

Understanding P-values and confidence intervals

Introduction

In the previous two chapters, we discussed using the results of randomized trials and observational studies to estimate treatment effects. We were primarily interested in measures of effect size and in problems with design (in randomized trials) and confounding (in observational studies) that could bias effect estimates. We did not spend much time considering the precision of our effect estimates or whether the apparent treatment effects could be a result of chance. The statistics used to help us with these questions — P-values and confidence intervals — are the subject of this chapter.

No area in epidemiology and statistics is so widely misunderstood and mistaught. We cover a more sophisticated understanding of P-values and confidence intervals in this text because 1) it is right, 2) it is important, and 3) we think you can handle it. After all, you have survived three chapters (3, 4, and 8) on using the results of diagnostic tests and Bayes's Theorem to update a patient's probability of disease. So now you are poised to gain a Bayesian understanding of P-values and confidence intervals as well. We will give you a taste in this chapter; those wishing to explore these ideas in greater depth are encouraged to read an excellent series of articles on this topic by Steven Goodman. (Goodman 1999a; Goodman 1999b; Goodman 2001)

Background: the null hypothesis, test statistics, and P-values

Before we can talk about what P-values and confidence intervals mean, we need to review classical statistical significance testing. The basic process is as follows:

1. State an appropriate “null hypothesis” (H_0), a hypothesis of “no effect,” the exact phrasing of which depends on the type of variables and the relationship between them that you wish to investigate. The null hypothesis will be something like: “there is no difference between the means in the two groups” or “the

response rates do not differ” or “there is no linear association between variables A and B.”

2. Choose α , the maximum probability of a Type 1 error that you are willing to tolerate. A “Type 1” error is when you reject the null hypothesis when it is true – that is, conclude that the difference you observed was not due to chance, when in fact it was. (A “Type 2” error is failing to reject the null hypothesis when it is false – that is, concluding that the difference could be due to chance when in fact it is not. The maximum probability of a Type 2 error is β .)
3. Use the results of the study to calculate the value of a “test statistic” with a known distribution if the null hypothesis is true. Examples are a t statistic, χ^2 statistic, or a regression coefficient divided by its standard error. The test statistic you use depends on the design of the study and the type of variables evaluated.
4. Use that test statistic to calculate a P-value. Classically, if the P-value is less than α , you reject the null hypothesis; however, authors of clinical research articles rarely explicitly reject or fail to reject the null hypothesis. More commonly, they will simply report the P-value and consider the result “statistically significant” if P is less than 0.05, otherwise not.

What do P-values really mean?

Unfortunately, the first thing we have to do is help you unlearn some of what you may have learned from other sources. There is no debate about whether the common misunderstanding of the meaning of P-values and confidence intervals is correct. The debate is entirely pedagogical. Does it matter? Is it worth the time it takes to explain what they actually mean? Can you understand the correct meaning of P-values and confidence intervals?

It is an issue one of us (TBN) has debated with some other teachers of evidence-based medicine. Their view is that it does not matter much if people misunderstand this material, because for most people, misunderstanding the correct interpretation of the P-value will not lead to huge misinterpretations of the literature. On the other hand, there are some issues that arise in nearly every research study (like whether it makes a difference if hypotheses are stated in advance, and whether you need to adjust for multiple comparisons) that simply do not make sense if you misunderstand the meaning of the P-value. To those teachers and students who insist that one can get along just fine not really understanding what P-values mean, we would point out that you also can get along fairly well believing that the sun revolves around the earth. It is much more satisfying, however, to learn (and teach) what is right. So here are correct and incorrect definitions of the P-value.

Correct Definition: A P-value is the probability of observing a value of the test statistic at least as extreme as that observed in the study, if in fact the null hypothesis is true.

Incorrect Definition: A P-value is the probability that the null hypothesis is true (i.e., that there is no difference between the groups, no relationship between the variables, etc.), given the results of the study. That is, if $P = 0.05$, there is a 5%

probability that the observed departure from the null hypothesis occurred by chance and a 95% probability that it did not and the observed difference is real.

The difference between the two definitions above may seem subtle, but it is important. An analogy with diagnostic testing can help make it clearer.

Using your understanding of diagnostic tests to understand P-values

Introduction to bayesian thinking: false-positive confusion

Remember the specious argument from Chapter 3, when we addressed what we called “false-positive” and “false-negative” confusion? Box 3.3 (about the need always to do a urine culture after a negative urinalysis) and Problem 3.3 (about the need always to culture the throat after a negative rapid test) were about false negatives. Recall that the faulty logic went something like this:

1. The sensitivity is 90%.
2. Therefore, the false negative rate is 10%.
3. Therefore, if the test is negative, there is a 10% chance that it is a false negative.

But, in fact, statement 3 was false, because in statement 2, “false negative” refers to $(1 - \text{Sensitivity})$, and in statement 3 it refers to $(1 - \text{Negative Predictive Value})$.

For this chapter, it is false-positive confusion that is most relevant. In the diagnostic testing setting, the false-positive confusion goes something like this:

1. The specificity of a test is 95%.
2. Therefore, the false-positive rate is 5%.
3. Therefore, if a patient has a positive result, there’s a 5% chance that it is a false-positive and the patient does not have the disease.
4. Therefore, if a patient has a positive result, there is a 95% chance that he does have the disease.

Once again, the problem is with statement 3 in which the probability of a positive result given no disease was converted into the probability of no disease given a positive result. That is, in the standard 2×2 table (Fig. 11.1), the usage of the term “false-positive rate” in statements 1 and 2 was $b/(b + d) = 1 - \text{Specificity}$. This corresponds to going vertically in the 2×2 table.

Then, in statement 3, we switched and started going horizontally, and the “false-positive rate” changed to $b/(a + b) = 1 - \text{Positive Predictive Value}$ (Fig. 11.2).

The “false-positive rate” that goes horizontally ($1 - \text{Positive Predictive Value}$) is more clinically relevant once you get a positive result; it is the probability that

		Gold Standard		Total
		Disease	No Disease	
Test	Positive	a True Positive	b False Positive	a + b
	Negative	c False Negative	d True Negative	c + d
		a + c	b + d	

Figure 11.1 When “false positive rate” refers to $1 - \text{specificity}$ or $b/(b + d)$, we are looking at the vertical “No Disease” column in the standard 2×2 table for a diagnostic test.

		Gold Standard		
		Disease	No Disease	Total
Test	Positive	a True Positive	b False Positive	a + b
	Negative	c False Negative	d True Negative	c + d
		a + c	b + d	

Figure 11.2 When “false positive rate” refers to $1 - \text{Positive Predictive Value}$ or $b/(a + b)$, we are looking at the horizontal “Test Positive” row of the standard 2×2 table for a diagnostic test.

your patient still does not have the disease, despite that positive result. However, we learned that it cannot be calculated from just (sensitivity and) specificity, because it depends on the prior probability of the disease.

Now consider the following argument:

1. We set α , the probability of a Type 1 error, at 5%.
2. Therefore, the probability of falsely concluding there is a difference, when in fact none exists, is 5%.
3. Therefore, if the P-value for our study is less than 0.05 and we reject the null hypothesis, the chance that we will be wrong is 5%.
4. Therefore, if the P-value is less than 0.05, there is at least a 95% chance that the difference between groups is not due to chance.

Can you see that this is exactly the same fallacy? Once again, the problem is with statement 3, although the ambiguity of statement 2 contributed to the problem. Statement 3 confuses the probability of the results given the null hypothesis with the probability of the null hypothesis given the results.

The key is that the P-value is a conditional probability; it is calculated assuming that the null hypothesis is true. In this way, it is like $1 - \text{Specificity}$, which is calculated conditional on not having the disease. $1 - \text{Specificity}$ is the probability of testing positive if you do not have the disease, whereas the P-value is the probability of observing an effect (or, to be more precise, a value of the test statistic) at least as extreme as that observed in the study if the null hypothesis were true.¹

This analogy between diagnostic and statistical tests can be visualized with a 2×2 table, similar to the ones we used for diagnostic tests (Fig. 11.3).

Just as was the case with diagnostic tests, what you really want is to go horizontally in this table – that is, what you want to know is the probability that there truly is a difference between groups, given the study results. But when you calculate a P-value, you are going vertically. That is, you assume the null hypothesis is true in order to calculate the P-value.

¹ For any one research question, there are many possible null hypotheses, and hence many test statistics that can be calculated. For example, there are test statistics to compare means, ranks, and standard deviations between groups, and they will not always give the same P-value. Note that it is also possible to calculate distributions of test statistics and P-values under assumptions other than the null hypothesis. For example, in an equivalency study, one might want to test the hypothesis that drug A is inferior to drug B by a specified amount. This is like calculating test characteristics for disease A vs. disease B, as opposed to Disease A present and absent. For example, “specificity” could be how often the test is negative in people with disease B rather than in everyone who does not have Disease A.

		TRUTH	
		Difference	No Difference
Study	Positive	Power ($1 - \beta$)	α
	Negative	β	$1 - \alpha$

Figure 11.3 The analogy between diagnostic and statistical tests can be visualized with a 2×2 table, similar to the one we used for diagnostic tests. Power ($1 - \beta$) is analogous to sensitivity and α is analogous to $1 - \text{Specificity}$.

We can summarize the Bayesian understanding of P-values exactly as we did when discussing diagnostic tests:

What you thought before + New information = What you think now

The new information, in this case, is the result of the study. The P-value is a measure of how consistent the result of the study is with the null hypothesis. However, it is not the posterior probability of the null hypothesis, because you cannot obtain a posterior probability without a prior probability.

Extending the analogy

The analogy between diagnostic tests and research studies can provide a lot of help understanding other aspects of P-values, too. A full analogy, adapted from an article Warren Browner and TBN wrote in 1987 (Browner and Newman 1987) is shown in Table 11.1.

Table 11.1. The analogy between diagnostic tests and research studies

Diagnostic test	Research study
Absence of Disease	Truth of null hypothesis
Presence of disease	Null hypothesis is false (e.g., real difference between groups exists)
Severity of disease in the diseased group	Magnitude of the true difference between groups
Cutoff for distinguishing positive and negative results	Alpha
Test result	P-value
Negative result (test within normal limits)	P-value exceeds alpha
Positive result	P-value less than alpha
Sensitivity	Power
False positive rate ($1 - \text{specificity}$)	Alpha
Prior probability of disease (of a given severity)	Prior probability of a difference between groups (of a given magnitude)
Posterior probability of disease, given test result	Posterior probability of a difference between groups, given study results

We can think of a research study as a diagnostic test to detect a difference (or association) between groups. Just as a sensitive test is more likely to find disease when it is present, a study with plenty of power (i.e., large sample size) is more likely to find a difference when it is present. In Chapter 5, we learned that many diseases are not homogenous, and that sensitivity would be expected to increase with the severity of disease. The analogy for research studies is that large differences between groups (i.e., strong associations) are easier to identify than small ones. Just as sensitivity depends on the severity of disease you wish to detect, power depends on the magnitude of the difference between groups you wish to detect; bigger differences, like more severe disease, are easier to find.²

When one does formal hypothesis testing for a research study, one compares the P-value from a study with a previously defined cut-off (α) for determining whether to reject the null hypothesis. This is analogous to deciding whether a test result falls within the “Normal Range.” Note that the more sure you want to be that a test is abnormal before labeling it as such, the wider your normal range will be. Similarly, the more sure you want to be that a P-value is inconsistent with the null hypothesis, the lower the alpha you will require.

Of course, simply comparing a P-value to alpha and reporting that it is lower (e.g., “ $P < 0.05$ ”) discards information. A P-value of 0.001 provides stronger evidence against the null hypothesis than a P-value of 0.049. This is similar to the point we made in Chapter 4, that dichotomizing WBC counts at 15,000 throws away information; a WBC count of 28,000 provides stronger evidence of bacteremia than a WBC count of 16,000.

Intentionally ordered tests and hypotheses stated in advance

If after a history and physical examination, you suspect a particular disease, and order a diagnostic test to confirm your hypothesis, a positive result is quite believable. This is because the disease you were testing for had a high prior probability. The posterior probability of disease depends only on the prior probability and the test result, however, not on whether you were smart enough to entertain the diagnosis in advance. Thus, the fact that a test was ordered by the third-year medical student with no particular suspicion of the disease does not mean the attending physician needs to assign a low prior probability when interpreting the result, if the history and physical examination immediately suggested the correct diagnosis to the attending.

Similarly, when testing research hypotheses, it is generally true that hypotheses stated in advance have higher prior probabilities than hypothesis arrived at after examining the data. But whether or not a hypothesis was stated in advance is not what is important. All that matters is the prior probability of the hypothesis being tested. Thus, if, after the data have been collected, some other study suggests a particular hypothesis, that hypothesis can be tested and will have a reasonable prior

² The analogy is not perfect, because for truly dichotomous disease states we need not specify a severity or stage of disease when estimating sensitivity, whereas we always must specify the magnitude of the difference we wish to detect when estimating power. This is because the degree of departure from the null hypothesis is not dichotomous.

probability, even if it was not stated in advance of the data collection. This happens in clinical medicine as well. A finding that the clinician either initially did not pay much attention to or dismissed as a red herring can suddenly provide evidence in favor of a disease when other findings pointing to that previously unconsidered disease become available.

Multiple hypotheses and multiple tests

It is well known that, if you look for enough different associations, either by selecting from multiple predictor and outcome variables or by restricting attention to various subgroups, it is easy to find statistically significant associations. The usual explanation for this is that, if there is a 5% chance of making a Type 1 error testing a single hypothesis, then if you test two (independent) hypotheses, the chance of such an error with either one would be closer to 10%; and if you test enough hypotheses, your chances of finding one or more with $P < 0.05$ approaches one. To address this issue, the Bonferroni correction is sometimes applied. The Bonferroni correction says that, if you want to test k different hypotheses and maintain a particular value for α , the Type 1 error rate for your whole study, you should use α/k as the Type 1 error rate for each individual hypothesis tested. Thus, if you wanted to test 2 hypotheses, you would require $P < 0.025$ before rejecting the null hypothesis; for 5 hypotheses, you would require $P < 0.01$, and so on.

The Bonferroni correction is overly conservative, because it does not account for the possibility that more than one of the null hypotheses can be falsely rejected.³ There are less conservative methods (such as the Holm, Student–Newman–Keuls or the Tukey tests) for adjusting the Type I error rate of each individual comparison when you are doing multiple comparisons (Glantz 2002). However, any adjustment to α for individual comparisons based on the overall α is problematic to apply. If you have collected your data and start running analyses, do you have to start counting every P-value your statistics package calculated as one of your hypotheses and reduce your value of α for individual comparisons accordingly? Must your level of α be forever affected? What if your colleague is running analyses as well? Do her hypothesis tests count against your α ? If the drug you are studying is associated with a bothersome side effect (e.g., cardiac arrhythmias), can you render the result not statistically significant by testing enough additional hypotheses about other side effects?

The problem with multiple hypothesis testing is that most of the multiple hypotheses have low prior probabilities. This is similar to the difference between a test that is intentionally ordered and one that pops up as abnormal on a twenty-test chemistry

³ To understand this, you need to understand the following probability theorem:

$$P(A \text{ or } B) = P(A) + P(B) - P(A \& B)$$

It makes sense to subtract $P(A \& B)$ because otherwise the probability gets counted twice. (Try drawing a Venn diagram to see this.) With the Bonferroni correction, event A is rejection of null hypothesis A and event B is rejection of null hypothesis B. $P(A) = P(B) = \alpha$, so $P(A \text{ or } B) = P(A) + P(B) - P(A \& B) = \alpha + \alpha - P(A \& B) = 2\alpha - P(A \& B)$. Of course, it is possible to falsely reject 2 different null hypotheses, so $P(A \& B) > 0$. Therefore, the probability of falsely rejecting either of the null hypotheses must be less than 2α .

panel. The interpretation of a particular statistical hypothesis test does not depend on how many other hypotheses were tested in the same study, just as the interpretation of a serum sodium level does not depend on whether you ordered an alkaline phosphatase on the same specimen. If clinical laboratories believed in the Bonferroni correction, they would widen the normal range of laboratory tests depending on how many tests were done on the same specimen. That being said, statistical approaches to avoid making too much of small P-values in the face of multiple comparisons are probably reasonable, because estimation of prior probabilities of hypotheses is a difficult and subjective process. But it is important to remember that, despite the aura of objectivity around statistical adjustment for multiple comparisons, no amount of statistical manipulation can ever get you a posterior probability without including some estimate of the prior probability. All of these methods basically aim to limit the probability of one or more Type 1 errors under the assumption that all of the null hypotheses are true.

Bias and laboratory error

Up to this point, when considering whether an observed difference between groups was real, we have been primarily concerned about chance as an alternative explanation for the study findings. However, another possible explanation could be bias: the results did not occur by chance, yet there may not be any real difference between groups. Bias in research studies is analogous to laboratory error. For example, in a patient reported to have an elevated potassium level, the specimen may have hemolyzed, and the true potassium level might be within the normal range. In fact, the patient could be hypokalemic; laboratory error can cause false-negative as well as false-positive results. Similarly, bias can hide real associations in addition to creating spurious ones.

Alternative diagnoses and confounding

An abnormal laboratory finding may not be due to a laboratory error, but still may not be due to the disease you are testing for; some other disease might be responsible. Similarly, a difference between groups (in an observational study) may not be due to chance, but still may not be due to a causal relationship; a third, confounding factor might be responsible. Thus, just as diagnosis requires a lot more than determining whether results of tests are normal, interpreting clinical research requires much more than ruling out chance and bias as the basis for associations.

Confidence intervals

There is no direct analogy between interpretation of results of diagnostic tests and of confidence intervals for research studies. Nonetheless, because confidence intervals are even more widely misunderstood than P-values, we review their meaning here.

The temptation is to say that there is a 95% probability that the true value of the parameter you are trying to estimate (e.g., the relative risk) lies within the 95% confidence interval (CI). As we have seen with P-values above, this is not the

case, because this is a statement of posterior probability, which you cannot make without knowing the prior probability.⁴ It turns out it is easier to say what confidence intervals do not mean than what they do mean. Confidence intervals do *not* indicate a range with a 95% probability of including the true value. What do they mean? We think a nonquantitative definition works best: *the confidence interval indicates a range of values consistent with what was observed in the study.*⁵ The higher the “level of confidence” (e.g., 99% vs. 95%), the wider the interval will be, corresponding to a looser definition of “consistent.” Of course, by chance alone, the true value might not be consistent with what was observed in the study, because the study happened to give the wrong answer.

The pedagogical debate we mentioned earlier, related to whether to try to give the correct interpretation of P-values, applies even more to confidence intervals, which are a bit harder to understand. For example, Douglas Altman, a bright light in the field of statistics and medicine, has written (Sackett et al. 2000; Guyatt et al. 2002):

A strictly correct definition of a 95% CI is, somewhat opaquely, that 95% of such intervals will contain the true population value. Little is lost by the less pure interpretation of the CI as the range of values within which we can be 95% sure that the population value lies.

We disagree. We think a lot is lost by the less pure interpretation, because different hypotheses have a wide range of prior probabilities. Therefore, the interpretation of the CI as the range of values within which we can be 95% sure that the population value lies is, in many cases, not even close.

Here is a simple example showing that the commonly used interpretation of the CI is wrong:

Picture a randomized trial, comparing Treatment A with Treatment B, that only has ten subjects per group. Four in each group die. The RR for mortality is 1.0 with a 95% CI of 0.34 to 2.9. You might believe that there is only a 5% chance that the true value is outside that CI, because it is fairly wide. But the 40% CI is a bit more narrow (0.75 to 1.33). Is there a 40% chance that it contains the true value? If so, there must be a 60% chance that the true value is outside that 40% CI – that is, that the true RR is <0.75 or >1.33 . In other words, there is a 60% chance that Treatment A either lowers mortality by 25% or increases it by 33%! But your study provided no information to suggest this was the case. How can a study that shows no difference between groups lead to a probability of 60% that there is at least a 25% difference in either direction?⁶

Let us just try to understand the strictly correct definition of the 95% CI. The idea that 95% of such intervals include the population value refers to a process, not to a particular interval. If we performed the study 100 times, we would expect that the 95% CIs of 95% of the studies would include the true value. But if we only did

⁴ Statisticians get around the fact that confidence intervals don’t mean what it seems like they should by creating their own definition of the word “confidence.” This definition makes the statement that you can be 95% *confident* that the true value lies within the 95% *confidence* interval both true and tautologous.

⁵ Another definition of the 95% confidence interval is *the range of hypotheses that would NOT have been rejected by this study at the 5% significance level ($\alpha = 0.05$)*. More generally, the $(1 - \alpha)$ confidence interval is the range of hypotheses that would not have been rejected at significance level α .

⁶ The answer is that the posterior probability that the true RR is <0.75 or >1.33 could be 60% only if the prior probability were more than that. Given no additional information about treatments A and B, there is no reason to presume that this is the case.

the study once, other information (i.e., a low prior probability of the result) might suggest that the 95% CI that we happened to get might not include the true value. In that case, the probability that the true value is within that study's 95% CI could be very different from 95%.

To understand this idea of a “process,” imagine that you have a bag filled with nineteen oranges and one grapefruit. The process of selecting one piece of fruit at random from the bag has a 95% chance of drawing an orange. If I tell you I have selected a piece of fruit, but give you no additional information, you will say that there is a 95% probability that I have an orange. However, if I tell you that it feels quite large for an orange, the probability that it is an orange decreases significantly. As with oranges versus grapefruits, in medicine, you usually have some prior information about the particular quantity you are trying to estimate. If the 95% CI that you obtained from your study seems as unlikely as a grapefruit-sized orange, then the probability that the interval contains the true parameter value is substantially less than 95%.

Again, we can summarize the Bayesian understanding confidence intervals similarly to that of P-values and diagnostic tests:

What you thought before + New information = What you think now

The “new information” in this case is the result of the study. The 95% CI is a range of parameter values consistent with the parameter estimate from the study, but it does not have a 95% probability of containing the true parameter value, because you cannot obtain posterior probability without prior probability.

Understanding and reporting negative studies: P-values, power, and confidence intervals

There is a trend toward eschewing P-values in favor of confidence intervals, the latter being felt to be more informative. Confidence intervals are, in fact, more informative; although, it isn't really a fair comparison, because confidence intervals have two numbers and the P-value is only one number. Confidence intervals are particularly useful for negative studies – they let you see how big an effect could have been missed.

Consider the reporting and interpretation of negative studies as a progression from the most elementary to the most sophisticated. We can present this the way Sackett et al. (1991) have presented interpretation of diagnostic tests, using a progression of colored karate belts.

We will use as an example a study of treatment of febrile infants with oral amoxicillin to prevent complications (like meningitis or infected joints or bones) of bacteremia (bacteria in the blood). The study included children 3–36 months old with fevers of at least 39°C (Jaffe et al. 1987). The authors reported that 27 of the 955 children in the study were bacteremic and that complications occurred in 2 of 19

(10.5%) bacteremic infants treated with amoxicillin compared with 1 of 8 (12.5%) bacteremic infants treated with placebo, a difference that was not statistically significant ($P = 0.9$). Note that there were more than twice as many bacteremic children in the amoxicillin group ($N = 19$) as in the placebo group ($N = 8$), presumably due to bad luck ($P = 0.07$), although a problem with the randomization is also possible.

White belt

The white belt just involves looking at the P-value to see whether it is ≥ 0.05 (or whatever alpha was chosen). Thus, a white belt reader would look at the study above and conclude, “amoxicillin doesn’t work,” because the P-value is far from significant. Many doctors and investigators have a white belt.

Yellow belt

The yellow belt involves considering not only the P-value, but also the power of the study. (Recall that the power is $1 - \beta$, the probability that the null hypothesis will be rejected, given that a true difference of a specified magnitude exists.) The power of a study is often included with a sample size calculation in the “Methods” section of a paper. In fact, some reviewers and editors insist on this, although in fact (as discussed below), it is not of much use to readers. The basic idea is that a negative study is not convincing if it was underpowered.

The study cited above was, in fact, underpowered. The authors state in the discussion that the power to detect a fourfold difference between groups in the odds of complications was only 24%. The authors’ conclusion that their “data do not support routine use of standard doses of amoxicillin . . .” is certainly reasonable, but that conclusion would also be true if they had studied 2 rather than 955 patients.

Green belt

The green belt is to examine the 95% CI for the RR or OR. In this case, the authors did present a 95% CI for the OR for complications.⁷ The point estimate of the OR was 1.2 with a 95% CI of 0.02 to 30.4. (This is actually the ratio of the odds of complications in the placebo group to the odds of complications in the amoxicillin group; they did not follow the convention of putting the odds in the active treatment group on top.) This tells you explicitly the range of values consistent with the study. One of us (TBN) was surprised that a negative study published in the *New England Journal of Medicine* would have such a wide confidence interval for its major outcome, and (with Dr. Robert Pantell) wrote a letter to the editor (Newman and Pantell 1988). The letter pointed out that a confidence interval for the OR that ranges from 0.02 to 30.4 suggests that the study provided virtually no information on the research question. True enough, but not the whole answer. Too bad we didn’t have a brown belt! Read on.

⁷ Why they presented the OR, and not the RR is not clear, as this was a randomized trial. It is especially puzzling because the 95% CI for the RR is quite a bit narrower! They also presented the risk difference (12.5% – 10.5% = 2%) and its confidence interval (–15% to +32%).

Blue belt

The blue belt does not apply to all studies, but does in this example. The key is to make sure that you do an intention-to-treat analysis. The analysis done by the authors compared complications *only among bacteremic patients!* But, as discussed in Chapter 9, the analysis should include all subjects randomized. At the time the amoxicillin was given, there was no way to know which children were bacteremic and which were not. Thus, benefits, risks, and costs might occur in nonbacteremic patients, and need to be compared between the entire amoxicillin group and the entire placebo group. The correct RR (keeping, for comparison purposes, the placebo group on top) is the ratio of 1/448 (the risk of complications in the placebo group) to 2/507 (the risk of complications in the amoxicillin group), which equals 0.57 with a 95% CI of 0.05 to 6.2.⁸

Brown belt

The confidence interval for the RR calculated in the “Blue belt” section is fine, but for making clinical decisions, it is really the absolute risk reduction (ARR), not the RR, that determines the balance of risks and benefits and hence clinical decisions. The brown belt involves calculating the (correct) ARR and its 95% CI. The ARR in this case was -0.17% . (Because the point estimate was an increase in risk with amoxicillin, the risk reduction is negative.) The 95% CI for the ARR goes from -0.9% to $+0.5\%$. That is, the upper limit of the 95% CI for the benefit of amoxicillin in this study is an absolute reduction in risk of complications of 0.5% . This, in turn, means that the lowest number needed to treat consistent with this study would be $1/0.5\% = 200$. If we are pretty sure that a NNT of 200 is too high, then the study makes us confident that we should not routinely treat febrile infants with amoxicillin.

If we are trying to use the study results to help with a clinical decision about a treatment, the ARR and its confidence interval are most useful. However, the relative risk reduction (RRR) tends to be more generalizable than the ARR. Thus, for patients at higher risk of bacteremia and/or complications, the NNT could easily be lower and whether they might benefit from treatment remains unknown.

Looking at the 95% CI for the ARR is a good idea for positive studies as well. The whole P-value and hypothesis-testing system is designed to determine the consistency of the data with an effect size of zero. But ruling out an effect size of zero is not as useful as ruling out an effect size that would be too small to warrant treatment. Thus, we could be fairly certain that a treatment has some small beneficial effect but still uncertain about whether to prescribe it. If a 95% CI not only excludes no effect, but also excludes benefits that are clinically trivial (i.e., that would lead to an NNT that is much too high), the study provides much stronger evidence of a clinically meaningful effect.

⁸ Note that the direction of the effect, albeit totally explicable by chance, is now in favor of placebo; the placebo group had a lower risk of complications than the amoxicillin group.

KEY POINT: The most important thing to look for when a study of a possible treatment shows no difference between groups is the confidence interval for the ARR, to see whether a clinically significant benefit (or risk) is consistent with the study results.

For a positive study, we want to look at the 95% CI for the ARR to see whether a clinically insignificant effect is consistent with the results.

Black belt

This one doesn't exist yet. We're saving it just in case we get more insights!

Once the study is completed, why are confidence intervals so much better than power?

When investigators set out to design a study, one of the things they need to do is estimate the required sample size. This generally involves making some guesses or assumptions about how the variables they wish to measure are distributed and the magnitude of the difference they hope to detect. Sometimes these assumptions turn out to be wrong. For example, perhaps when the investigators of the study of amoxicillin described above did their sample size calculation, they estimated that they would have many more bacteremic patients than they found. Once the study is done, however, it really doesn't matter what the investigators *thought* they were going to find when they designed the study. What they actually did find is what determines how informative the study is. This is much better expressed using a 95% CI, which incorporates evidence generated by the study. The power calculation, done before the results of the study were available, was based on less information.

For a negative study, the confidence interval is particularly useful when there is a trend in the data. For example, a little known result from the International Reflux Study in Children, a study of medical vs. surgical treatment of vesicoureteral reflux, (reflux of urine from the bladder up the ureter to the kidney; Weiss et al. 1992) is the “within-groups” relationship between pyelonephritis (kidney infections) and renal scarring. Standard teaching is that scarring in children with reflux is due to infection – that is, to reflux of infected urine into the kidney. But in that study, children with one or more episodes of acute pyelonephritis were *less* likely to develop renal scarring: RR = 0.28; 95% CI = 0.07 to 1.14. The fact that the point estimate is far below one and that the upper limit of the 95% CI only goes to 1.14 is much more informative than just a statement that the results were not statistically significant. This study may not have had power to detect a doubling of the risk of scarring with infection, but because of how the results came out, it suggests that an effect of that magnitude is highly unlikely.⁹

⁹ If you're really paying attention, you might argue that the prior probability that the relative risk would be less than 1 was very small, and we'd have to agree. That's why we said we can be pretty sure the risk is not *doubled*, rather than just saying we can be pretty sure it's within the 95% CI. You might also ask why we present the RR and its confidence interval, rather than the ARR. The reason is that the RR is better for assessing causality (the goal here), while the ARR is better for clinical decision making.

Useful shortcut: confidence intervals for small numerators

A situation that arises frequently in clinical research is that you observe either no instances of the outcome (called “events” in probability lingo) or a very small number of them. Years ago, Hanley and Lippman–Hand (1983) wrote a classic paper about zero numerators, called “If Nothing Goes Wrong, Is Everything All Right?”. They described the “Rule of Three,” which states that, if you observe zero events out of N trials (e.g., no deaths in $N = 100$ people on a drug), then the upper limit of the 95% CI for the true event rate is about $3/N$ (Box 11.1).

Example: A new drug is given to 60 people. It seems to work, and has no serious adverse effects. The authors conclude it is “safe and effective.” The upper limit for the 95% CI for any serious adverse effect is about $3/60$, or 5%.

Box 11.1: Derivation of the Rule of Three

If p is the probability of an event, x is the number of events, and N is the number of trials, your goal is to find the value of p such that $P(x = 0) = 0.05$.

$$P(x = 0) = (1 - p)^N = 0.05$$

$$N \ln(1 - p) = \ln(0.05)$$

MacLaurin Series Expansion:

$$\ln(1 - p) = -p - p^2/2 - p^3/3 - p^4/4 - \dots \approx -p, \text{ for small } p$$

$$N(-p) \approx \ln(0.05)$$

$$Np \approx \ln(20)$$

$$Np \approx 2.996$$

$$p \approx 3/N$$

(The “3” in the “Rule of 3” comes from the natural logarithm of 20, which is 3, or equivalently, from the natural logarithm of $1/20$, which is -3 .)

The “Rule of Three” for 0 numerators has analogs for slightly higher numerators, too (Newman 1995). Basically, for numerators of 0, 1, 2, 3, and 4, the numerator for the upper limit of the 95% CI is somewhere around 3, 5, 7, 9, and 10, respectively (Table 11.2). These numbers are not exact, but they are close enough, and a whole lot easier to do in your head or on your calculator than exact confidence intervals.

It is easier to illustrate this shortcut with examples than to explain it.

1. Three deaths are observed in 500 patients on a new drug. What is the upper limit of the 95% CI for the death rate?

The short cut for 3 is to use 9 as the numerator for the upper limit of the 95% CI. So it would be $\sim 9/500$, or 1.8%. (The exact binomial answer is 1.74%.)

Table 11.2

Observed numerator	Approximate numerator for upper limit of 95% CI
0	3
1	5
2	7
3	9
4	10

2. One case of HIV is found among 101 household contacts. What is the upper limit of the 95% CI for the risk of HIV among contacts?

For a numerator of 1, you use 5. So the upper limit of the 95% CI is $\sim 5/101 = 5\%$. (The exact binomial answer is 5.4%.)

3. A laboratory test done on 50 patients with disease is found to be 98% sensitive. What is the lower limit of the 95% CI for sensitivity?
- First you need to figure out that there must have been $49/50 (= 0.98 \times 50)$ positive tests.
 - Therefore, the false-negative rate was $1/50$.
 - The upper limit of the 95% CI for false-negative rate of $1/50$ is about $5/50$, or 10%.
 - Therefore, lower limit of 95% CI for sensitivity is $100\% - 10\% = 90\%$. (Exact binomial answer is 89.4%.)

Summary of key points

- P-values are sometimes misinterpreted as the probability that the null hypothesis (e.g., of no difference between groups) is true. But because P-values are calculated conditional on the null hypothesis, they cannot provide the probability that it is true.
- Confidence intervals provide a range of values consistent with results of the study, but it is not true that a 95% CI from a study has a 95% probability of containing the true value of the parameter being studied.
- 95% CI for negative studies are more useful than power, because they include information obtained from the study results
- In negative studies, look at the confidence interval for the absolute risk reduction (ARR), to see whether a clinically significant benefit (or risk) is consistent with the study results.
- The 95% CIs of the ARR for positive studies are most convincing when they not only exclude a null effect, but also exclude effects too small to be clinically meaningful.
- The “Rule of 3” for 0 numerators can be used to estimate the upper limit of the 95% CI for studies with no events. The rule can be extended to a “Rule of 3, 5, 7, 9, and 10” for numerators of 0, 1, 2, 3, and 4.

References

- Browner, W. S., and T. B. Newman (1987). "Are all significant P values created equal? The analogy between diagnostic tests and clinical research." *JAMA* **257**(18): 2459–63.
- Glantz, S. A. (2002). *Primer of Biostatistics*. New York, NY, McGraw-Hill, Medical Pub. Div.
- Goodman, S. N. (1999a). "Toward evidence-based medical statistics. 1: The P value fallacy." *Ann Intern Med* **130**(12): 995–1004.
- Goodman, S. N. (1999b). "Toward evidence-based medical statistics. 2: The Bayes factor." *Ann Intern Med* **130**(12): 1005–13.
- Goodman, S. N. (2001). "Of P-values and Bayes: a modest proposal." *Epidemiology* **12**(3): 295–7.
- Guyatt, G., D. Rennie, et al. (2002). *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice*. Chicago, IL, AMA Press.
- Hanley, J. A., and A. Lippman-Hand (1983). "If nothing goes wrong, is everything all right? Interpreting zero numerators." *JAMA* **249**(13): 1743–5.
- Jaffe, D. M., R. R. Tanz, et al. (1987). "Antibiotic administration to treat possible occult bacteremia in febrile children." *N Engl J Med* **317**(19): 1175–80.
- Newman, T. B. (1995). "If almost nothing goes wrong, is almost everything all right? Interpreting small numerators." *JAMA* **274**(13): 1013.
- Newman, T. B., and R. H. Pantell (1988). "Occult bacteremia in febrile children." *N Engl J Med* **318**(20): 1338–9.
- Sackett, D. L., R. B. Haynes, et al. (1991). *Clinical Epidemiology: A Basic Science for Clinical Medicine*. Boston, MA, Little Brown.
- Sackett, D. L., R. B. Haynes, et al. (2000). *Evidence-Based Medicine: How to practice and teach EBM*, 2nd Ed. Edinburgh: Churchill Livingstone: 233.
- Weiss, R., J. Duckett, et al. (1992). "Results of a randomized clinical trial of medical versus surgical management of infants and children with grades III and IV primary vesicoureteral reflux (United States). The International Reflux Study in Children." *J Urol* **148**(5 Pt 2): 1667–73.

Chapter 11 Problems

1. Evaluate each of the following statements about statistical inference as true or false. Briefly explain your answer.
 - a) If the P-value = 0.05, then there is a 95% probability that the results did not occur by chance.
 - b) The null hypothesis generally states that there is a difference between the groups.
 - c) If the P-value is sufficiently high, the null hypothesis is not rejected.
 - d) The 95% CI is the range of values with a 95% probability of containing the true (population) value.
2. You may have been told or observed that, when a test is ordered as part of a chemistry panel of 20 tests, an abnormal result is more likely to be false positive than when the test is ordered by itself. Is this correct? Why?
3. A randomized trial from Italy (Veronesi et al. 2003) compared sentinel-node biopsy (just removing one underarm lymph node to see if it has cancer in it) with routine axillary dissection (opening up the armpit and trying to remove all of the nodes) in 516 women with primary breast cancer tumors ≤ 2 cm in diameter.

They found significantly less swelling, pain, scarring, and numbness or tingling in the women in the sentinel-node group. There also were fewer unfavorable events and deaths in that group, as shown in the table below:

	Axillary Dissection	Sentinel-Node Biopsy
Number of subjects	257	259
Adverse events other than death (metastases, recurrences, etc.)	21	13
Deaths	6	2

The authors' conclusion was: "Sentinel-node biopsy is a safe and accurate method of screening the axillary nodes for metastasis in women with a small breast cancer."

- a) The point estimate for the death rate in the sentinel-node biopsy group was $2/259 = 0.77\%$. Use the shortcut in the chapter to estimate the upper limit of the 95 % CI for this estimate.

An accompanying editorial, however, was critical of the Italian study because of its small sample size (Krag and Ashikaga 2003). It cited two other trials in process as having appropriate sample sizes: one with power to detect about a 2% (absolute) difference in survival and the other with power to detect a 5% difference. As the editorialists put it:

The era in which randomized clinical trials are dominated by a single institution – an approach that was perhaps justifiable in the past – is now over, since virtually no single institution can enroll enough patients to allow detection of small differences between two study groups . . .

The conclusion that sentinel-node surgery does not result in reduced survival and therefore that it is a safe procedure, equivalent to axillary dissection, must await the completion of larger clinical trials with sufficient power.

Assume that, as suggested by the editorialists, a $\geq 2\%$ absolute difference in total mortality would be clinically significant. Partial output from Stata (csi command) to compare total mortality in the two groups is shown below. (The sentinel-node group is considered "exposed" and "cases" are deaths.)

csi 2 6 257 251

	Exposed	Unexposed	Total	
Cases	2	6	8	
Noncases	257	251	508	
Total	259	257	516	
Risk	.007722	.0233463	.0155039	
	Point estimate		[95% CI]	
Risk difference	−0.0156243		−0.0369425	.0056939
Risk ratio	0.3307593		0.0673847	1.623539
	$\chi^2(1) = 2.06$		$\text{Pr} > \chi^2 = 0.1509$	

- b) Based on the 95% CI, is a clinically significant ($\geq 2\%$) increase in mortality with sentinel-node biopsy consistent with the findings?
 - c) Imagine that you went through your answer to part (b) with the editorialists, and they remained skeptical. How would you explain their continued skepticism in Bayesian terms?
4. A case-control study (Foxman and Frerichs 1985) of urinary tract infections (UTIs) among female college students found that, among women who had sexual intercourse less than once a week, diaphragm use (compared with oral contraceptive use) was associated with increased odds of UTI (OR = 7.0; 95% CI 0.04 to 625).
- a) Based on this result, what can you say about the posterior probability of the hypothesis that diaphragm use causes UTI in these women?
 - b) Why do you think the confidence interval is so wide?
5. A Glaxo-Smith-Kline-funded study compared the antidepressants paroxetine and imipramine with placebo in a randomized, double-blind study in adolescents with major depression (Keller et al. 2001). The “Results” section of that paper states:

Serious adverse effects occurred in 11 [of 93] patients in the paroxetine group, 5 [of 95] in the imipramine group, and 2 [of 87] in the placebo group . . . The serious adverse effects in the paroxetine group consisted of headache during discontinuation taper (1 patient) and various psychiatric events (10 patients) . . . Of the 11 patients, only headache (1 patient) was considered by the treating investigator to be related to paroxetine.

The “Discussion” states:

Because these serious adverse events were judged by the investigators to be related to treatment in only 4 patients (Paroxetine, 1; imipramine, 2; placebo, 1), causality cannot be determined conclusively.

The last sentence of the abstract is:

CONCLUSIONS: Paroxetine is generally well tolerated and effective for major depression in adolescents.

Although no P-values for adverse events are presented in the paper, we calculated the P-value for serious adverse events, comparing paroxetine with placebo, and it is 0.014 (two-tailed).

- a) The calculation above entirely ignores the fact that there was an imipramine group. If that group is included, the investigators would want to make three comparisons: paroxetine vs. imipramine, paroxetine vs. placebo, and imipramine vs. placebo. Using the Bonferroni correction for testing these three hypotheses at $\alpha = 0.05$, a P-value of $0.05/3 = 0.0133$ would be required to reject the null hypothesis, and results above would not be statistically significant. Do you think the Bonferroni correction is appropriate in this case? Why or why not?

- b) The authors indicate that the treating physicians generally did not attribute adverse effects in the paroxetine group to paroxetine, and for this reason “causality cannot be determined conclusively.” Do you agree? Explain.
6. We have talked about the analogy between diagnostic tests to identify a disease and clinical research studies to identify a causal relationship between predictor and outcome. What is the expected effect of the Bonferroni correction used to adjust for multiple hypothesis testing on the *sensitivity* and *specificity* of a research study for identifying a causal relationship?

References for problem set

- Foxman, B., and R. R. Frerichs (1985). “Epidemiology of urinary tract infection: I. Diaphragm use and sexual intercourse.” *Am J Public Health* **75**(11): 1308–13.
- Keller, M. B., N. D. Ryan, et al. (2001). “Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial.” *J Am Acad Child Adolesc Psychiatry* **40**(7): 762–72.
- Krag, D., and T. Ashikaga (2003). “The design of trials comparing sentinel-node surgery and axillary resection.” *N Engl J Med* **349**(6): 603–5.
- Veronesi, U., G. Paganelli, et al. (2003). “A randomized comparison of sentinel-node biopsy with routine axillary dissection in breast cancer.” *N Engl J Med* **349**(6): 546–53.