Australia	To help ease some misinterpretations of results to provide more accurate conclusions, demonstrate particular types of misconceptions and present figures and simulations that should lead to fewer misconceptions.	Both	Low powered study	Final year undergraduates and postgraduate environmental science students	a C% CI will capture about C% of replication means
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of CIs	Empirical	Questionnaire	Researchers and students—all in the field of psychology	Assign probabilities to parameters or hypotheses, which is something that in the frequentist framework is not allowed.
Callaert (2007)	Belgium	Address statistical reasoning at a first course in statistics level, in higher education	Theoretical	Theoretical	Students	Another frequent type of error shows up when the student "mechanically" knows that he should switch from probability to confidence but where he keeps writing inequalities about a fixed but unknown population proportion pi
Reaburn (2014)	Australia	To learn of students' difficulties of understanding CIs and beliefs about them. This is then used to develop improved teaching programs	Empirical	Cyclical questionnaire	Students from one-semester introductory statistics	95% of the sample means are included in the confidence interval" or "95% of the population will be in the stated interval
Reaburn (2014)	Australia	To learn of students' difficulties of understanding CIs and beliefs about them. This is then used to develop improved teaching programs	Empirical	Cyclical questionnaire	Students from one-semester introductory statistics	95% of the population visited a doctor between seven and eleven times.
Foster (2014)	UK	Examine the problem of the frequent misinterpretation of the nature of confidence intervals, as an aspect of the learning of mathematical definitions. This while noticing the tension between parroting mathematically rigorous (although uninternalized) statements and expressing imperfect but developing understandings.	Empirical	Questionnaire	High school and college teachers of mathematics	About 95% of the time the true population mean lies inside the confidence interval.
Foster (2014)	UK	Examine the problem of the frequent misinterpretation of the nature of confidence intervals, as an aspect of the learning of mathematical definitions. This while noticing the tension between parroting mathematically rigorous (although uninternalized) statements and expressing imperfect but developing understandings.	Empirical	Questionnaire	High school and college teachers of mathematics	I'm 95% sure that the confidence interval contains the true population mean.
Foster (2014)	UK	Examine the problem of the frequent misinterpretation of the nature of confidence intervals, as an aspect of the learning of mathematical definitions. This while noticing the tension between parroting mathematically rigorous (although uninternalized) statements and expressing imperfect but developing understandings.	Empirical	Questionnaire	High school and college teachers of mathematics	There is a 95% chance that the true population mean is inside the confidence interval.
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of CIs	Empirical	Questionnaire	Researchers and students—all in the field of psychology	We can be 95 % confident that the true mean lies between 0.1 and 0.4
Lyu, Peng and Hu (2018)	China and Germany	To introduce raw data that is available for anyone interested in examining how students aswell as researchers misinterpret inter alia CIs, and how NHST and CIs affect the interpretation of study- or research results	Empirical	Questionnaire	Psychological researchers in different positions in related fields of psychology.	We can be 95 % confident that the true mean lies between 0.1 and 0.4
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of CIs	Empirical	Questionnaire	Researchers and students—all in the field of psychology	If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4.

Lyu, Peng and Hu (2018)	China and Germany	To introduce raw data that is available for anyone interested in examining how students aswell as researchers misinterpret inter alia CIs, and how NHST and CIs affect the interpretation of study- or research results	Empirical	Questionnaire	Psychological researchers in different positions in related fields of psychology.	If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4.
Canal and Gutiérrez (2010)	Colombia	They take notion from a former study where the authors tries to see what a sample of experts and university students actually understand by CIs	Empirical	Questionnarie	Engineering and Business Administration undergraduate students of several universities in Cali and experts in the same subjects	In a 95% CI, 95% of the population (weights) are within the interval
Canal and Gutiérrez (2010)	Colombia	They take notion from a former study where the authors tries to see what a sample of experts and university students actually understand by CIs	Empirical	Questionnarie	Engineering and Business Administration undergraduate students of several universities in Cali and experts in the same subjects	In a 95% CI, most of the population (weights) is within the interval
Crooks, Bartel and Alibali (2019)	USA	To assess conceptual knowledge of CIs in undergraduate as well as graduate psychology students	Empirical	Lab setting	Psychology graduate students and undergraduate students	A 95% confidence interval indicates that there is a 95% chance that the sample mean equals the population mean.
Eliason (2018)	USA	To analyze pre-service teachers' conception about CIs	Empirical	Pretest and interviews	Students in mathematics education course for pre-service secondary teachers focused on the teaching and learning of statistics and probability	The Range Spanned By a Confidence Interval with a 95% Confidence Level is the Range that Contains 95% of the Population Data.
Gutiérrez and Yáñez (2009)	Colombia	Research what a sample of experts and university students actually understand about CIs	Empirical	Questionnaire	Experts (professionals devoted to statistics, or its teaching) and senior statistics students	Most of weights are between 42 and 48 pounds
Gutiérrez and Yáñez (2009)	Colombia	Research what a sample of experts and university students actually understand about CIs	Empirical	Questionnaire	Experts (professionals devoted to statistics, or its teaching) and senior statistics students	The probability that the interval includes the population mean is 95%
Navruz and Delen (2014)	USA	To show that CIs are useful and valuable in research studies when used in the correct form and with correct interpretations	Theoretical	Theoretical	NA	there is 95% probability that our 95% CI captures the true population parameter
Hazra (2017)	India	To instruct how to use CIs confidently	Theoretical	Theoretical	NA	95% of the sample data lie within that interval.
O'Brien and Yi (2016)	Canada	To bring up and describe some basic principles of CIs and their interpretation	Theoretical	Theoretical	NA	we are "95% confident" that the true mean lies within the interval
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of Cis	Empirical	Questionnaire	Researchers and students—all in the field of psychology	The probability that the true mean is greater than 0 is at least 95 %
Lyu, Peng and Hu (2018)	China and Germany	To introduce raw data that is available for anyone interested in examining how students aswell as researchers misinterpret inter alia CIs, and how NHST and CIs affect the interpretation of study- or research results	Empirical	Questionnaire	Psychological researchers in different positions in related fields of psychology.	The probability that the true mean is greater than 0 is at least 95 %
García-Pérez and Alcalá-Quintana (2016)	Spain	To re-analyze Hoekstra et al's study of interpretation of CIs, and discuss the two interpretations of CIs and why misinterpretation can't be inferred from endorsement of some of the items	Both	(replication) questionnaire	First year and master students	The probability that the true mean is greater than 0 is at least 95 %
Miller and Ulrich (2016)	New Zealand and Germany	To question Hoekstra et al's study (2014) where they got the result that first year & master students and researchers "have no reliable	Theoretical	Theoretical	First year, master students & researchers	The probability that the true mean is greater than 0 is at least 95 %

		knowledge about the correct interpretation of CIs"				
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of CIs	Empirical	Questionnaire	Researchers and students—all in the field of psychology	The probability that the true mean equals 0 is smaller than 5 %
Lyu, Peng and Hu (2018)	China and Germany	To introduce raw data that is available for anyone interested in examining how students aswell as researchers misinterpret inter alia CIs, and how NHST and CIs affect the interpretation of study- or research results	Empirical	Questionnaire	Psychological researchers in different positions in related fields of psychology.	The probability that the true mean equals 0 is smaller than 5 %
García-Pérez and Alcalá-Quintana (2016)	Spain	To re-analyze Hoekstra et al's study of interpretation of CIs, and discuss the two interpretations of CIs and why misinterpretation can't be inferred from endorsement of some of the items	Both	(replication) questionnaire	First year and master students	The probability that the true mean equals 0 is smaller than 5 %
Miller and Ulrich (2016)	New Zealand and Germany	To question Hoekstra et al's study where they got the result that first year & master students and researchers "have no reliable knowledge about the correct interpretation of CIs"	Theoretical	Theoretical	First year, master students & researchers	The probability that the true mean equals 0 is smaller than 5 %
Naimi and Whitcomb (2020)	USA	To bring up and demonstrate key properties of frequentist CIs, to elucidate interpretations and explain common misunderstandings	Theoretical	Theoretical	NA	there is a 95% probability that the true odds ratio lies between 1.02 and 1.31 in this example

Almost all citations use the same formulation for the same problem. Either they use specific numbers for the range, which is the same as writing the interval or they explicit write interval. Most of them phrase a "95 % confident", "95% of the data or population", "most of the data" or "95 % probability". In the table we have fourteen theoretical, twelve empirical and three articles that are both. As mentioned earlier the articles that doesn't have a specific population target is called NA. Almost all theoretical articles don't have any population target except four, where one has students, the other one has folks, and two other articles has researchers and students. All empirical articles have target populations, and as we can see in the table we have a mix within the articles. Some examples of usual target populations from the table above are students, researchers, experts, and teachers whom comes from different subjects and education levels. In the articles that are empirical, *questionnaires* are the most frequent research design. It seems to be in the end of the table that nine articles are not as similar to each other compared to the ones present before them.

The misinterpretations that are mentioned by more than one article is presented below:

We have these two misinterpretations:

- *"The probability that the interval includes the sample mean is 95%"*
- *"The probability that the interval includes the population mean is 95%"*

These are mentioned both by Canal and Gutiérrez (2010) *and* Gutiérrez and Yáñez (2009).

Another misinterpretation is:

- *"There is a 95 % probability that the true mean lies between 0.1 and 0.4"*

This one is mentioned in four different articles, which are those by Hoekstra, Morey, Rouder et. al. (2014); Lyu, Peng and Hu (2018); Morey Hoekstra, Rouder et al. (2016); Miller and Ulrich (2016); García-Pérez and Alcalá-Quintana (2016).

The misinterpretation:

- *"We can be 95 % confident that the true mean lies between 0.1 and 0.4"*

Is mentioned by García-Pérez and Alcalá-Quintana (2016); Morey, Hoekstra, Rouder et al. (2016); Hoekstra, Morey, Rouder et. al. (2014); Miller and Ulrich (2016); Lyu, Peng and Hu (2018).

Then we have the following misinterpretation that is mentioned by several articles:

- *"If we were to repeat the experiment over and over, then 95 % of the time the true mean falls between 0.1 and 0.4"*

This one is mentioned by Morey, Hoekstra, Rouder et al. (2016); García-Pérez and Alcalá-Quintana (2016); Hoekstra, Morey, Rouder et. al. (2014); Miller and Ulrich (2016); Lyu, Peng and Hu (2018).

Then we have:

- "*The probability that the true mean is greater than 0 is at least* 95 %"

Which is mentioned by Hoekstra, Morey, Rouder et. al. (2014); Lyu, Peng and Hu (2018); García-Pérez and Alcalá-Quintana (2016); Miller and Ulrich (2016).

Then we have the following misinterpretation which is mentioned by the same authors as the misinterpretation above:

- "*The probability that the true mean equals 0 is smaller than 5 %"*

In the article of Morey et al. (2016) i.e., the one we got the inspiration to use three of our five fallacies, they argue that advocacy of confidence interval is not based on a principled understanding of confidence interval theory but rather on a folk understanding. They also present three fallacies which underlie the folk understanding of confidence intervals and place these in the philosophical and historical context of proper confidence interval theory.

Masson and Loftus (2003) stated, with regard to 95% confidence intervals, *"in the absence of any other information, there is a 95 % probability that the obtained confidence interval includes the population mean."* In 2014 Cumming wrote *"[w]e can be 95 % confident that our interval includes [the parameter] and can think of the lower and upper limits as likely lower and upper bounds for [the parameter]."* These presented interpretations of confidence intervals are not right, and we refer to the mistake the authors made as "Fundamental Confidence Fallacy" (Morey et al., 2016).

The Fundamental Confidence Fallacy (FCF) is the first misinterpretation presented in table 1.6 above and is explained as followed:

- "*If the probability that a random interval contains the true value is X%, then the plausibility or probability that a particular observed interval contains the true value is also X%; or, alternatively, we can have X% confidence that the observed interval contains the true value*."

What we know before observing the data - that the confidence interval has a fixed chance of having the true value, given that the necessary assumptions hold - and what we know after is easily confused. In the frequentist confidence interval theory, the interval either contains the true value or not. The theory says nothing about the probability that a particular, observed confidence interval contains the true value (Hoekstra et al., 2014).

As we can see in Table 1.6 a lot of articles mention a faulty interpretation about what the confidence interval contains, which is the same problem as the Fundamental Confidence Fallacy. In the paper *Statistical tests, P-values, confidence intervals, and power: a guide to misinterpretations* (2016) the researchers examine several common misinterpretations of confidence intervals and one of them, which is the same problem as CFC, is one of the misinterpretations presented in the table. The misinterpretation is as followed:

- *"The specific 95 % confidence interval presented by a study has a 95 % chance of containing the true effect size"*

As mentioned previously, a reported confidence interval is a range between two numbers. Whether an observed interval contains the true effect is either 100 % which means that the true effect is within the interval or 0%, which means that the true effect is not in the interval. What the 95 % refers to is only how often 95 % confidence intervals computed from many studies would contain the true size if all the assumptions used to compute the interval were correct.

Still this interpretation of an interval, 95% probability of containing the true value, can be computed but require both the assumptions that we needed when we conducted a confidence interval and further assumptions that can be summarized in what can be called a prior distribution. This leads to intervals that are called Bayesian posterior or credible intervals (Greenland et al., 2016).

In the theory common misinterpretation of a 95 % confidence interval is this:

- *"there is a 95% probability that the true (unknown) effect estimate lies within the 95% CI."*

As we can see, this misinterpretation is very similar to the misinterpretations from the table. And as mentioned earlier, with a frequentist approach, which confidence intervals are from, this interpretation is not accurate. While the population parameter is treated as fixed, this means the value is either inside the interval with 100 % probability or outside the interval

which is a 0 % probability. So, we cannot explain an observed interval as a certain % of containing the value. Although, in the Bayesian approach the misconception could be considered a correct interpretation. Confidence intervals and credible intervals are different, and they should therefore be interpreted differently, which was also shown under Theory 2.2. The citation from Karlen (2002) and Whitener (1990) in Table 1.6, show that confidence intervals are interpreted as credible intervals in their articles.

Another common misinterpretation mentioned in the article *Statistical tests, P-values, confidence intervals, and power: a guide to misinterpretations* (2016) is:

- *"An observed 95 % confidence interval predicts that 95 % of the estimates from future studies will fall inside the observed interval."*

There are several ways this statement is wrong. 95% is not how frequently that one interval that is presented will contain future estimates under the model, it is how frequently some other unobserved intervals will contain the true effect. As mentioned earlier, an observed interval contains the true effect or not. The 95% is referring only to how often 95% confidence intervals computed from very many studies would contain the true effect if all the assumptions used to compute the intervals were correct.

In the *Bayesian approach* an interpretation of a 95% credible interval would be:

"there is a 95% probability that the true (unknown) effect estimate would lie within the interval, given the evidence provided by the observed data."

A common *misinterpretation* of the 95% confidence interval is the following:

"there is a 95% probability that the true (unknown) effect estimate lies within the 95% CI."

In a *frequentist approach* this interpretation is not accurate. The population parameter is treated as a fixed (unknown) value, meaning the fixed value is either inside the interval with 100 % probability or outside the interval which is a 0 % probability (Hespanhol et al., 2019). In the Bayesian approach we can describe the parameters with probability distribution

because of the fact that the parameters in this approach is treated as random variables (Lesaffre & Lawson, 2012; O'Neill, 2012).

In the end of table 1.6 we have some misinterpretations that differ from the other in their formulation which are the following:

- "*The probability that the true mean is greater than 0 is at least 95 %*"
- "*The probability that the true mean equals 0 is smaller than 5 %*"
- "*there is a 95% probability that the true odds ratio lies between 1.02 and 1.31 in this example*"

The first two misinterpretations is about which probability the true mean has. As we mentioned before the parameters in the Bayesian approach can be described with probability distribution but that is not possible in the frequentist approach since the parameters are not random. The third misinterpretation above is relatively similar to the previous misinterpretations but differ regarding that it has "the true odds ratio" instead of, for example, just mean. Same as before, this is a wrong interpretation because either the parameter lies in the interval or not.

4.2 Precision fallacies

The second table include articles where the misinterpretation is regarding precision concerning the confidence interval and is called *precision fallacies*. In Table 1.7 we present 8 articles with 22 misinterpretations.

Table 1.7. Precision fallacies

Authors	Origin	Purpose	Theoretical and/or empirical	Research design	Target population	Misinterpretation of confidence interval
Morey, Hoekstra, Rouder, Lee and Wagenmakers (2016)	Europe and USA	Bring up examples of confidence intervals with different properties aswell as showing why these are incorrect	Theoretical	Theoretical	Folk (no specific target population)	The width of a confidence interval indicates the precision of our knowledge about the parameter. Narrow confidence intervals correspond to precise knowledge, while wide confidence errors correspond to imprecise knowledge
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	A 95% CI is double the length of a 50% CI
Fidler (2006)	Australia	To discuss reform efforts in statistical reporting regarding psycology, medicine and ecology, and through data on students understanding of CIs, explain how to improve statistics education in psycology	Theoretical	Theoretical	Students in psycology (and medicine, ecology)	90% CI wider than 95% CI (for same data)
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an educational tool that could be used to confront particularly misconceptions about CI distributions	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology, Medicine and other science disciplines	90% CI is wider than a 95% CI for the same data.
Fidler (2006)	Australia	To discuss reform efforts in statistical reporting regarding psycology, medicine and ecology, and through data on students understanding of CIs, explain how to improve statistics education in psycology	Theoretical	Theoretical	Students in psycology (and medicine, ecology)	CI width increases with sample size
Fidler (2006)	Australia	To discuss reform efforts in statistical reporting regarding psycology, medicine and ecology, and through data on students understanding of CIs, explain how to improve statistics education in psycology	Theoretical	Theoretical	Students in psycology (and medicine, ecology)	Unsure of relationship between CI width and sample size
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	As N increases, length does not change
Fidler (2006)	Australia	To discuss reform efforts in statistical reporting regarding psycology, medicine and ecology, and through data on students understanding of CIs, explain how to improve statistics education in psycology	Theoretical	Theoretical	Students in psycology (and medicine, ecology)	CI width unaffected by sample size
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	As N decreases C% level decreases as N increases C% level increases
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an educational tool that could be used to confront particularly misconceptions about CI distributions	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology, Medicine and other science disciplines	CI width increases as N increases
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology,	Change in N has little effect on CI width.

		educational tool that could be used to confront particularly misconceptions about CI distributions			Medicine and other science disciplines	
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	As N increases, length increases, and as N decreases, length decreases
Canal and Ruiz (2015)	Colombia	To collect information about and better understand the mental structures and mechanisms that students develop around confidence intervals	Empirical	Interview and questionnaire	Students	larger sample size means greater accuracy in the estimation and therefore the probability that the interval contains the population mean is greater
Canal and Ruiz (2015)	Colombia	To collect information about and better understand the mental structures and mechanisms that students develop around confidence intervals	Empirical	Interview and questionnaire	Students	a greater level of confidence, gives greater (value) accuracy
Eliason (2018)	USA	To analyze pre-service teachers' conception about CIs	Empirical	Pretest and interviews	Students in mathematics education course for pre-service secondary teachers focused on the teaching and learning of statistics and probability	Statisticians Must Take Multiple Samples in Order to Compute a Reliable Confidence Interval
Fidler and Cumming (2005)	Australia	To help ease some misinterpretations of results to provide more accurate conclusions, demonstrate particular types of misconceptions and present figures and simulations that should lead to fewer misconceptions.	Both	Low powered study	Final year undergraduates and postgraduate environmental science students	CI width would increase with increases in sample size
Fidler and Cumming (2005)	Australia	To help ease some misinterpretations of results to provide more accurate conclusions, demonstrate particular types of misconceptions and present figures and simulations that should lead to fewer misconceptions.	Both	Low powered study	Final year undergraduates and postgraduate environmental science students	CI width was unaffected by sample size
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an educational tool that could be used to confront particularly misconceptions about CI distributions	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology, Medicine and other science disciplines	the relationship between the width of a confidence interval and the confidence level itself
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	As C% level increases, length decreases
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	As confidence level increases, CI width decreases (for the same data)
Canal and Ruiz (2015)	Colombia	To collect information about and better understand the mental structures and mechanisms that students develop around confidence intervals	Empirical	Interview and questionnaire	Students	If the confidence level is higher in the range, its values will be very close to the mean, therefore, there will be a minimum error in calculating the population mean
Naimi and Whitcomb (2020)	USA	To bring up and demonstrate key properties of frequentist CIs, to elucidate interpretations and explain common misunderstandings	Theoretical	Theoretical	NA	the upper and lower bounds of the interval do provide information regarding confidence

In Table 1.7 we have three theoretical, four empirical and one article that is both empirical and theoretical. The empirical articles have used interviews, surveys or/and questionnaires as type of research design. All articles except two have students as target population, but in different subjects and levels of education. All other articles have folks as target population, except for one, that has no specific target population.

There are some misinterpretations from different articles that are mentioned in the table more than once which are the following:

In Fidler (2006) is the misinterpretation:

- *"90% CI wider than 95% CI (for same data)"*

And in Kalinowski (2010) there is:

- *"90% CI is wider than a 95% CI for the same data*"

As we can see there is only small differences in formulation between these misinterpretations.

We also have these misinterpretations that are almost identical, with just a small difference in formulation.

- *"CI width unaffected by sample size"*, Fidler (2006)
- *"CI width was unaffected by sample size"*, Fidler and Cumming (2005)

It seems that some articles are more like each other than others. Although all articles are about misinterpretations of confidence intervals, that somehow are concerning precision. Relatively many of the articles are about the misinterpretations concerning the relationship between sample size (also called N) and the width, length, or range of the confidence interval. As we explain in the theory, we have the same confidence coefficient $1 - \alpha$ independently of an increase in N. The increase in N leads to a shorter confidence interval which gives us more reliance in our estimated parameter $\bar{x}$, which is an estimate of μ. The length of a confidence interval can, by decreasing the confidence coefficient, also be shortened with a fixed sample size N but this results in a shorter confidence interval with loss in confidence.

There are also quite a lot of articles that have misinterpretations concerning what is inside the confidence interval (the confidence level itself) and the relationship to the width of the interval. For example, the first misinterpretation in the Table 1.7 from Morey et al. (2016). The misinterpretation is explained as followed:

- *"The width of a confidence interval indicates the precision of our knowledge about the parameter. Narrow confidence intervals correspond to precise knowledge, while wide confidence errors correspond to imprecise knowledge."*

This second fallacy is from the article by which we got the inspiration to use precision as a category. The fallacy is called *The Precision fallacy*. The precision of an estimate and the size of a confidence interval has no necessary connection. We cannot interpret confidence intervals as indicating the precision of our estimate. Confidence intervals are often claimed by proponents to be more useful for evaluating the precision with which a parameter can be estimated. In the article by Morey et al. (2016) they mentioned two examples. The first one is from Cumming (2014) who wrote:

- "*[l]ong confidence intervals (CIs) will soon let us know if our experiment is weak and can give only imprecise estimates*"

The second one is Young and Lewis (1997) where they stated that:

- "*[i]t is important to know how precisely the point estimate represents the true difference between the groups. The width of the CI gives us information on the precision of the point estimate*"

Even though all misinterpretations in this table are similar we can see that the results are not as clear as the table with probability fallacies. For example, regarding the misinterpretations that are about the relationship between the width of the confidence interval and sample size, N. Below, five different misinterpretations from four different articles are presented, that is from our table above.

- *"CI width would increase with increases in sample size"* - Fidler and Cumming (2005)
- *"As N increases, length increases, and as N decreases, length decreases"* - Kalinowski, Lai, and Cumming (2018)
- *"CI width unaffected by sample size"* - Fidler (2006)
- *"As N increases, length does not change"* - Kalinowski, Lai, and Cumming (2018)
- *"Change in N has little effect on CI width"* - Kalinowski (2010)

It seems to be more dimensions of misinterpretations regarding precision fallacies compared to probability fallacies. Even though the misinterpretations are different we can summarize it as difficulties to understand the relationship between width and sample size.

4.3 Likelihood fallacies

The third table will include articles where the confidence interval has been misinterpreted because of likelihood related reasons and is called *likelihood fallacies*. In Table 1.8 the articles are presented in the same way as the previous tables. Here, we have included 8 articles and extracted 17 misinterpretations.

Table 1.8. Likelihood fallacies.

Authors	Origin	Purpose	Theoretical and/or empirical	Research design	Target population	Misinterpretation of confidence interval
Morey, Hoekstra, Rouder, Lee and Wagenmakers (2016)	Europe and USA	Bring up examples of confidence intervals with different properties aswell as showing why these are incorrect	Theoretical	Theoretical	Folk (no specific target population)	A confidence interval contains the likely values for the parameter. Values inside the confidence interval are more likely than those outside. This fallacy exists in several varieties, sometimes involving plausibility, credibility, or reasonableness of beliefs about the parameter
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	Likelihood decreases in a linear way as we move away from the sample mean
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	Linear reduction in likelihood across a CI.
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	Everything inside the CI is equally likely
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	Everything outside the CI is equally unlikely
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	All points inside a CI are equally likely to land on the μ
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	All points outside a CI are equally unlikely to land on the μ
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an educational tool that could be used to confront particularly misconceptions about CI distributions	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology, Medicine and other science disciplines	that a confidence interval has a uniform likelihood distribution across, where many of these show a cliff effect (a sudden major drop in likelihood at each limit of the interval)
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	Cliff effect
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	There is a likehood cliff at the end of a CI (both 50% and 95% CIs)
Kalinowski, Lai, Fidler and Cumming (2010)	Australia	Review statistical cognition studies and explain the contributions of both the quantitative and qualitative components	Both	Interactive survey and interview	Final year honors undergraduates and graduate students	a uniform distribution, implying that each point in the CI is equally likely to land on μ and each point out of the interval is equally unlikely to land on μ, with a sharp cliff separating the two

Crooks, Bartel and Alibali (2019)	USA	To assess conceptual knowledge of CIs in undergraduate as well as graduate psychology students	Empirical	Lab setting	Psychology graduate students and undergraduate students	A confidence interval gives you the range of possible values for the sample mean.
Fidler (2006)	Australia	To discuss reform efforts in statistical reporting regarding psycology, medicine and ecology, and through data on students understanding of CIs, explain how to improve statistics education in psycology	Theoretical	Theoretical	Students in psycology (and medicine, ecology)	Plausible values for sample mean
Hazra (2017)	India	To instruct how to use CIs confidently	Theoretical	Theoretical	NA	A CI is a range of plausible values for the sample
Fidler and Cumming (2005)	Australia	To help ease some misinterpretations of results to provide more accurate conclusions, demonstrate particular types of misconceptions and present figures and simulations that should lead to fewer misconceptions.	Both	Low powered study	Final year undergraduates and postgraduate environmental science students	CI provided "plausible values for the sample mean"
Kalinowski, Lai and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	A CI is a range of plausible values for the sample mean
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an educational tool that could be used to confront particularly misconceptions about CI distributions	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology, Medicine and other science disciplines	CIs are a range of plausible values for the sample mean

We have three theoretical, five empirical and two articles that are both. Regarding the theoretical articles' target populations, one article has no specific population and is called NA, the second has students and the last one has folks. The rest of the articles have students as target population. The level of education and subjects differ between the articles' students. The three theoretical articles do not have any research design, so it only says theoretical in the table. The empirical articles have four surveys and one lab setting. One of the articles with both has made a low powered study, and the other has interactive survey aswell as interview. Worth pointing out from the table is that most of the misinterpretations are from Kalinowski, Lai, and Cumming (2018).

These misinterpretations presented in table 1.8 all have to do with likelihood i.e., which specific values we get and whether we get them or not. The first misinterpretation comes from Morey et al. (2016) and is the last one from the article that inspired a category. This fallacy is called *The Likelihood fallacy* (TLF) and is expressed as followed:

- *"A confidence interval contains the likely values for the parameter. Values inside the confidence interval are more likely than those outside. This fallacy exists in several varieties, sometimes involving plausibility, credibility, or reasonableness of beliefs about the parameter."*

In the confidence procedure we have a fixed average probability that the true value is included. Although, this does not mean that reasonable values are included. The researchers argue that there is not any support, from the confidence intervals, of any reasonable belief about the parameter. As we can see from the table, the last five articles misinterpret confidence intervals like TLF. They all use the formulation "*plausible values"*.

In Pawel Kalinowskis' article *identifying misconceptions about confidence intervals* (2010) he found that students hold several misconceptions about confidence intervals. Many of the students believed the misconceptions that a confidence interval has a uniform likelihood distribution across, where many of these show a cliff effect (a sudden major drop in likelihood at each limit of the interval). We can from the table see that other authors than Kalinowski (2010) have identified this problem. We both have the part about confidence intervals having a uniform distribution, and cliff effect. Kalinowski, Lai, and Cumming (2018) have misinterpretations about points being/ not being equally likely to be in the confidence interval. They also mentioned misinterpretations about linear reduction in likelihood concerning the confidence interval.

4.4 Overlaps fallacies

The fourth table is called *overlaps fallacies* and include articles where the misinterpretation of confidence intervals have been made because of overlaps. The table consist of 8 articles and 10 misinterpretations.

Table 1.9. Overlaps fallacies

Authors	Origin	Purpose	Theoretical and/or empirical	Research design	Target population	Misinterpretation of confidence interval
Greenland, Senn, Rothman, Carlin, Poole, Goodman, and Altman (2016)	USA, Luxembourg, Australia, and UK	To form a discussion of the basic statistics and through this provide definitions. Also, present misinterpretations of p-values, CIs and power	Theoretical	Theoretical	NA	If two confidence intervals overlap, the difference between two estimates or studies is not significant
Kalinowski, Lai, and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	Overlap misconception
Tan and Tan (2010)	Singapore	Discussing the correct interpretation of CIs, and highlighting some common misunderstandings	Theoretical	Theoretical	NA	If the upper limit of the 95% CI of one group just touches the lower limit of the 95% CI of the other group, the p-value for the difference between the 2 groups is 0.05
Altman (2005)	UK	To discuss interpretations of CIs and some common misuses, focused on the principles rather than the mathematics of CIs	Theoretical	Theoretical	NA	common misuse of CIs in a comparative study is the presentation and comparison of separate CIs for each group rather than consideration of a CI for the contrast. This practice leads to inferences based on whether the two CIs, such as for the means in each group, overlap; or whether one group has a CI including the value for no effect whereas the other does not
Cumming and Fidler (2005)	Australia	To explore three CI problems: an incorrect belief about CI overlap and its relation to statistical significance; failure to distinguish between CIs and standard error bars; neglect of the importance of research design in applying and interpreting intervals	Theoretical	Theoretical	NA	'just touching' 95% CIs (i.e., CIs that just do not overlap) are equivalent to a statistically significant difference (at $p<.05$) between point estimates
Fidler (2006)	Australia	To discuss reform efforts in statistical reporting regarding psycology, medicine and ecology, and through data on students understanding of CIs, explain how to improve statistics education in psycology	Theoretical	Theoretical	Students in psycology (and medicine, ecology)	statistical nonsignificance is equivalent to evidence of 'no effect.'
Altman (2005)	UK	To discuss interpretations of CIs and some common misuses, focused on the principles rather than the mathematics of CIs	Theoretical	Theoretical	NA	A nonsignificant result (i.e., with $p > 0.05$) that the groups are ''the same
McCormack, Vandermeer and Allan (2013)	Canada	To review how researchers can look at very similar data but have different conclusions based purely on an over-reliance of statistical significance and a false understanding of CIs	Theoretical	Theoretical	NA	Overlap misconceptions
Belia, Fidler, Williams and Cumming (2005)	Australia	To investigate researchers' understanding of confidence intervals (CIs) and standard error (SE) bars	Empirical	Questionnaire and adjustable applet	Authors of empirical articles about confidence intervals (and SE bars), published in journals in psychology, behavioral neuroscience or medicine	error bars, whether a 95% CI or SE bars, just touch when means are just statistically significantly different (p .05)
Belia, Fidler, Williams and Cumming (2005)	Australia	To investigate researchers' understanding of confidence intervals (CIs) and standard error (SE) bars	Empirical	Questionnaire and adjustable applet	Authors of empirical articles about confidence intervals (and SE bars), published in journals in psychology, behavioral neuroscience or medicine	overlap of CIs or SE bars

These misinterpretations have to do with overlaps of confidence intervals i.e., whether for example intervals "touch each other" and if we have significance results or not. Some of the citations don't specifically mention *overlaps* but all the misinterpretations are of that concern. We have six theoretical and two empirical articles, no articles with both. Five of the six theoretical articles have no specific target population and are called NA and the one that has a target population has students. One of the empirical articles has students as target population and the other one has authors. None of the theoretical articles have a research design, and concerning the empirical articles, one has questionnaire and adjustable applet, and the other article has survey as research design.

There are some misinterpretations that are identical presented in the Table 1.9 with the same reason as before. The misinterpretations are these:

- *"Overlap misconception"* - Kalinowski, Lai, and Cumming (2018)
- *"Overlap misconceptions"* - McCormack, Vandermeer and Allan (2013)

In Table 1.9 we present citations where confidence intervals are interpreted by using incorrect conclusions about overlaps. From the paper *Statistical tests, P-values, confidence intervals, and power: a guide to misinterpretations* (2016) they present what they argue is a common misinterpretation, which is the first one presented in Table 1.9 which is:

- *"If two confidence intervals overlap, the difference between two estimates or studies is not significant."*

If you were to take two 95 % confidence intervals from different subgroups or studies, they may overlap substantially but the test for difference between them could still produce $P<0.05$. If they fail to overlap, we would be using the same assumptions we used to estimate the confidence intervals and find $P<0.05$ for the difference, and one of the 95 % intervals may contain the point estimate from the other. We can also see in table 1.9 that other articles have identified this misinterpretation as well, for example Altman (2005) or Tan and Tan (2010).

4.5 Miscellaneous fallacies

The last table presented contains articles with misinterpretations of confidence intervals of other sorts than those in the previous four fallacies. This table is called *miscellaneous fallacies*. As mentioned earlier the table consist of 18 articles and 33 misinterpretations.

Table 1.10. Miscellaneous fallacies.

Authors	Origin	Purpose	Theoretical and/or empirical	Research design	Target population	Misinterpretation of confidence interval
Greenland, Senn, Rothman, Carlin, Poole, Goodman, and Altman (2016)	USA, Luxembourg, Australia, and UK	To form a discussion of the basic statistics and through this provide definitions. Also, present misinterpretations of p-values, CIs and power	Theoretical	Theoretical	NA	An effect size outside the 95 % confidence interval has been refuted (or excluded) by the data
O'Brien and Yi (2016)	Canada	To bring up and describe some basic principles of CIs and their interpretation	Theoretical	Theoretical	NA	The CI is invalid if the sample selection is not completely random
Greenland, Senn, Rothman, Carlin, Poole, Goodman, and Altman (2016)	USA, Luxembourg, Australia, and UK	To form a discussion of the basic statistics and through this provide definitions. Also, present misinterpretations of p-values, CIs and power	Theoretical	Theoretical	NA	If one 95 % confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of CIs	Empirical	Questionnaire	Researchers and students—all in the field of psychology	The "null hypothesis" that the true mean equals 0 is likely to be incorrect
Lyu, Peng and Hu (2018)	China and Germany	To introduce raw data that is available for anyone interested in examining how students aswell as researchers misinterpret inter alia CIs, and how NHST and CIs affect the interpretation of study- or research results	Empirical	Questionnaire	Psychological researchers in different positions in related fields of psychology	The "null hypothesis" that the true mean equals 0 is likely to be incorrect
Callaert (2007)	Belgium	Address statistical reasoning at a first course in statistics level, in higher education	Theoretical	Theoretical	Students	The first difficulty relates to the randomness of intervals in repeated samples
Fidler and Cumming (2005)	Australia	To help ease some misinterpretations of results to provide more accurate conclusions, demonstrate particular types of misconceptions and present figures and simulations that should lead to fewer misconceptions.	Both	Low powered study	Final year undergraduates and postgraduate environmental science students	replication means don't differ much from the original mean
Miller and Ulrich (2016)	New Zealand and Germany	To question Hoekstra et al's study where they got the result that first year & master students and researchers "have no reliable knowledge about the correct interpretation of CIs"	Theoretical	Theoretical	First year, master students & researchers	The "null hypothesis" that the true mean equals 0 is likely to be incorrect
García-Pérez and Alcalá-Quintana (2016)	Spain	To re-analyze Hoekstra et al's study of interpretation of CIs, and discuss the two interpretations of CIs and why misinterpretation can't be inferred from endorsement of some of the items	Both	(replication) questionnaire	First year and master students	The "null hypothesis" that the true mean equals 0 is likely to be incorrect
Belia, Fidler, Williams and Cumming (2005)	Australia	To investigate researchers' understanding of confidence intervals (CIs) and standard error (SE) bars	Empirical	Questionnaire and adjustable applet	Authors of empirical articles about confidence intervals (and SE bars), published in journals in psychology, behavioral neuroscience or medicine	misconceptions held by many researchers about the relation between error bars and replication
Kalinowski, Lai and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	A 50% CI indicates lack of data
Kalinowski, Lai and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	A 50% CI means the μ could land anywhere
Kalinowski, Lai and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience	A CI is a range of individual scores

					and medicine and other	
O'Brien and Yi (2016)	Canada	To bring up and describe some basic principles of CIs and their interpretation	Theoretical	Theoretical	NA	Misinterpret the confidence interval as a statement about the unknown parameter
Hoekstra, Morey, Rouder et. al. (2014)	Netherlands and USA	To learn more about researchers' interpretation of CIs	Empirical	Questionnaire	Researchers and students—all in the field of psychology	A CI can be used to evaluate only the procedure and not a specific interval
Lyu, Peng and Hu (2018)	China and Germany	To introduce raw data that is available for anyone interested in examining how students aswell as researchers misinterpret inter alia CIs, and how NHST and CIs affect the interpretation of study- or research results	Empirical	Questionnaire	Psychological researchers in different positions in related fields of psychology.	A CI can be used to evaluate only the procedure and not a specific interval
Fidler (2006)	Australia	To discuss reform efforts in statistical reporting regarding psycology, medicine and ecology, and through data on students understanding of CIs, explain how to improve statistics education in psycology	Theoretical	Theoretical	Students in psycology (and medicine, ecology)	CI is a range of individual scores within one standard deviation
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an educational tool that could be used to confront particularly misconceptions about CI distributions	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology, Medicine and other science disciplines	CIs are range of individual scores
Kalinowski (2010)	Australia	Present a taxonomy of CI misconceptions, explore faulty conceptual models which could be the source of some misconceptions, and propose an educational tool that could be used to confront particularly misconceptions about CI distributions	Empirical	Survey	Fourth year undergraduate and postgraduate students from Psychology, Ecology, Medicine and other science disciplines	CIs are a range of individual scores within one standard deviation
Crooks, Bartel and Alibali (2019)	USA	To assess conceptual knowledge of CIs in undergraduate as well as graduate psychology students	Empirical	Lab setting	Psychology graduate students and undergraduate students	A confidence interval gives you the range of the individual scores
Crooks, Bartel and Alibali (2019)	USA	To assess conceptual knowledge of CIs in undergraduate as well as graduate psychology students	Empirical	Lab setting	Psychology graduate students and undergraduate students	A confidence interval gives you the range of the individual scores within one standard deviation of the population mean
Finch and Cumming (2009)	Australia	To discuss the meaning and interpretation of CIs in single studies, to review and integrate	Theoretical	Theoretical	NA	A CI can be wrongly interpreted as describing the range of observed values in the data
Finch and Cumming (2009)	Australia	To discuss the meaning and interpretation of CIs in single studies, to review and integrate	Theoretical	Theoretical	NA	CI is described as a range of values for a sample summary statistic (such as the sample mean)
Naimi and Whitcomb (2020)	USA	To bring up and demonstrate key properties of frequentist CIs, to elucidate interpretations and explain common misunderstandings	Theoretical	Theoretical	NA	we can interpret confidence interval estimates by invoking properties that apply to the estimand
Eliason (2018)	USA	To analyze pre-service teachers' conception about CIs	Empirical	Pretest and interviews	Students in mathematics education course for pre-service secondary teachers focused on the teaching and learning of statistics and probability	A Confidence Interval Allows Us To Make Inferences About More Than the Mean of the Population
Naimi and Whitcomb (2020)	USA	To bring up and demonstrate key properties of frequentist CIs, to elucidate interpretations and explain common misunderstandings	Theoretical	Theoretical	NA	Calculations for Confidence Intervals are Not Based on Actual Sampling Distributions
Callaert (2007)	Belgium	Address statistical reasoning at a first course in statistics level, in higher education	Theoretical	Theoretical	Students	The second difficulty relates to the non-randomness of a confidence interval after the sample has been taken
Navruz and Delen (2014)	USA	To show that CIs are useful and valuable in research studies when used in the correct form and with correct interpretations	Theoretical	Theoretical	NA	CIs do not do more than the NHSST
Fidler and Cumming (2005)	Australia	To help ease some misinterpretations of results to provide more accurate conclusions, demonstrate particular types of	Both	Low powered study	Final year undergraduates and postgraduate environmental science students	CI is the "range", or "truncated range of individual scores

		misconceptions and present figures and simulations that should lead to fewer misconceptions.				
Kalinowski, Lai and Cumming (2018)	Australia	To explore students interpretation of the relative likelihood of capturing a population parameter at diverse points of a CI in two studies	Empirical	Survey	Final year honors undergraduates, graduate students and post-graduate students in psychology, social science, neuroscience and medicine and other	Standard shape regardless of C% level
Finch and Cumming (2009)	Australia	To discuss the meaning and interpretation of CIs in single studies, to review and integrate	Theoretical	Theoretical	NA	The parameter is inside the CI
Yang (2011)	Taiwan	To illustrate partcharacteristics underlying mathematics teachers' alternative understanding/misunderstanding of CI related concepts	Emperical	Questionnaire	Mathematics teachers	some teachers made statistical inference mainly based on distributions of sampled values
Callaert (2007)	Belgium	Address statistical reasoning at a first course in statistics level, in higher education	Theoretical	Theoretical	Students	misinterpretation of the conditionality and the probability of the intersection event

We have seven theoretical, nine empirical and two articles that are both theoretical and empirical. None of the theoretical articles have a research design. Concerning the empirical articles' research designs there are four with questionnaire, one with both questionnaire and adjustable applet, one with pretest and interviews, one lab setting, and two surveys. The two articles that are both empirical and theoretical have low powered study and (replication) questionnaire as research design. The target populations of the theoretical articles were students in two of them, one with students and researchers and the rest didn't have any specific population. In the empirical articles we had one with researchers and students, two with researchers, one with authors, one with teachers and four with students. The articles which were both empirical and theoretical had students as target population in both articles.

In the table we find all kinds of misinterpretations that didn't fit in any of the other tables. To structure the relatively many kinds of misinterpretations we have tried to divide them into different groups, presented in the table below. The groups were created depending on how many other misinterpretations seemed alike. We still have one group that isn't divided, but instead is called *unspecified* and consist of the misinterpretations that after all haven't been like other misinterpretations. The rest of the groups have been made because we found similarity between the misinterpretations, by for example their focus in the misinterpretations or on similar use of words or meaning. Each column will be presented in chronological order.

Table 1.11. Groups for the category Miscellaneous fallacies

Unspecified	Parameter	Replication	Interval vs. procedure	Hypothesis	Sample	Individual scores
An effect size outside the 95 % confidence interval has been refuted (or excluded) by the data	The parameter is inside the CI	misconceptions held by many researchers about the relation between error bars and replication	A CI can be used to evaluate only the procedure and not a specific interval	The "null hypothesis" that the true mean equals 0 is likely to be incorrect	The CI is invalid if the sample selection is not completely random	CIs are range of individual scores
CIs do not do more than the NHSST	Misinterpret the confidence interval as a statement about the unknown parameter	replication means don't differ much from the original mean	A CI can be used to evaluate only the procedure and not a specific interval	The "null hypothesis" that the true mean equals 0 is likely to be incorrect	Calculations for Confidence Intervals are Not Based on Actual Sampling Distributions	CI is a range of individual scores within one standard deviation
we can interpret confidence interval estimates by invoking properties that apply to the estimand	A 50% CI means the μ could land anywhere			The "null hypothesis" that the true mean equals 0 is likely to be incorrect	The second difficulty relates to the non-randomness of a confidence interval after the sample has been taken	A CI is a range of individual scores
A Confidence Interval Allows Us To Make Inferences About More Than the Mean of the Population				The "null hypothesis" that the true mean equals 0 is likely to be incorrect	The first difficulty relates to the randomness of intervals in repeated samples	CIs are a range of individual scores within one standard deviation
Standard shape regardless of C% level					some teachers made statistical inference mainly based on distributions of sampled values	A confidence interval gives you the range of the individual scores
A 50% CI indicates lack of data					CI is described as a range of values for a sample summary statistic (such as the sample mean)	A confidence interval gives you the range of the individual scores within one standard deviation of the population mean
If one 95 % confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one						CI is the "range", or "truncated range of individual scores
A CI can be wrongly interpreted as describing the range of observed values in the data						
misinterpretation of the conditionality and the probability of the intersection event						

In the first column, *unspecified*, we have nine misinterpretations. The first misinterpretation comes from the article *Statistical tests, P-values, confidence intervals, and power: a guide to misinterpretations* (2016) and was presented there as a common misinterpretation. The misinterpretation is the following:

- *"An effect size outside the 95 % confidence interval has been refuted (or excluded) by the data."*

The combination of the data with the assumptions, as well as the arbitrary 95 % criterion, is required to declare an effect size past the interval, which is more or less incompatible with the observations. The confidence interval is, just like the P value, calculated from various assumptions. Violating this may lead to the presented result. Also another misinterpretation in this column comes from the same article, that argue this one also is a common one, and is the following:

- *"If one 95 % confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one"*

The precision of a statistical estimation, when the model is correct, is measured directly by the width of the confidence interval. There is no inclusion or exclusion of the null or any other value.

The second column is called *parameter*. As we can see the misinterpretations have to do with the parameter in some way. This column has some connection to what is said in the theory. Our population parameter is treated as a fixed (unknown) value, which means that the fixed value is either inside the interval or outside the interval. We can not know if the parameter is inside the confidence interval or not, we just know it is either inside or outside. In the theory it says that a confidence interval should not be used as a statement about the parameter, as it relates to the specific sample at hand. Instead, it provides a statement about the conception of the procedure of drawing such intervals repeatedly.

In the third column, *replication*, we have two misinterpretations. These misinterpretations seem to arise from misunderstanding with connection about replication.

The fourth column is called *Interval vs. procedure*. The misinterpretations are identical but are from two different articles (same reason as earlier). As mentioned in the theory, a computed confidence interval should not be interpreted, according to the confidence interval theory. And the true interpretation of confidence limits, according to Neyman and others is "*The result of a procedure that will contain the true value in a fixed proportion of samples*".

The fifth column is called *Hypothesis*. As we can see the misinterpretations are identical (the reason is the same as earlier).

The sixth column is called *Sample* and consist of misinterpretations that have some connection to samples in some meaning.

Our last, seventh column is called *Individual scores* and consist of seven different misinterpretations with some kind of relationship to individual scores. Some of the misinterpretations are identical, with the same reason as before.

4.6 Summary of the individual tables

Two types of misconceptions Fidler (2005) present is relational and definitional. Definitional misconceptions refer to what a confidence interval measures, estimates or its inferential nature. Relational misconceptions, on the other hand, refer to expectations of relationships between confidence level, width and sample size. The results from this study could by divided into the two types of misconceptions that were presented by Fidler (2005), which were *definitional misconceptions* and *relational misconceptions*. The first kind refers to what a confidence interval measures, estimates or its inferential nature, while the other one refers to expectations of relationships between confidence level, width and sample size. In all categories we could identify both of these kinds.

We first had four categories where articles that had misconceptions alike each others were placed. Then we had one last category *miscellaneous fallacies* which consisted of the less usual or mentioned misinterpretations, at least for this study, we didn't find misinterpretations similar enough to conduct a new specified category. We would say the common mistakes are the ones mentioned in the first four categories and the most common one we would argue is the category *probability fallacies* with the reason that it is the most comprehensive category if we ignore the last category.

4.7 Some words about the article by Hoekstra et.al. (2014)

Some of the misinterpretations are relatively recurrent and comes from the article by Hoekstra et.al. (2014) called *Robust misinterpretation of confidence intervals*. We find this specific article worthy of a few words since it is *recurrently used in this study* and it *has received some critics from other authors*.

In a survey conducted in 2014 by Morey, Hoekstra, Rouder and Wagenmakers (HMRW) six statements of misunderstanding of confidence intervals were given to students as well as researchers. Four of the statements assign probabilities to parameters or hypotheses, which is something that in the frequentist framework is not allowed. The other two statements mention the boundaries of the CI, whereas a CI can be used to evaluate *only* the procedure and not a specific interval. The results could be taken as evidence that large misunderstandings of confidence intervals exist even within the target population. In 2015, this survey was targeted with critique from Miller and Ulrich (MU), whom argued that some of the statements used in the survey could be viewed as *"*appropriate under other meanings of 'probability' that are in common use*"*, in other words they argue the conclusion that HMRW made is incorrect.

As a response to the critique, an article called *Continued misinterpretation of confidence intervals: response to Miller and Ulrich* was made in 2015, which continually argued that not only students and researchers had many misunderstandings of confidence intervals, but also methodologists such as Miller and Ulrich. While MU argues that the interpretation of "probability" is the problem, HMRW argues that the reference class problem prevents any kind of unique association of long-run frequency in individual observed intervals (Morey et al., 2015). The reference class problem will emerge by any attempts to associate characteristics in individual events regarding long-run frequencies (Reichenbach, 1949; Venn, 1888; von Mises, 1957).

5. Discussion

The aim in the paper which was to perform an explorative literature review that consist of articles which studies how confidence intervals are misinterpreted has been fulfilled. Based on the results in this paper we can state that some misinterpretations are more common than others. The common misinterpretations of confidence intervals are category one to four, where the first category *probability fallacies*, is the most common one. The less common misinterpretations, based on this paper, are the ones presented in the fifth category.

There are some misinterpretations from different articles that are mentioned in the table more than once. This is because it's an area of research where researchers cite each other. In this paper we have several times seen that the same misinterpretations have appeared in different articles. Perhaps however, this is more common in this paper due to the methodology used when selecting articles.

Some of the misinterpretations that we have encountered in the articles have been difficult to interpret. The first category, the one concerning probability fallacies, was relatively straightforward and easy to grasp, and was therefore not so difficult to interpret. Categories two to four were more difficult to interpret compared to the first one since they were a bit more complex and had more dimensions. In the first category most of the misinterpretations were more or less identical in their formulations and they referred more clearly to the same problem, which means it was easier to identify what the misinterpretations were about. The last category, *Miscellaneous fallacies*, was the most difficult to interpret. The reason is mostly because the misinterpretations were more diffuse as there was an uncertainty regarding why a misinterpretation was a wrong interpretation.

The paper consisted of subjective actions concerning the reasons mentioned in *3.3 Comments concerning the method*. These may have affected the results. It is difficult to know whether we have missed relevant misinterpretations and how widely this paper covers the area of misinterpretations. There were also some comments about the restrictions and possible ways to conduct a review. But even though we may not have covered all relevant misinterpretations in this paper, it is clear that category one especially, but also two, three and four are common misinterpretations.

References

Altman, D. G. (2005). Why we need confidence intervals. *World Journal of Surgery*, *29*(5), 554-556. https://doi.org/10.1007/s00268-005-7911-0

Belia, S., Fidler, F., Williams, J., & Cumming, G. (2005). Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, *10*(4), 389-396. https://doi.org/10.1037/1082-989X.10.4.389

Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research*, *468*(3), 885-892. https://doi.org/10.1007/s11999-009-1164-4

Briggs, W. M. (2012). It is time to stop teaching frequentism to non-statisticians. *arXiv preprint arXiv:1201.2590*. Retrieved from https://arxiv.org/abs/1201.2590

Callaert, H. (2007). Understanding confidence intervals. *Proceedings of the 5th Congress of the European Society for Research in Mathematics Education*, 692–701.

Canal, G. Y., & Gutiérrez, R. B. (2010). The confidence intervals: a difficult matter, even for experts. *Proceedings of the Eighth International Conference on Teaching Statistics*, 1-4.

Canal, G.Y. & Ruiz, L.R. (2015). On the meaning that teachers in training will give to the accuracy of a confidence interval and its relationship with the sample size and level of confidence. *Proceedings of the 2015 Satellite Conference of the International Association for Statistical Education (IASE)*, 1-6.

Crooks, N. M., Bartel, A. N., & Alibali, M.W. (2019). Conceptual Knowledge of Confidence Intervals in Psychology Undergraduate and Graduate Students. *Statistics Education Research Journal*, *18*(1), 46-62.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and non-central distributions. *Educational and Psychological Measurement*, *61*(4), 532–574. https://doi.org/10.1177/0013164401614002

Cumming, G., & Fidler, F. (2005). Interval estimates for statistical communication: Problems and possible solutions. *IASE / ISI Satellite*, 1-7.

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall?. *Psychological Methods*, *11*(3), 217-227.

Eliason, K. L. (2018). *Addressing pre-service teachers' misconceptions about confidence intervals* [Thesis, Brigham Young University]. Brigham Young University ScholarArchive. https://scholarsarchive.byu.edu/etd/6917.

Fidler, F. (2005). *From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology* [Unpublished PhD thesis, University of Melbourne].

Fidler, F. (2006). Should psychology abandon p-values and teach CIs instead? Evidence-based reforms in statistics education. *Proceedings of the seventh international conference on teaching* statistics, 1-6.

Fidler, F, & Cumming, G. (2005). Teaching confidence intervals: Problems and potential solutions. *Proceedings of the 55th international statistics institute session*, 1-5.

Fidler, F., & Loftus, G. R. (2009). Why figures with error bars should replace p-values: Some conceptual arguments and empirical demonstrations. *Journal of Psychology*, *217*(1), 27–37. https://doi.org/10.1027/0044-3409.217.1.27

Finch, S., & Cumming, G. (2009). Putting research in context: Understanding confidence intervals from one or more studies. *Journal of Pediatric Psychology*, *34*(9), 903-916. https://doi.org/10.1093/jpepsy/jsn118

Foster, C. (2014). Confidence Trick: The Interpretation of Confidence Intervals. *Canadian Journal of Science, Mathematics and Technology Education*, *14*(1), 23-34. https://doi.org/10.1080/14926156.2014.874615

Harlow, L. L. (1997). Significance Testing in Introduction and Overview. In L. L. Harlow, S., A. Muliak., & J. H. Steiger (Eds.), *What If There Were No Significance Tests*? (1-17). Mahwah, NJ, USA: Lawrence Erlbaum.

Hazra, A. (2017). Using the confidence interval confidently. *Journal of Thoracic Disease*, *9*(10), 4125-4130. https://doi.org/10.21037/jtd.2017.09.14

Hespanhol, L., Vallio, C. S., Costa, L. M., & Saragiotto, B. T. (2019). Understanding and interpreting confidence and credible intervals around effect estimates. *Brazilian Journal of Physical Therapy*, *23*(4), 290-301. https://doi.org/10.1016/j.bjpt.2018.12.006

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E. J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hogg, R. V., & Tanis, E. A. (2005). *Probability and statistical inference*. Upper Saddle River, NJ: Prentice Hall.

García-Pérez, M. A., & Alcalá-Quintana, R. (2016). The Interpretation of Scholars' Interpretations of Confidence Intervals: Criticism, Replication, and Extension of Hoekstra et al. (2014). *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.01042

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P-values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gutiérrez, B. R., & Yáñez, C. G. (2009). Experts and students' conceptions regarding confidence intervals. *Heurística*, *16*, 5–12.

Kalinowski, P. (2010). Identifying misconceptions about confidence intervals. *Proceedings of the eighth international conference on teaching statistics*, 1-5.

Kalinowski, P., Lai, J., & Cumming, G. (2018). A cross-sectional analysis of students' intuitions when interpreting CIs. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00112

Kalinowski, P., Lai, J., Fidler, F., & Cumming, G. (2010). Qualitative Research: An Essential Part of Statistical Cognition Research. *Statistics Education Research Journal*, *9*(2), 22-34.

Karlen, D. (2002). Credibility of confidence intervals. *Advanced Statistical Techniques in Particle Physics, Proceedings*, 53-57.

Karolinska institutet. (2021, december 8). *Systematisk litteraturöversikt som examensarbete*. https://kib.ki.se/soka-vardera/systematiska-oversikter/systematisk-litteraturoversikt-som-examensarbete.

Karolinska institutet. (2021, november 15). *Systematiska översikter*. https://kib.ki.se/soka-vardera/systematiska-oversikter.

Lang, J. M., Rothman, K. J., & Cann, C. I. (1998). That confounded. *P*-value. *Epidemiology*, *9*, 7–8.

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, *5*(6), 161–171. https://doi.org/10.1111/1467-8721.ep11512376

Lyu, Z., Peng, K., & Hu, C-P. (2018). P-Value, Confidence Intervals, and Statistical Inference: A New Dataset of Misinterpretation. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00868

Lyu, X., Xu, Y., Zhao, X., Zuo, X., & Hu, C. (2020). Beyond psychology: Prevalence of p value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology*, *14*. http://doi.org/10.1017/prp.2019.28

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee M. D., & Wagenmakers E-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1)**,** 103–123. https://doi.org/10.3758/s13423-015-0947-8

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, *57*(3), 203–220. https://doi.org/10.1037/h0087426

McCormack, J., Vandermeer, B., & Allan, G. M. (2013). How confidence intervals become confusion intervals. *BMC Medical Research Methodology*, *13*(1), 1-6. https://doi.org/10.1186/1471-2288-13-134

Miller, J., & Ulrich, R. (2016). Interpreting confidence intervals: A comment on Hoekstra, Morey, Rouder, and Wagenmakers (2014). *Psychonomic Bulletin & Review*, *23*(1), 124-130. https://doi.org/10.3758/s13423-015-0859-7

Morey, R. D., Hoekstra, R., Rouder, J. N., & Wagenmakers, E-J. (2016). Continued misinterpretation of confidence intervals: response to Miller and Ulrich. *Psychonomic Bulletin & Review, 23*(1), 131–140. https://doi.org/10.3758/s13423-015-0955-8

Naimi, A. I., & Whitcomb, B. W. (2020). Can confidence intervals be interpreted?. *American Journal of Epidemiology*, *189*(7), 631-633. https://doi.org/10.1093/aje/kwaa004

Navruz, B., & Delen, E. (2014). Understanding confidence intervals with visual representations. *Abant İzzet Baysal Üniversitesi Eğitim Fakültesi Dergisi*, *14*(1), 346-360.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A. Mathematical and Physical Sciences, 236*(767), 333-380. https://doi.org/10.1098/rsta.1937.0005

O'Brien, S. F., & Yi, Q. L. (2016). How do I interpret a confidence interval?. *Transfusion*, *56*(7), 1680-1683. https://doi.org/10.1111/trf.13635

O'Neill, P. D. (2002). A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Mathematical Biosciences, 180*, 103-114. https://doi.org/10.1016/S0025-5564(02)00109-8

Reaburn, R. (2014). Students' understanding of confidence intervals. *Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS-9)*, 1-4.

Rothman, K. J. (1978). A show of confidence. *New England Journal of Medicine*, *299*(24), 1362–1363.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129. https://doi.org/10.1037/1082-989X.1.2.115

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik., & J. H. Steiger (Eds.), *What if there were no significance tests?* (37-64). Mahwah, NJ, USA: Lawrence Erlbaum.

Tan, S. H., & Tan, S. B. (2010). The correct interpretation of confidence intervals. *Proceedings of Singapore Healthcare*, *19*(3), 276-278.

Yang, K. L. (2011). Mathematics teachers make statistical inference based on the distribution of sampled values. *Proceedings of the British Society for Research into Learning Mathematics*, *31*(3).

Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, *75*(3), 315-321.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594