P-values and confidence intervals? Think again

Abstract

Typical coursework in quantitative science includes inferential frequentist statistics, and many graduates master the technical side: they know how to compute *p*-values and confidence intervals. Unfortunately, much less emphasis is placed on correct interpretation of these numbers and on attempts to understand their relevance (or irrelevance) to the business of causal inquiry. In this text (originally written as a book chapter), I tell the technical story of *p*-values and confidence intervals and review common misinterpretations. Despite their widespread use, both *p*-values and confidence intervals do not add much to the study of causation.

This text is a hybrid of teaching material and a commentary on many previous publications—too many to mention…

Infinite replications

The story begins with a theoretical idea called *exact* replication, which is never possible. Suppose that one particular study, call it study #1, has yielded an estimate ($Estimate_1$) of some causal parameter—for instance, the rate ratio for death in some trial. Now, assume it were possible to replicate the study *exactly* many times and compute an estimate from each replication. If so, we would have obtained a sequence of many estimates as shown below:

#1 Study→Data→ $Estimate_1$
#2 (replication) Study→Data→ $Estimate_2$
#3 (replication) Study→Data→ $Estimate_3$
#4 (replication) Study→Data→ $Estimate_4$
.
.
.
#n (replication) Study→Data→ $Estimate_n$

Continuing this mental exercise an infinite number of times, we may define a theoretical quantity—the average of all these estimates—and call it the expected value. In simplified symbolic language:

Expected value =
$\lim (Estimate_1 + Estimate_2 + \ldots + Estimate_n)/n$
when $n \rightarrow \infty$

Although the expected value is an unknown abstract quantity, it will help us to define two central ideas—bias and randomness—which will be explored in the following sections. But the essence is this. Bias alludes to any discrepancy between the expected value and the parameter of interest, whereas randomness alludes to any discrepancy between the study's estimate and the expected value (Figure 1). (In formal language, what I called "estimate" should be called a point estimate, to distinguish it from an interval estimate.)

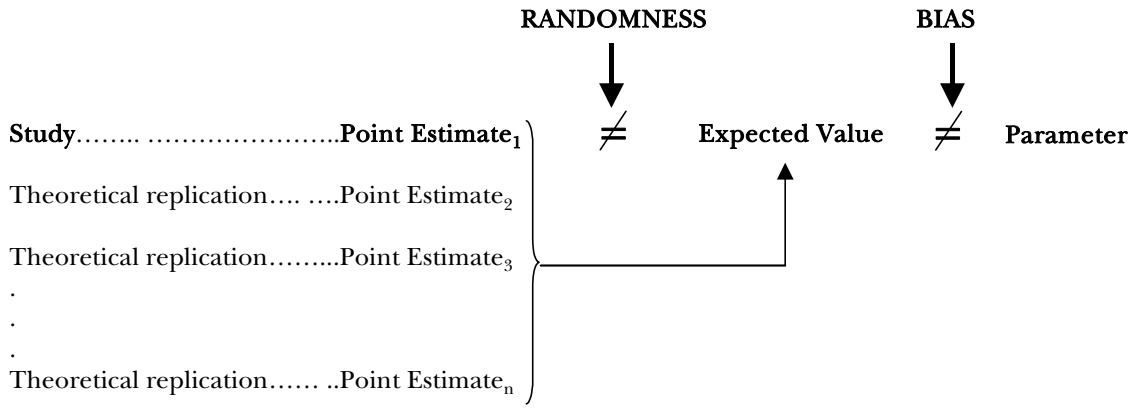


Figure 1. A graphical illustration of bias and randomness.

Biased estimate or biased estimator?

Consider a hypothetical trial of a drug called streptokinase for patients with ischemic stroke and assume that the rate ratio for death for the causal contrast between streptokinase and a sugar pill was, unfortunately, 1.47. Is that number a biased estimate of the causal parameter? Is 1.47 an invalid number?

The answer to these questions might cause a little surprise. Even if some kind of bias undermines the validity of the rate ratio from that trial, it is not the number that is invalid but the questions I have just asked. It is incorrect to ask whether a number is biased because the term applies to the *process* that generated the number, not to the number that was generated by the process—a subtle but an important distinction. At one time or another almost every mind confuses the two ideas because we all crave the answer to the question about the estimate itself: "Is this number the right number?" Unfortunately, the scientific method offers no direct answer to this question and the questions it does offer to answer might seem irrelevant on the surface.

We say that the process that generated a point estimate is *biased* when the expected value of that process differs from the parameter of interest. Bias implies that if we were able to replicate the study infinite times and compute the average of infinite point estimates, this average will not hit on the Truth.[1] Conversely, the process that generated a point estimate is called *unbiased* if the expected value will be equal to the parameter we are trying to estimate. In the interest of brevity, statisticians have invented a one-word substitute for "the process that generated the point estimate"—*estimator*—but we should not confuse an estimator with an estimate. A rate ratio of 1.47 from the streptokinase trial is an estimate, whereas the process that generated this number is an estimator. Since the whole exercise is mental, we can also envision an estimator that is biased a lot and one that is biased a little, depending on the distance between the expected value and the value of the parameter. Notice again that I am alluding here to the difference between *two unknown numbers*, not between the result of the trial (1.47) and the unknown causal parameter.

With these definitions in mind, it is easy to understand, for example, why confounders are a source of bias. If we perpetually replicate the process that generated a confounded measure of effect, the average of all estimates will still be confounded; it will not be equal to the causal parameter.

Notice the difference between bias in science and bias in informal language. When the layperson says that a number is biased, he or she indeed delivers two messages, one of which says that the outcome is wrong and the other explains why it is wrong ("because the process was biased"). Unlike the layperson, however, a scientist knows that every estimate is likely to be wrong regardless of whether the estimator was biased, so the layperson's dual assertion is vague. If she implies that an unbiased process would have yielded the right answer (the parameter), then she is wrong: no estimate is likely to hit on the Truth, and an estimate from an unbiased process could be far from the Truth. If the layperson does not imply that an unbiased process would have yielded the right answer, why blame a biased process for yielding a wrong answer? If all estimates are wrong, does it matter *why* they are wrong?

It matters. To the scientist.

We are witnessing again the disparity between the scientific method and its outcome; between our power to choose what to do and our lack of power to guarantee a correct answer. Although nothing can guarantee that an estimate will be close to the parameter of interest, we are still committed to coherent reasoning of our actions. We still have to rationalize the method by which we derive the point estimate—the estimator, that is. If Truth seeking is the goal, the preference for an unbiased estimator seems natural.

How do we know that an estimator is unbiased? Well, we never know. We try honestly and meticulously to prevent and remove bias and then conjecture that we have obtained an estimate from an unbiased estimator. In causal inquiry, for example, we display our assumptions in a causal diagram and try to remove bias by design and analysis. That's all that science has to offer to causal inquiry and that's a lot more than is offered by methods such as coin tossing, guessing, and witchcraft. Not because the right answer is waiting at the end of the road but because the road itself is constructed of superior reasoning.

It is time, perhaps, to bring to light common criticism of the scientific method and then bury it for good. It is called "the moment reflection argument", popping up whenever we rationalize a method by saying "assuming that". Three examples are listed below but there are many more:

- A moment reflection would suggest it is absurd to assume that any causal diagram corresponds to the Truth.

- A moment reflection would suggest it is absurd to assume that an estimator from an observational study is unbiased.

- A moment reflection would suggest that all statistical models are wrong.

I am willing to accept the moment reflection argument on epistemological level, if the intention is to remind us that scientific knowledge remains conjectural forever. On methodological level, however, the argument is no more than arrogant truism. And the best rebuttal may be a question: How come we have accumulated so much *conjectural knowledge* about causal reality despite this discouraging truism?

[1] Formally, the expected value is defined on an infinite-size study, rather than on an infinite number of replications. But the difference is subtle.

"Random error"

I dislike the word "random" because it rings like chaotic behavior, as in "a random act of violence". In scientific inquiry, however, randomness often alluded to reality that evolved from probabilistic rules rather than from chaos. A so-called random act of violence, for example, is realization of a propensity to act violently, and a random sample of U.S. residents is realization of a procedure called random sampling. That we cannot know what reality a process will yield does not imply underlying chaos.

Much of inferential statistics is built on the idea of probabilistic realization. Once you assume that probabilistic rules have governed the estimate from a study, you can use the data to calculate things such as standard error, *p*-value, and confidence interval, which will be discussed later. The interesting question, however, is not how the math works but what source of randomness explains why a point estimate from an unbiased estimator rarely, if ever, hits on the parameter. And at a deeper level the key questions are these:

- What drives an estimate away from the expected value?
- Why should endless theoretical replication produce a distribution of point estimates rather than repeatedly produce the same estimate?
- What replication do we have in mind?

Sometimes, we can answer these questions mechanistically, pointing to a probabilistic process that we have initiated—a process that generated one point estimate and could have generated others. The classical example is random sampling. If we estimate the proportion of smokers in some town from one random sample of 100 residents, we may equate replication with theoretical re-sampling of 100 residents that will likely yield a different sample and a different point estimate. Endless replication will therefore generate a sampling distribution—a probability distribution of many point estimates that are scattered around the true proportion of smokers in that town. In that distribution, some ranges of estimates will show up more often than others. For example, it is more probable to get estimates in the immediate range of the true proportion than estimates at the tails of the distribution.

Many courses in elementary statistics present similar examples as the foundation of "inferential frequentist statistics", the school of statistics that speaks the language of *p*-values and confidence intervals. Now, if you don't see the relevance to causal inquiry, let me reassure you that you haven't missed any subtle point. The story as told is irrelevant indeed.

How, then, do we answer the three questions in the context of causal inquiry? Why should endless theoretical replication generate a distribution of point estimates around a causal parameter rather than repeatedly hit on the target? And what replication do we have in mind, anyway?

There are several possible answers, most of which are linked to our choice between two models of causation: determinism and indeterminism. I will not discuss them here, however.

Standard error

The variance (VAR) or the standard deviation (SD) describe the spread of the values of a variable around its arithmetic mean. To compute the variance, we subtract the mean from each observed value, square the result (the difference), and take the average of all those squared differences. The variance is therefore the average squared difference from the mean whereas the standard deviation is simply the square root of the variance: $SD = VAR^{½}$. (If the word *standard* doesn't sound intuitive, think about *standardized* instead: By computing the deviation from the mean we standardize each value to the mean. Now, instead of thinking about the observed values of a variable, think by analogy about many estimates from an estimator, which we could have obtained from many theoretical replications. And instead of thinking about the mean of a variable, think about the mean of these estimates, which is the expected value. Next, think about the spread of these estimates around their expected value (the so-called sampling distribution) and about the variance and the standard deviation as measures of that spread. The analogy I have just described is shown next.

Variable	≡	Estimator
Mean value	≡	Expected value
Observed values	≡	Point estimates
Distribution	≡	Sampling distribution
Variance	≡	Variance
Standard deviation	≡	Standard error (standard deviation of the estimates)

It is not too difficult to grasp the ideas of "distribution" and "spread" of a variable because we can display the values in a graph and see the spread. But when we come to the sampling distribution of point estimates around the expected value, things get a little abstract. The distribution is formed in our mind by some kind of theoretical replication; the expected value remains unknown; and of all possible point estimates only one is known: the estimate from the study. So, whenever I lose the thread of reasoning, I look back at the analogy above.

Two other sources of confusion make matters worse: the word "sampling" in "sampling distribution" and the term standard error for a standard deviation. As I mentioned before, sampling distribution is a statistical term for any probability distribution that can be formed by replication. (Perhaps "replication distribution" would have been a better jargon.) But why call the standard deviation of a replication distribution "standard error"? What's the error? Well, if the estimator is unbiased, its expected value is also the value of the parameter and, therefore, any estimate that does not hit on the expected value is an error; it misses the truth.

To sum up, the standard error measures some aspects of randomness. Probabilistic rules cause the points estimates—one known and others theoretical—to form a sampling distribution around the expected value, and the standard error is the standard deviation of that distribution. When the sampling distribution is wide, the standard error is large; when it's tight, the standard error is small. We prefer the latter, of course: we want a hit on the expected value.

Sample size

Our preference for a small standard error rests, again, on reasoning for choosing a method, not on knowing what the outcome might be. Facing a choice between a biased estimator and an unbiased one, we readily opt for the latter (even though we are not guaranteed a better return.) Facing a choice between an estimator that produces a wide sampling distribution and an estimator that produces a tight one, we go for the latter again. But nothing, of course, guarantees that an estimate from a tight distribution will be close to the parameter; we might obtain a tail value—by chance.

In the battle for truth we have two enemies to fight: bias and randomness. Fortunately, we may use a powerful weapon against randomness—a large sample. The larger the sample, the tighter the sampling distribution, and the smaller the standard error. Which brings up a mysterious behavior of Nature. For some reason, she demands that we pay to gain insight into her secrets, never sending a free lunch our way. If you want to estimate a parameter, you have to invest some effort by conducting a study. And the more you invest (the larger the sample), the more you *might* get in return. There is power in numbers.

Two principles guide causal inquiry: strive for an unbiased estimator of the causal parameter and strive for an estimator that produces a tight sampling distribution—for a small standard error, that is. Unfortunately, however, the two goals tend to conflict, creating "the bias-variance tension" (or the "bias-standard error tension".) A familiar example shows up in the context of confounding. To eliminate confounding bias by conditioning, we may stratify the sample on confounders, compute stratum-specific estimates, and collapse these estimates into one average. In this process, we often sacrifice the size of the standard error because each stratum contains only a fraction of the sample and re-assembling the pieces into one average might not remedy the damage of splitting. Had we wanted, on the other hand, to minimize the standard error, we should have kept the sample intact. But to keep the sample intact means to allow the bias of confounding. What remedy is offered by Nature? You guessed it. You can always pay more: get a larger sample.

It is natural to ask next "How large of a sample is large enough?" and I am tempted to reply with a question: "How rich of a person is rich enough?" Neither material richness nor richness of scientific knowledge comes with a price cap. There is no price cap for knowledge because certain knowledge is unattainable—Nature has not put it for sale.

Probability density functions

We are a few sections away from a critical discussion of the *p*-value, a randomness-related machinery that has transformed much of causal inquiry into an automated procedure. Data go in, a *p*-value comes out, and a verdict is issued: "not statistically significant", "almost statistically significant", "statistically significant", or "highly statistically significant". But before we dive into the p-value swamp—and try to get out—let's find out how we get to that magic number. The explanation, unfortunately, is a little long. Hang on.

The road starts with a sampling distribution (or a replication distribution, if you prefer) of point

estimates around the expected value of an estimator. To keep things simple, let's consider again an example of random sampling, and make three assumptions:

The parameter of interest is descriptive, the percentage of smokers in some town, and its value is 50%. We estimate that (unknown) number from a random sample of 100 residents. We are allowed 100 replications, which means 100 samples of 100 residents and, therefore, 100 estimates of the percentage of smokers. (Note that after each replication, we return the sample back to the pool before sampling again.)

One way to find a pattern in the data is to order the list of estimates from the lowest value (possibly 0% of smokers in some samples) to the highest value (possibly 100%) and calculate the frequency of estimates that fall in various ranges. For example, we might find out that one tenth of the samples shows a percentage of smokers in the interval [0%, 25%] and that one-half shows a percentage in the interval [25%, 50%]. Evidently, each of these frequencies should somehow reflect a probability: the probability of getting a sample in which the percentage of smokers resides in some interval.

Figure 2 shows an example of hypothetical, but possible, results. Ranges of estimates that are close to the parameter show up more often than remote ones, as we expect, and the sum of the four bars is equal to 1, as it should.

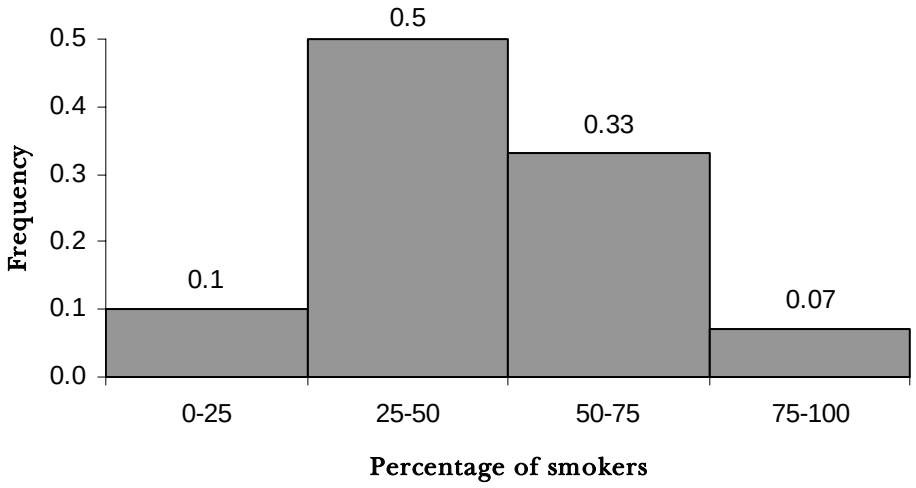


Figure 2. Hypothetical frequency (proportion) of samples showing a percentage of smokers in four successive ranges. Each range is 25 percentage points wide.

Let's shrink, next, the intervals from a width of 25 percentage points to a width of 10-percentage points, and thereby create a sequence of ten successive intervals between 0% and 100% (Figure 3).

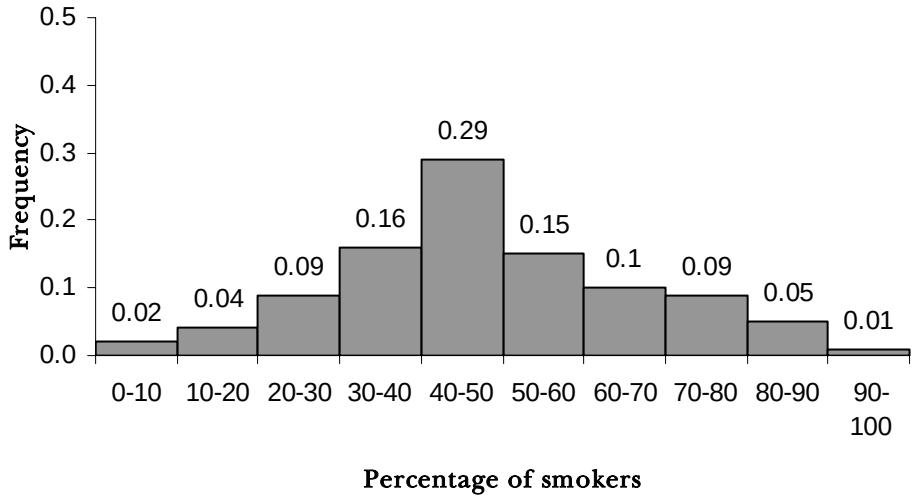


Figure 3. Hypothetical frequency (proportion) of samples showing a percentage of smokers in ten successive ranges. Each range is 10 percentage points wide.

As before the height of each bar estimates the probability of getting a percentage that belongs to that 10-percentage point range. The total bar area is, again, equal to 1.

Imagine now that you continue this exercise in the following way: 1) Instead of sampling 100 residents just 100 times, you sample 100 residents infinite times; 2) As you proceed with re-sampling, you continually shrink the width of the intervals until the tops of adjacent bars blend, and it's hard to tell where one bar ends and another begins. At the end of this mental exercise all you can see is a silhouette—a curve that separates the area below (which still covers a probability of 1) from the area above (Figure 4.) The curve and the area below are called *probability density function*.

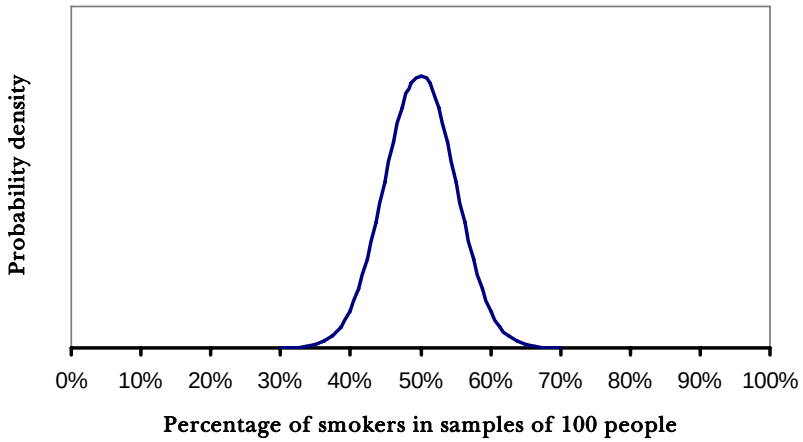


Figure 4. A probability distribution of point estimates from 100-people samples around a hypothetical percentage of smokers in some town (50%). The curve and the area below are called "probability density function".

If you were not born with excessive passion for mathematical ideas, the name might sound intimidating. But it might turn out less intimidating after I clarify the word "function" and explain how to read the graph of Figure 4 (if you don't know already.)

"Function" is an input-output idea in math. We say that Y is a function of X, or Y = f (x) in notation, if we can compute the value of Y from the value of X, display the points (X, Y) in a graph, and connect them by a line. A kernel of these ideas applies here as well.

If you choose a percentage on the X-axis of Figure 4 (say, 40%), draw a vertical line until you meet the curve, and turn left to the Y-axis, you may find the Y-value of that point. In a probability density function, however, the value of Y is not interesting. The Y-axis does not contain the information of interest because we drew a curve to approximate the top border of a sequence of tiny bars, not to connect a sequence of data points. Unlike a classical function, the information of interest is depicted as an *area* under the curve, but that area is zero for X=40% or for any other percentage! In geometry, a vertical line does not occupy any area.

To obtain useful information from the graph of a probability density function, we must specify *an interval* for X and "read" (calculate somehow) the area under the curve for that interval. In Figure 4 the simplest example is the interval [0%, 100%] for which the area under the curve depicts a probability of 1. For any other interval, the area under the curve occupies a fraction of the total area and that fraction makes up a probability value. For instance, about 95 percent of the total area is contained between 40% and 60%, which means that the probability of getting a percentage of smokers in that interval is about 0.95. Of one million samples, about 0.95 million will contain 40% to 60% of smokers.

The analogy between a probability density function and a "classical" function is summarized in Table 1.

Table 1. Analogy between a probability density function and a classical function, as depicted on a graph

Classical function (for example, Y=2X)	**Probability density function** (for example, a sampling distribution)
Input: X, a number on the X-axis	Input: the interval [a, b] on the X-axis
Desired output: Y, a number on the Y-axis	Desired output: the probability of "falling" in the interval [a,b]. 0<Probability≤1
Graphical output: read on the Y-axis by finding the point (X,Y) on the line that depicts the function	Graphical output: the area under the line between "a" and "b", expressed as a fraction of the total area under the line

Where does "density" come from?

Figure 5 shows, again, a hypothetical sampling distribution of random samples of 100 residents (left panel) and another hypothetical distribution of samples of 50 residents (right panel). These two distributions differ but in either case the total area under the curve still displays a probability of 1. Now, think for a moment about probability of 1 as a mass-like quantity, perhaps some gel, and about the total area under the curve as a "smear" of 1 unit of gel. When the distribution is tighter, as in the left panel, the smear must be *denser* at the center and *lighter* at the tails. For example, the area under the curve for the interval [40%, 60%] contains "more probability" in the left function than in the right one; the left function is denser at the center. In subject matter language, the meaning is this: The probability of getting an estimate in the range of 40% to 60% is larger for a sampling distribution of 100 residents than for a sampling distribution of 50 residents.

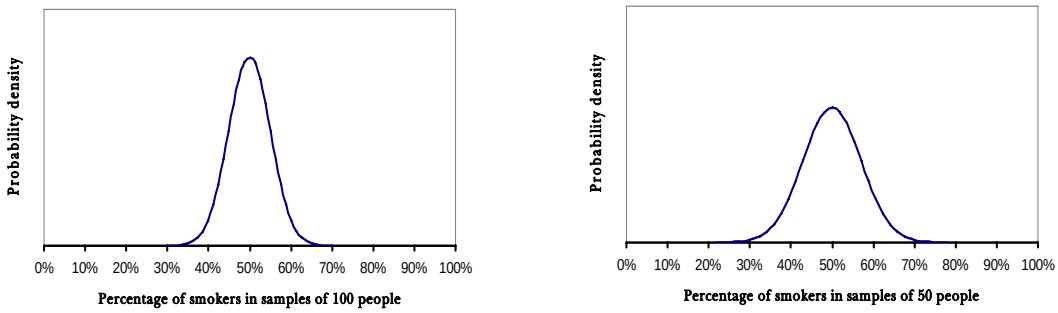


Figure 5. Two probability density functions of point estimates around a hypothetical percentage of smokers in some town (50%): left panel, 100-resident samples; right panel, 50-resident samples.

Converting an estimate to a Z-value

Our discussion so far has been purely theoretical, setting all practical matters aside. You cannot, however, take the collection of estimates from some estimator and display its sampling distribution, because you got only one number: the estimate from the study. And even if it were possible to replicate some process several times, it is still impossible to replicate any process infinite times and display a probability density function of the estimator. Fortunately however, statistical theory has found a two-step solution to this problem: First, statisticians found several well-characterized probability density functions. Second, they showed us how to convert the unknown sampling distribution of various estimators to one of those well-described functions. My abstract words call for an example.

You might have learned about the standard normal distribution, an example of a probability density function for what is called the *Z statistic* (Figure 6.) Details aside, the curve of this function is symmetrical and bell-shaped, and we know the area below for any interval of interest. For example, about 95% of the area under the curve (probability of 0.95) is located in the interval [Z=–2, Z=2] and about 2.5% of the area (probability of 0.025) is located in the interval [2, ∞), a tail area. Now, it turned out that you can transform the sampling distribution of the mean difference to the *Z* distribution by following three steps:

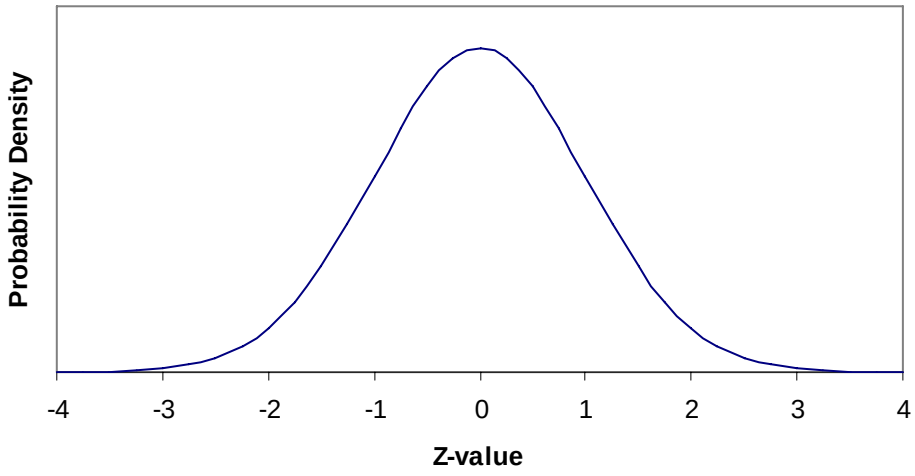


Figure 6. The standard normal distribution: a well-characterized probability density function

First, draw in your mind a sampling distribution of many estimates of the mean difference, from an unbiased estimator, around the unknown causal parameter—the true mean difference. What shape does it have? If you pictured a bell-shaped distribution, similar to the *Z*-distribution, you were right. Second, make up a number for the causal parameter, say 0. That number resides at the center of your sampling distribution. Third, subtract the value of the causal parameter (0 in this case) from each estimate and divide the result by the standard error of the sampling distribution.

If your sample was "large enough", you have just computed (in your mind) a list of *Z*-values. In notation:

$$\frac{\text{Estimated mean difference} - 0}{\text{SE (mean difference)}} \approx Z\text{-value} \qquad \text{Equation 1}$$

But this doesn't get us too far. We have only one estimate from that hypothetical distribution (the mean difference from the study), and we don't know the standard error. Fortunately, again, statistical theory has found a way around one of these problems: we can *estimate* the standard error of the sampling distribution of the mean difference using data from a single sample.

We are almost there. Take the point estimate of the mean difference (from a large study), divide it by the (estimated) standard error and you get a Z-value. Then, look at the Z-distribution, find the location of your Z-value on the horizontal axis of that probability density function, and tell yourself quietly: "Assuming the true mean difference is zero and my estimator is unbiased, this is the location of my estimate on a sampling distribution of the mean difference." We'll see later what use you might make of that whispering.

What about the sampling distribution of ratio measures of effect, such as the odds ratio or the rate ratio?

Unlike the mean difference, ratio measures of effect do not display a nice bell-shaped curve, but we can easily handle the problem by switching to a logarithmic scale. For example, if the sample is "large enough" the sampling distribution of the log(OR) looks bell-shaped, just like that of the mean difference, so we can convert the *log* of the estimated OR to a Z-value as before:

If log (OR_{causal}) happened to be zero (implying OR_{causal}=1: no effect), then,

$$\frac{\text{Estimated log(OR)} - 0}{\text{SE [log(OR)]}} \approx \text{Z-value} \qquad \text{Equation 2}$$

The last formula is a bit scary but all that I have done is to substitute "log (OR)" for "mean difference". Again, there are ways to estimate the "standard error of the sampling distribution of the log of the odds ratio" (a breathtaking phrase).

Converting an estimate to a χ^2 value

When we wish to convert the sampling distribution of ratio measures of effect, another probability density function, the chi-squared distribution, often becomes handy. The chi-squared statistic is, however, a trickier concept than the Z-statistic because it is a family of distributions, each linked to something called "degrees of freedom". And "degrees of freedom" is one of those statistical ideas that are difficult to grasp. But there is no need to worry: I will mention only the first member of the family—a chi-squared statistic on one degree of freedom (written $\chi^2_{1\ d.f.}$). Its probability density function is shown in Figure 7. Keep in mind that the area outlined by the curve, the X-axis, and the Y-axis contains a probability of 1.

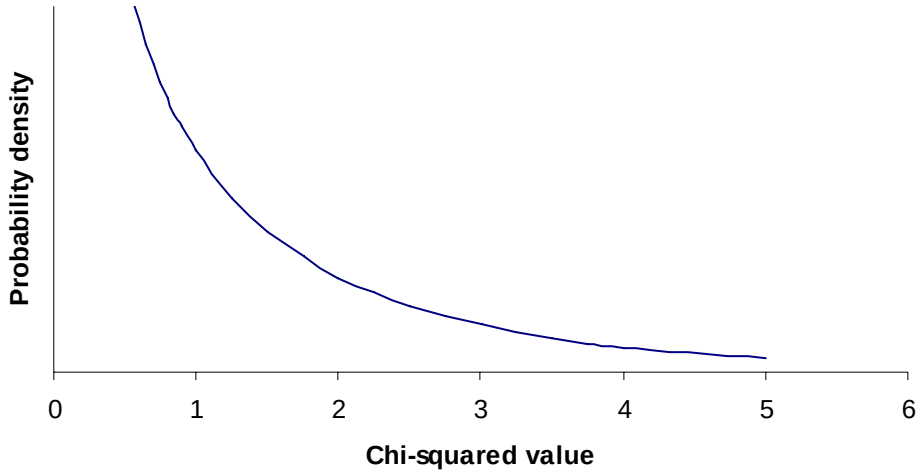


Figure 7. The chi-squared distribution (1 degree of freedom)

It turns out that we can convert ratio measures of effect, such as the odds ratio, into a chi-squared value (1 d.f.) in the following way:

If log (OR_{causal}) happened to be zero (implying OR_{causal}=1: no effect), then,

$$\frac{[\text{Estimated log(OR)}]^2}{\text{Var [log(OR)]}} \approx \chi^2_{1\text{ d.f.}} \qquad \text{Equation 3}$$

Take a moment to compare equations 3 and 2. The left hand side of equation 3 is the square of the left hand side of equation 2 and, therefore, the same relation must hold for the right hand sides—that is, $\chi^2_{1d.f.} = Z^2$. So, we have just learned that the Z-statistic and the chi-squared statistic on 1 degree of freedom are related.

One last technical point, which we'll need later, has to do with the relation between the areas under the chi-squared curve and the Z-curve. If $\chi^2_{1d.f.}$=4, for example, the tail area in the interval [4, ∞) contains a probability of 0.05. But if you switch to a Z-distribution by taking the square root of 4, you will find that the right tail interval [2, ∞), or the left tail interval (–∞,–2], contain only half that probability—only 0.025. Therefore, the right tail on the chi-squared distribution (which is not symmetrical) corresponds to two tails on the bell-shaped Z-distribution. In simplified statistical notation, my words say the following.

$$\Pr(\chi^2_{1d.f.} > 4) = \Pr(Z > 2) + \Pr(Z < -2) = 2\Pr(Z > 2)$$

And in general $\Pr(\chi^2_{1d.f.} > C) = \Pr(Z > \sqrt{C}) + \Pr(Z < -\sqrt{C}) = 2\Pr(Z > \sqrt{C})$

Now we are ready to march into the *p*-value world.

Computing a *p*-value

Suppose that a randomized trial of some drug (streptokinase) has, surprisingly, produced an odds ratio for death of 1.47 and a hazard ratio of 1.44. Table 2 shows the conversion of these estimates to a chi-squared value.

Table 2. Computing two *P*-values

	Odds Ratio (OR)	Hazard Ratio (HR)
point estimate	1.47	1.44
Log_e (point estimate)	0.385	0.365
SE [$\log_e$(…ratio)]	0.25	0.19
Var [$\log_e$(…ratio)]	0.0625	0.036
$\chi^2_{1d.f.}$	2.37	3.70
p-value	0.12	0.05

The *p*-values in the last row are simply the area under the curve of the chi-squared distribution to the right of each chi-squared value. Specifically, the probability of the interval [2.37, ∞) is 0.12 whereas the probability of the interval [3.70, ∞) is 0.05. To see what these numbers might mean in the context of the streptokinase trial, let's focus on the hazard ratio column (*p*=0.05).

Picture the bell-shaped sampling distribution of the $\log_e$(HR) and assume that zero resides at its center (which means HR_{causal}=1). Then, transform that distribution in your mind to a Z distribution and ask yourself which value on the horizontal axis of the Z-distribution corresponds to the hazard ratio of 1.44 from the trial. The answer of 1.9 can be computed from the data in Table 2. Since the point estimate of 1.44 corresponds to a chi-squared value of 3.7, it must also correspond to Z=1.9 (or –1.9) which is the square root of 3.7. You can also, of course, calculate the Z-value directly according to equation 2:

$$\frac{\text{Estimated log(HR)} - 0}{\text{SE [log(HR)]}} = \frac{\log(1.44)}{0.19} = \frac{0.365}{0.19} = 1.9$$

If you look up a table of the Z-distribution, you will find that the area under the curve to the right of 1.9 (or to the left of –1.9) contains a probability of about 0.025. So where does that *p*-value of 0.05 (Table 2) come from?
The answer has to do with the relation I mentioned earlier between the area under the curve of the chi-squared distribution and the area under the curve of the Z distribution. A *p*-value of 0.05 for a chi-square value of 3.7 indeed quantifies the area to right of 3.7, but that area matches *two* tails on the symmetrical Z-distribution: Z>1.9 *and* Z<–1.9.

Now we need to retrace our steps to the hazard ratio:

Z= 1.9 *corresponds to* log (HR)= 0.365 *which corresponds to* HR=1.44
Z=–1.96 *should correspond to* log (HR)= –0.365 *which corresponds to* HR=0.69
(Note that 0.69 is just the inverse of 1.44: 1/1.44=0.69).

Therefore, Pr (Z>1.9 or Z<–1.9) = 0.05 on the Z-distribution

implies

Pr [log (HR) > 0.365 or log (HR) < –0.365] on the bell-shaped distribution of log (HR)

which implies

Pr (HR>1.44 or HR<0.69) = 0.05 on the sampling distribution of the hazard ratio.

The last derivation captures the meaning of the *p*-value. In words: If HR_{causal}=1, if the estimator is unbiased, and if we have some replication in mind, the following may be said about that sampling (replication) distribution: The probability of getting a point estimate in the interval [1.44, ∞) *or* in the interval [0, 0.69] is 0.05.
A parallel statement may be made about the sampling distribution of the odds ratio from that trial: The probability of getting a point estimate in the interval [1.47, ∞) or in the interval [0, 0.68] is 0.12.

What use can we make of these probability statements?

Rather than answering this question, I will tell you first what you cannot use them for. You are *not* allowed to make any of the following statements—they are all false:

- The probability of getting a hazard ratio of 1.44 by chance is 0.05.
- The probability that HR_{causal}=1 is 0.05.
- If we claim that HR_{causal} does not equal 1, the probability of our making an erroneous claim is 0.05.
- Since the *p*-value in the hazard ratio column is relatively small (0.05), the point estimate (1.44) is credible; it should be taken seriously.
- Since the *p*-value in the odds ratio column is relatively large (0.12), the point estimate (1.47) is not credible; it should be discarded.

I think I have listed above all or most of the statements we would like to make with the help of the *p*-value, but unfortunately none is technically permissible. What is left is two options: we may use the *p*-value as a measure of evidence against the so-called "null hypothesis" (for example, against the hypothesis that $HR_{causal}=1$), *or* we may use the *p*-value as a rule for deciding between two hypotheses: $HR_{causal}=1$ and $HR_{causal}\neq 1$. The first idea was proposed by Fisher; the second, by Neyman with help from Egon Pearson. All three statisticians are counted among the founders of frequentist statistics (during the first half of the twentieth century), but they did not speak in one voice. In fact, Fisher and Neyman were academic enemies, a point long forgotten by those who put one foot in Fisher's world and another in Neyman's (to be explained shortly.)

Fisher's proposal: *p*-value as evidence

If $HR_{causal}=1$ and the estimator is unbiased, the probability of getting a point estimate in the interval [1.44, ∞) or in the interval [0, 0.69] is 0.05. Since this probability is small—said Fisher and all of his followers—only two explanations may hold: either $HR_{causal}=1$ and something rare has happened in the study or our estimate does *not* belong to a sampling distribution around 1 (that is, HR_{causal} is *not* 1.) And since the *p*-value is relatively small, only 0.05, the latter explanation should be better: $HR_{causal}\neq 1$. The smaller the *p*-value, the stronger the evidence against $HR_{causal}=1$, argued Fisher. How small is small enough? No arbitrary number is sacred, of course, but for some reason of all unsacred numbers 0.05 seems to have magical properties. For millions around the world, a *p*-value of 0.05 makes the difference between gathering enough evidence and not gathering enough evidence.

What if the *p*-value from the streptokinase trial were "large"—say, 0.8?

The Fisherian answer is clear: We have learned nothing from the study and have wasted our time. The study provides no basis for any statement about HR_{causal} because the lack of evidence against $HR_{causal}=1$ does not imply positive evidence for $HR_{causal}=1$. (Just as the lack of evidence to incriminate a defendant does not imply his innocence.) Unfortunately, however, this part of Fisher's thought is often forgotten by those who rely on the *p*-value when they claim that "A does not cause B", or that "A is not associated with B". Maybe A does not cause B, indeed, but a large *p*-value cannot form the basis of that claim—according to Fisher.

Fisher's proposal to consider the *p*-value as a measure of evidence against the causal null sounds simple and intuitive, and surely appeals to human psychology. But it contains three kinds of shortcomings: technical, logical, and scientific.

On a technical level, our *p*-value of 0.05 is a statement about the probability of data given that the causal null is true, or in notation: $\Pr(\text{data} \mid HR_{causal}=1) = 0.05$. Read: probability of data is 0.05, given that $HR_{causal}=1$. To speak about evidence against the null, however, we have to reverse the order and talk about the probability of the causal null being true given the data. (Again, as in court: The probability that the defendant is innocent, given evidence gathered at the crime scene.) But you cannot simply turn a probability on its head and write something like $\Pr(HR_{causal}=1 \mid \text{data}) = 0.05$. If you wish to talk about the probability of a hypothesis, rather than the probability of data, you have to switch to another school of statistical thought (Bayesian statistics), but that school doesn't speak the language of *p*-values. Furthermore, and contrary to prevailing thinking, simulation work has shown that a *p*-value around the magical threshold of 0.05 provides only weak evidence against the null hypothesis. If *p*-values are to be used as evidence at all, 0.05 is bad cutoff for "enough evidence"; 0.01 would be better.

On a logical level, two problems are hidden in Fisher's logic. First, it is problematic to argue from low probability of data under some hypothesis to falsehood of that hypothesis. A point estimate of 1.44 may indeed be rare if $HR_{causal}=1$ but it is a valid member of the sampling distribution and it does not logically contradict the truth of $HR_{causal}=1$. No solid bridge connects the idea of implausibility of data, which is a fuzzy psychological state, and falsehood of a hypothesis, which is a piece of reality.

Second and even more important, our *p*-value of 0.05 does *not* quantify the probability of observing a hazard ratio of 1.44, as so many erroneously think. That probability remains unknown (and is technically zero on a density function.) The number 0.05 is the probability of getting point estimates at the two tails of the sampling distribution—*estimates that are larger than 1.44 or smaller than 0.69*. But none of these estimates was computed, of course! Fisher's logic rests on low probability of imaginary results and his argument should be summarized as follows: We conclude that $HR_{causal}=1$ is false not because we know how improbable the estimate 1.44 is (if $HR_{causal}=1$), but because we know how improbable it is to get *unobserved* estimates in the intervals [1.44, ∞) and [0, 0.69]. I can't say that the argument is unequivocally wrong, but its logic is far from certain

in many minds. For example: is it logical to use the p-value (the entire tail area of the distribution) as the basis for our claim that we have observed one rare point estimate (if $HR_{causal}=1$)? Or: if $HR_{causal}\neq 1$, as we try so hard to claim, how can we logically talk about the probability distribution of estimates around $HR_{causal}=1$? If $HR_{causal}\neq 1$, the distribution from which our *p*-value was computed does not exist at all.

Besides technical and logical shortcomings, Fisher's proposal contains a scientific flaw, which in my view is fatal. Fisher (and Neyman too) mistakenly assumed that empirical science should issue a verdict on the causal null, trying to show (or decide) that it's false. Yet that verdict adds little to conjectural knowledge of causal reality. But before discussing the scientific angle, let's see what Neyman has proposed to science.

Neyman's proposal: rules for decision-making

Neyman did not consider studies to be evidence-generating machines, nor was he focused on methods to falsify the causal null. For Neyman, the inference from any study is a matter of decision: do we decide that $HR_{causal}=1$ or do we decide that $HR_{causal}\neq 1$?

Whichever decision we make, we are taking the risk of making a mistake. If we decide that $HR_{causal}\neq 1$ we might commit one kind of a mistake, called type I error: erroneously claiming that the causal parameter is not 1 when in fact it is. On the other hand, if we decide that $HR_{causal}=1$ we might commit another kind of a mistake, called type II error: erroneously claiming that the causal parameter is 1 when in fact it is not. Of course, regardless of which decision we make, we cannot know whether we have made a mistake.

Realizing that ignorance cannot be conquered, Neyman proposed what he considered to be a sound approach to decision-making: a method to control the frequency of mistakes "in the long-run". His method deals with both kinds of errors, but I will highlight the first.

In the context of the streptokinase trial, Neyman prescribes the following actions:

1. Before the trial is initiated, state the type I error frequency you are willing to tolerate *in science*. Call it α (expressed as a proportion). Likewise, state the type II error, β.

2. After the trial is completed, compare the *p*-value to α and use the following two rules of decision:
    a. If $p\leq\alpha$, reject $HR_{causal}=1$, accept $HR_{causal}\neq 1$, and report your type I error frequency, α.
    b. If $p>\alpha$, reject $HR_{causal}\neq 1$, accept $HR_{causal}=1$, and report your type II error frequency, β.

(I should note that Neyman rejected any reference to *p*-values, and would have stated the inequalities above in the language of the Z-statistic or the chi-squared statistic, using something called a "critical value of the statistic" instead of α. Although the two approaches are miles apart philosophically, they are practically exchangeable in many examples.)

If you consistently follow these rules and use the same value of α, your type I error frequency will not exceed that α. Of all *true* null hypotheses that happen to be tested, only $100x\alpha$ percents will be rejected (erroneously, of course, since they are true.) But be careful to not confuse the correct denominator "all true null hypotheses that happen to be tested" with a false one: "all null hypotheses that end up rejected". Unfortunately, we are able to count only the latter—decisions to reject—because we can't tell which tested null hypotheses are true ones (Figure 8.)

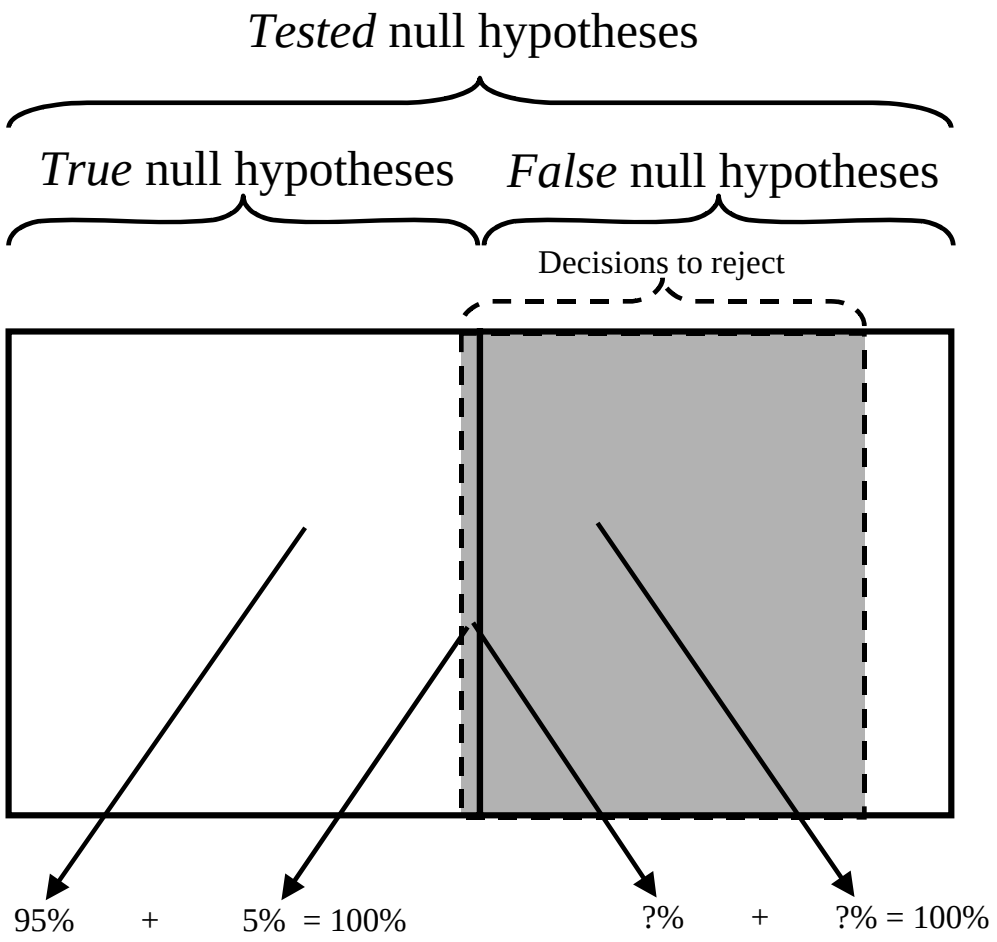


Figure 8. A graphical illustration of the type I error frequency (α=0.05). Notice the unknown frequency of an erroneous decision to reject the null among all such decisions. The ratio of true null hypotheses to false ones is also unknown (illustrated here as equal size squares.)

Question: Suppose you followed Neyman's rules in the streptokinase trial and ended up rejecting $HR_{causal}=1$ and accepting $HR_{causal}\neq 1$. What is the probability that you committed the error called type I—mistakenly rejecting the causal null?

Well, that question embeds two faults: First, it does not have an answer because the α value refers to *true* null hypotheses, not merely to rejected ones. Second, in the world of frequentist statistics, where probability thrives on empirical or theoretical replication, this question must be declared meaningless. There is no repeated decision in one trial, and therefore, there is no frequency of wrong decisions and no type I error frequency. We decided once and either have made a mistake or have not. So keep in mind two *wrong* answers to the question above, which many would give: 1) the *p*-value; 2) the α value. To reiterate: neither the *p*-value nor the value of α quantifies the type I error in the streptokinase trial.

The last exchange says it all about Neyman's proposal. Unlike most of his colleagues in science, Neyman was not interested in the inference from any particular study. He was interested in worldwide statistical behavior that could control the overall frequency of erroneous inference from data. But no scientist who respects her profession would subscribe to Neyman's proposal had she been bluntly told what she is committed to do: follow a rule of decision to serve science as a whole, but learn nothing from your own study! Neyman's proposal is not a statistical tool at the service of scientists; it is a prescription for how all of science should be conducted. And a bad one—to many minds.

The point I have just made may be sharpened with two extreme examples: If we state that α=0.05 before the trial is initiated, both a *p*-value of 0.049 and a *p*-value of 0.0001 would lead to the same decision. Both should be reported as "The null was rejected; α=0.05" (and neither *p*-value should be reported at all!) In contrast, a *p*-value of 0.049 and a *p*-value of 0.051 would lead to opposite decisions: one should be reported again as "The null was rejected; α=0.05" but the other should be reported as "The null was accepted" (or not rejected.) Again, according to Neyman neither *p*-value should be reported.

Neyman's proposal for science might occasionally sound appealing: forget about evidence and just follow a prescribed behavior to limit the frequency of type I error "in the long run". But when reality comes into play, we often find ourselves saying just the opposite: forget about that rigid, mechanical rule and just look at the evidence. And the evidence in the last two examples—the *p*-values—is hard to ignore: the numbers 0.049 and 0.051 are not that different, yet they lead to opposite decisions, whereas the numbers 0.049 and 0.0001 are *very* different, yet that difference makes no difference. As a result of this cognitive dissonance, statistical practice has turned into an exercise in changing hats: When we wish to play by the evidence, we put on Fisher's hat (the shortcomings of his proposal notwithstanding), and when we wish to play by a rule of decision, we put on Neyman's hat (his indifference to evidence notwithstanding). And sometimes, fortunately, we can pretend to wear both hats at the same time. Now, I am moving uncomfortably in my seat: is this dancing in two incompatible worlds called the method of causal inquiry? Are we allowed to look at the *p*-value, decide which world better fits our psychological wellbeing, and shamelessly call it the method of science?

If you eventually decide to join one of these worlds, keep Table 3 in mind. And never allow your mind to switch to the enemy's column.

Table 3. A contrast of Neyman's proposal with Fisher's

Results and questions	Fisherian reply	Neymanian reply
What is your pre-study statement about error frequencies?	Silence	α, say 0.05 β, say 0.10
p=0.051 What is your inference?	Some evidence against $HR_{causal}=1$	$HR_{causal}=1$ Type II error frequency: 10%
p=0.049 What is your inference?	Some evidence against $HR_{causal}=1$	$HR_{causal}\neq1$ Type I error frequency: 5%
p=0.001 What is your inference?	Strong evidence against $HR_{causal}=1$	$HR_{causal}\neq1$ Type I error frequency: 5%
p=0.8 What is your inference?	Silence	$HR_{causal}=1$ Type II error frequency: 10%

The scientific flaw

Suppose we rejected the null theory, $HR_{causal}=1$, in the streptokinase trial by following Neyman's rule of behavior or Fisher's approach to evidence. What have we learned about the value of the causal parameter besides that it's different from 1? The surprising answer is—nothing!

We may be tempted to think that we have somehow advanced the status of the point estimate from untrustworthy to trustworthy, but there is no greater falsehood than this. Even if the point estimate could help us to draw inference about the truth of $HR_{causal}=1$, we cannot turn back and draw inference from $HR_{causal}\neq1$ to the truth of $HR_{causal}=1.44$. If we say that argument A ("$HR_{estimated}=1.44$") implies argument B ("$HR_{causal}\neq1$"), we cannot recruit argument B in support of argument A (unless you have other support for B.) To do so is to argue that A implies A, which is either trivial or nonsense. In short, $HR=1.44_{estimated}$ cannot imply or certify or endorse itself by implying $HR_{causal}\neq1$.

That faulty chain of inference, from $HR_{estimated}=1.44$ to $HR_{causal}\neq1$ to $HR_{causal}=1.44$, is imprinted in millions of minds thanks to awkward terminology called "statistical significance". When the *p*-value is small enough, we are told to say that the point estimate is "statistically significant", which sounds like declaring the point estimate as valid or credible or "real". But that was not the original meaning of the statistical term, when coined many years ago. At that time, "significant" was derived from "signified", and the phrase implied that the point estimate *signified evidence against the null*—not that it had any intrinsic significance. As I just wrote, the inference rides on a

one-way road of (questionable) evidence and logic: from $HR_{estimated}$=1.44 to a small *p*-value to $HR_{causal}\neq1$. The road ends at $HR_{causal}\neq1$ and no U-turn is logically permitted.

Which brings us to the scientific flaw. Fisher and Neyman and many others have placed so much emphasis on falsifying a causal null because they mistakenly equated causal inquiry with sorting causes from non-causes: with knowing whether the contrast between streptokinase and placebo has *any* effect on death. This tenet undoubtedly corresponds to the layperson's view of causation, but is scientifically wrong. We learn almost nothing from "$HR_{causal}\neq1$" because that knowledge entails every other possible value of the causal parameter, including those that are trivially small. And nobody, of course, would claim to have learned much from $HR_{causal}\neq1$ if the (unknown) value of HR_{causal} were sufficiently small, say 0.99.

One escape route might be to claim that we are not interested in furnishing evidence against *precise* causal null, such as $HR_{causal}=1$. We are really interested in evidence against a null of the form $0.98<HR_{causal}<1.02$, and under certain conditions evidence against precise null also carries to a small interval null. This is technically true, but the epistemological problem does not vanish. If we state that the interval of no interest is [0.98, 1.02], do we seriously claim that knowledge of $HR_{causal}=1.02$ is uninteresting and that knowledge of $HR_{causal}=1.021$ is suddenly interesting? Where continuity of numbers exists, dichotomy must logically fail. To sum up: to know something about hidden causal reality amounts to knowing (conjecturally) the value of the causal parameter, and a *p*-value does not deliver these goods. From a scientific viewpoint it is useless. Rejection of the null does not add knowledge.

Much of my criticism of Fisher's proposal, Neyman's proposal, and *p*-values is not originally mine, and you can hardly find written rebuttals—perhaps because it's hard to rebut the arguments. But it does not matter at all, because no words will eliminate *p*-values and null hypotheses testing from science. They simply serve too many psychological needs and too many societal needs. So rather than trying to fight them through logic and reasoning, we should invest effort in studying why they are hopelessly entrenched in people of reason. "The *p*-value addiction" should make a good PhD dissertation in sociology or psychology.

Confidence interval

Every elementary course in statistics explains how to compute a 95% confidence interval around a point estimate from a bell-shaped distribution. For a large enough sample, take the standard error (SE) and multiply it by 1.96. Subtract the result from the point estimate and you get the lower end; add to the point estimate, and you get the upper end. For example:

Lower limit of the mean difference:
mean difference – 1.96xSE(mean difference)

Upper limit of the mean difference:
mean difference + 1.96xSE(mean difference)

The idea is fairly simple. When the sampling distribution is bell-shaped and resembles that of the Z-statistic, 95% of the area under the curve is contained between (–1.96xSE) and (+1.96xSE).

Things get a little more complex when we switch to ratio measures of effect. The hazard ratio, for example, does not display a bell-shaped distribution, but the *log* version does. We may, therefore, compute the 95% confidence interval for log(HR) in the same way:

Lower limit of the log(HR):
log(HR)–1.96xSE[log(HR)]

Upper limit of the $\log_e$(HR):
log(HR) + 1.96xSE[log(HR)]

To return to the hazard ratio scale and get the limits for the HR itself, we have to raise *e* to the power of these terms:

Lower limit of the HR: $e^{\log(HR)-1.96\,xSE[\log(HR)]}$

Upper limit of the HR: $e^{\log(HR)+1.96\,xSE[\log(HR)]}$

The multiplier 1.96 is derived from the confidence level—here 95%. If you want an 80% confidence interval, the multiplier will be smaller and if you want 99%, it will be larger. The larger the percentage, the wider the interval, which makes an intuitive sense: if you wish to have more confidence, you have to throw a wider net. These days it's hard to find anything in the literature besides 95% confidence intervals, but there is nothing more sacred in 95% than, say, 93.3%.

To illustrate the arithmetic, let's follow again the estimates from the streptokinase trial. The hazard ratio for death was 1.44 and the estimated standard error was 0.19. Therefore, the lower limit of a 95% confidence interval is about 1.0 ($e^{\log(1.44)-1.96x0.19}$) and the upper limit is about 2.1 ($e^{\log(1.44)+1.96x0.19}$).

That was the easy part of the story. What is not so easy is to inject a truthful interpretation into what we have done and to avoid false ones. The truth, unfortunately, is this: If the estimator is unbiased; if you were repeatedly generating point estimates from that estimator; and if you were to construct a 95% confidence interval around each point estimate, you would have gotten many different intervals. Of these, 95% would have contained the causal parameter, HR_{causal}. Five percents of the intervals would not.

We have, however, only one point estimate (1.44) and one confidence interval [1.0, 2.1], so what may we say about the relation between the interval at hand and the causal parameter? The embarrassing answer is—nothing. But since that answer is hard to swallow, three interpretations were made up: one rides on the word "confidence"; another rides on the word "plausibility"; and a third rides on the word "precision". As you will see shortly, all three lead nowhere.

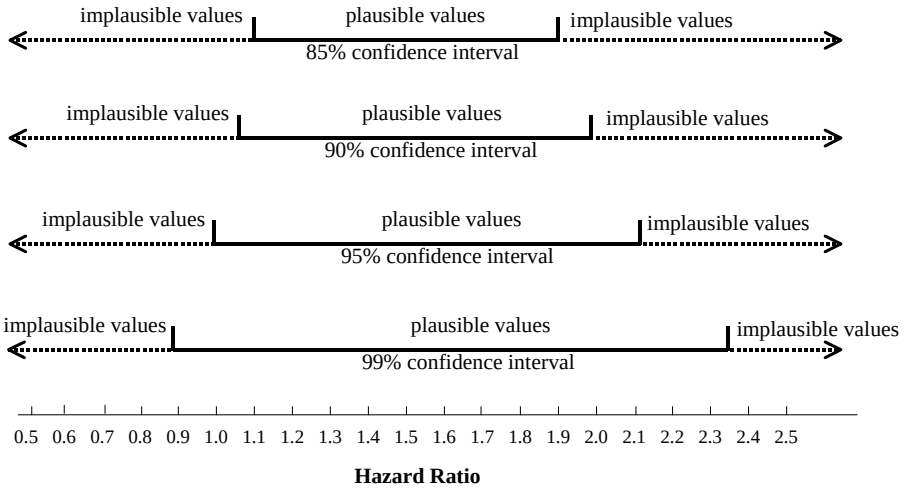
According to the first interpretation we are taught to say that "we have 95% confidence that the interval [1.0, 2.1] contains the causal parameter", which sounds very close to certainty. But where is that 95% coming from? What is the reason for the confidence?

In the business enterprise, we may say that we are 95% confident that a business deal will be successful. In statistics, however, percentage confidence must originate in the math of probability. And there hides the devil. In the frequentist world, the world in which the confidence interval was computed, the probability that the parameter lies in one particular interval cannot be 95%. The causal parameter is a number, such as 1.2 or 0.9 or 1.0, and a number does not form a probability distribution: either the number belongs to the interval [1.0, 2.1] or it does not. To talk about the probability that the causal parameter resides in one specific interval is not different from talking about the probability that the causal null is true. And as you saw earlier, "Pr (HR_{causal}=1 | data) = 0.05" does not belong in frequentist statistics. Nor does its equivalent in the confidence interval department: "Pr (1.0≤HR_{causal}≤2.1 | data) = 0.95" is a forbidden statement.

A second interpretation of a confidence interval appeals to the idea of "plausibility". Instead of saying that we are 95% confident that the causal parameter resides in the interval [1.0, 2.1], we say that the interval displays a *plausible* range for the causal parameter. Values within are plausible and values outside are not. This interpretation survives a little longer than the first one but only because it thrives on the seducing power of the word "plausible". One absurdity becomes evident as soon as we turn to values at or around the limits. That dichotomy forces us to claim that 1.0 and 2.1 are both plausible values of the causal parameter whereas 2.2 is not. Yet 2.2 is a very close neighbor of 2.1—a much closer neighbor of 2.1 than is 1.0—so how can we seriously contemplate the numbers 1.0 and 2.1 in one breadth and build a wall between 2.1 and 2.2 in another?

Another embarrassing finding has to do with the relation between the plausible range and the confidence level. Figure 9 shows the contrast between plausible values and implausible ones according to four levels of confidence. It is easy to see how fragile the idea becomes. To increase or decrease the range of plausible values all that we need to do is to change the level of confidence. In fact, we can shrink or stretch that range to our liking by setting the confidence level lower or higher, respectively. For example, what is not plausible for a 95% confidence interval will become plausible for 97%, or for 99.9%, and so on.

Figure 9. Plausible values and implausible values according to four levels of confidence



The third interpretation capitalizes on the word "precision": the tighter the confidence interval, the more precise the estimate. But what does the adjective "precise" precisely mean in the context of a point estimate? If it means "closer to the truth", as it sounds indeed, then the claim is false. Regardless of the width of the confidence interval, we have no way of knowing whether one estimate lies close to the truth or far from the truth, and no semantic twist can change this unpleasant fact. The *estimator* may be more precise or less precise—referring to the spread of the sampling distribution—but these adjectives cannot be attached to a single estimate. An estimate cannot be called "precise", just as it cannot be called "unbiased".

Furthermore, the adjective "imprecise" is sometimes dishonestly used to turn a 95% confidence interval into a *p*-value because the two are mathematically related: If the null value is contained in the interval, the *p*-value is larger than 0.05, and if the null value is contained in the interval and is close to one end, then the *p*-value is just larger than 0.05. Now, instead of saying that "the result was borderline statistically significant" (whatever that means), one could direct the spotlight to the point estimate and say that it is "imprecise": almost kosher but not quite so. To some minds, such practice amounts to abusing two ideas in one sentence: the idea of a confidence interval and the idea of a *p*-value. I agree.

Where do we go from here?

One route is to close the books on inferential frequentist statistics and switch to another school called Bayesian statistics, which allows us to compute probabilities of hypotheses and to construct credible intervals—the Bayesian counterparts of confidence intervals. Bayesianism, however, is more than a statistical tool and carries its own bag of epistemological and technical difficulties. For example, the probability that a hypothesis is true is not derived from data alone; it is also conditional on prior probability that the hypothesis is true (which means prior beliefs). Having no interest in using data to transform pre-study beliefs about a causal parameter into post-study beliefs, I reject that philosophy of science. Nor does it matter to me what goes into those prior beliefs—whether it is expert opinion, a flat prior, or semi-empirical content. To my mind, scientific knowledge is conjectural (and therefore, fallible) knowledge of hidden reality; it is not the probability that conjectural knowledge is true.

If you wish to stay within the frequentist school, you have to realize that both a *p*-value and a confidence interval are derived from simple arithmetic on two numbers: the point estimate and the standard error. To get the *p*-value, we start by dividing the point estimate by the standard error and to get a 95% confidence interval, we take the point estimate and add and subtract (1.96xSE). We should not be surprised, perhaps, that no arithmetic can extract more information than whatever is already contained in the point estimate and the standard error. So instead of looking for a miraculous tool for statistical inference, let's return to the starting point.

- The key result of a study is the point estimate—a conjectural value of the causal parameter. Neither a p-value, nor a decision about the truth of the causal null, nor a confidence interval may assume a more important epistemological role than that of the point estimate.

- In our search for the causal parameter, we should strive for an estimate from an unbiased estimator. We can never be certain, however, that the estimator is unbiased: We cannot know whether the expected value is equal to the causal parameter or how close they might be.

- In our search for the causal parameter, we should strive for an estimate from an estimator whose sampling distribution is tight. Regardless of how tight that distribution is, however, we can never be certain that our estimate lies close to the expected value. No confidence interval and no credible interval can change that cruel reality.

- In our search for the truth we often have to choose between the magnitude of bias and the size of the variance (the bias-variance tension). No general rule can tell us the "right" balance between the two.

If you accept these tenets of the scientific method, the role of the standard error becomes clear—and modest. Derived from the sample size and inversely related to it, the standard error informs us about the sampling distribution of theoretical estimates, not about the estimate at hand. It is no more than an aid in comparing the quality of several estimators (of the same causal parameter); in choosing between competing estimates; and in temporarily accepting or rejecting a point estimate. Of course, no naïve rule of inference follows the size of standard error, just as no rule of inference can tell us that an estimator is truly unbiased. But in either case the empirical road remains open. If you reject a point estimate on the account of testable bias, your criticism may be tested. And if you reject a point estimate on the account of a large standard error, you or others may get another estimate from a larger sample. That societal rules might prevent us from doing so is a technical

matter—not an epistemological problem. That no final verdict is possible reassures us that the method accords with the conjectural nature of scientific knowledge.

Confidence limit difference and confidence limit ratio

No words will convince scientists and statisticians to replace *p*-values and 95% confidence intervals with something so trivial and non-decisive as the standard error. The battle was lost long ago. If a plain standard error is to return to science eventually, it must ride on the back of a confidence interval: it must present itself as some derivation from a confidence interval (even though the truth is the other way around…). To that end, a clever idea was proposed in 2001.

Since most of us associate a confidence interval with "precision", what matters in precision is the width of the error—not the bounds on the error! Therefore, when the estimator displays a bell-shaped sampling distribution, anyone who takes precision seriously should report the width of the confidence interval: the *confidence limit difference* (CLD). For example, the width of the 95% confidence interval for a mean difference (from a large enough sample) is

upper limit minus lower limit =

(mean difference+1.96xSE(mean difference))–
(meandifference–1.96xSE(mean difference)) =

2x1.96xSE(mean diff.) = 3.92xSE(mean diff.)

This little trick has taken us back to the scale of a standard error, using a multiplier that makes no conceptual difference. The reference point is, of course, zero: a sampling distribution from an infinite sample will have zero standard error. The smaller the quantity (3.92xSE), the tighter the sampling distribution.

Most causal parameters, however, are measured on a ratio scale and their sampling distribution is not bell-shaped, which creates a technical problem. Being a standard deviation, the standard error does not describe well the spread of a skewed distribution, so we have to start with the log version again. For example, for a binary causal contrast, the width of the confidence interval for the log of the hazard ratio is

upper limit minus lower limit =

log(HR) + 1.96xSE[log(HR)] –
{log(HR) – 1.96xSE[log(HR)]} =

2x1.96xSE[log(HR)] = 3.92x SE[log(HR)]

That solution will do, but, unfortunately, we have not used the confidence limits for the hazard ratio itself, as many would like. A little trick will solve the problem: if we *divide* the upper limit for the hazard ratio by the lower limit, we get $e^{3.92xSE[\log(HR)]}$ as shown below:

Upper limit / Lower limit =

$$\frac{e^{\log(HR)+1.96xSE[\log(HR)]}}{e^{\log(HR)-1.96xSE[\log(HR)]}} = e^{3.92xSE[\log(HR)]}$$

The ratio of the upper to lower limit is naturally called the *confidence limit ratio* (CLR). If you examine the formula, you will see that it is simply *e* raised to the power of what we called the "confidence limit difference" on the log scale. Now the reference point has changed to 1. With an infinite sample, SE[log(HR)]=0 and $e^{3.92x0}$=1. The closer the CLR to 1, the tighter the sampling distribution.

To sum up: The frequentist version of statistics can tell us about randomness-related properties of the estimator—of the process that generated the point estimate. It cannot tell us whether the estimate has hit on the causal parameter or by how much it has missed it. Any complaints should be filed with Nature.

Suggested Reading

Anderson DR et al. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wild Life Management 2000;64:912-923

Berger JO, Sellke T. Testing a point null hypothesis: the irreconcilability of P-values and evidence. Journal of the American Statistical Association 1987;82:112-122

Berger JO. Could Fisher, Jeffreys, and Neyman have agreed on testing? Statistical Science 2003;18:1-32.

Cohen J. The earth is round (p< .05). American Psychologist 1994; 49:997-1003

Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. British Medical Journal 1986;292:746-750

Giere RN. Empirical probability, objective statistical methods, and scientific inquiry. In: Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science, Vol. II, pages 63-101.

Gillies D. Philosophical theories of probability. Routledge, New York, 2000

Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate (with discussion). American Journal of Epidemiology 1993;137:485-496

Goodman SN. Toward evidence-based medical statistics. 1: the P value fallacy. Annals of Internal Medicine 1999; 130:995-1004

Goodman SN. Of P-values and Bayes: a modest proposal. Epidemiology 2001;12:295-297

Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. The American Statistician 2003;57:171-178.

Lang JM, Rothman KJ, Cann CI. That confounded P-value. Epidemiology 1998;9:7-8

Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? Journal of the American Statistical Association 1993;88:1242-9.

Lindley DV. The philosophy of statistics. The Statistician 2000; 49:293-337

Oakes M. Statistical Inference. Epidemiology Resources Inc., Chestnut Hill, MA, 1990

Poole C. Beyond the confidence interval. American Journal of Public Health 1987;77:195-199

Poole C. Confidence intervals exclude nothing. American Journal of Public Health 1987;77:492-493

Poole C. Feelings and frequencies: two kinds of probability in public health research. American Journal of Public Health 1988:78;1531-1533

Poole C. Low P-values or narrow confidence intervals: which are more durable? Epidemiology 2001;12:291-294

Rothman KJ. A show of confidence. New England Journal of Medicine 1978;299:1362-1363

Rothman KJ. Significance questing. Annals of Internal Medicine 1986;105:445-447

Sellke T, Bayarri MJ, Berger JO. Calibration of p-values for testing precise null hypothesis. American Statistician 2001;55:62-71