CPH 931 — Fall 2008 — Dr. Charnigo

Lecture 10

Lecture 10A: Avoiding common misunderstandings about p-values

Preface. Parts of Lecture 10A are based on pages 329 through 338 of the *Users' Guides to the Medical Literature* (2005) edited by Guyatt and Rennie. Other parts are based on "A Dirty Dozen: Twelve P-Value Misconceptions" by Goodman (published in *Seminars in Hematology*, 2008). Sometimes I refer to "A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure" by Cohn et al (published in the *New England Journal of Medicine*, 1991).

Introduction. We are all familiar with p-values. A p-value less than 0.05 is statistically significant and permits rejection of a null hypothesis. If a p-value less than 0.05 is acquired when comparing two groups on the primary endpoint of a study, then the study is positive. Conversely, a p-value greater than 0.05 is not statistically significant, does not allow rejection of a null hypothesis, and yields a negative study if acquired when comparing two groups on the primary endpoint.

Despite being ubiquitous in the medical literature, p-values remain a source of misunderstandings. Three of the most common misunderstandings are as follows.

1. The p-value is the probability that the null hypothesis is true.

2. When comparing two groups, a p-value greater than 0.05 means that there is no important difference between the groups.

3. When comparing two groups, a p-value less than 0.05 means that there is an important difference between the groups.

After reviewing some hypothesis testing concepts, I will explore these three misunderstandings in detail.

Null and alternative hypotheses. The null and alternative hypotheses represent two competing theories about the state of nature. For example, the null hypothesis in Cohn et al (1991) is that enalapril and hydralazine-isosorbide dinitrate are equally effective in preventing short-term mortality among males receiving digoxin and diuretic therapy for heart failure, while the alternative hypothesis is that enalapril and hydralazine-isosorbide dinitrate are not equally effective.

Generally speaking, the null hypothesis represents a neutral stance that we suspect is false but are unwilling to abandon unless the evidence against it is convincing. On the other hand, the alternative hypothesis represents a non-neutral stance that we suspect is true but are unwilling to adopt unless the evidence against the null hypothesis is compelling.

Note that there is an asymmetry here: a decision is made according to whether there is enough evidence against the null hypothesis rather than whether there is enough evidence favoring it. This is not unlike how criminal trials are conducted in the United States: a decision is made according to whether there is enough evidence against the defendant (innocent until proven guilty) rather than whether there is enough evidence favoring him (guilty until proven innocent).

What is a p-value? A p-value is the probability of getting results at least as extreme as those actually observed if the null hypothesis is true.

As a simple illustration, consider a fictional scenario in which 10 people are given two medications ("Drug 1" and "Drug 2") in random order and then asked to state which medication they prefer. The null hypothesis is that the medications are equally effective, while the alternative hypothesis is that they are not.

Table 1:

Prefer Drug 1	Probability	p-value
10	1/1024	2/1024
9	10/1024	22/1024
8	45/1024	112/1024
7	120/1024	352/1024
6	210/1024	772/1024
5	252/1024	1024/1024
4	210/1024	772/1024
3	120/1024	352/1024
2	45/1024	112/1024
1	10/1024	22/1024
0	1/1024	2/1024

Assuming a forced choice (i.e., ruling out answers like "I have no preference") and that the null hypothesis is true, the most likely outcome is that 5 people will prefer Drug 1 and that 5 people will prefer Drug 2. The probability that this will happen is $252/1024 = 0.246$. On the other hand, that all 10 people will prefer Drug 1 is an extremely unlikely outcome. The probability that this will happen is $1/1024$.

If all 10 people prefer Drug 1, then the p-value is $2/1024 = 0.002$. This is the probability that the number of people who prefer Drug 1 will be either 10 or 0. Note that 0 is as extreme a result as 10: if 0 people prefer Drug 1, then 10 people prefer Drug 2.

If 8 people prefer Drug 1, then the p-value is $112/1024 = 0.109$. With the small sample size of 10, having 80% of the people vote for Drug 1 is not strong enough evidence against the null hypothesis for us to conclude that Drug 1 is more effective!

Exploring the first misunderstanding: "The p-value is the probability that the null hypothesis is true." Central to this misunderstanding is the attempt to assign probabilities to the state of nature. To understand this, consider a simple question: "What is the probability that the first customer who entered Rite Aid yesterday was female?"

One's initial reaction may be to answer "50%". However, the correct answer is "Either 100% or 0%, but I can't say which because I wasn't there." That answer may sound strange, but think about it. If we reviewed the security tapes, then we would know the gender of the first customer. Is the gender of the first customer not determined just because we haven't seen the security tapes? That is an absurd proposition because "yesterday" already happened. Yet, the incorrect answer of "50%" implies that the gender of the first customer is not determined as long as we haven't seen the security tapes.

The bottom line is that <u>our ignorance of the state of nature does not make it random</u>. The probability that the null hypothesis is true is either 100% or 0%, but we can't say which. We can make an educated guess about the state of nature by collecting data and quantifying the evidence against the null hypothesis.

Exploring the second misunderstanding: "When comparing two groups, a p-value greater than 0.05 *means that there is no important difference between the groups."* Behind this misunderstanding is the failure to distinguish between "absence of evidence" and "evidence of absence". Consider the actual data from the study by Cohn et al (1991).

During follow-up 132 of 403 patients (32.8%) assigned to enalapril died, compared to 153 of 401 patients (38.2%) assigned to hydralazine and isosorbide dinitrate. The estimated risk difference of −5.4% is accompanied by

Table 2:

	Death	No Death	Row Total
Enalapril	132	271	403
Hydralazine	153	248	401
Column Total	285	519	804

a p-value of 0.109. Based on the data in Table 2, we cannot conclude that enalapril is more effective than hydralazine-isosorbide dinitrate.

Now suppose that we had the following data.

Table 3:

	Death	No Death	Row Total
Enalapril	264	542	806
Hydralazine	306	496	802
Column Total	570	1038	1608

All I did was double the number in each cell, so the estimated risk difference remains -5.4%. But now the p-value is 0.023, and we can conclude that enalapril is more effective than hydralazine-isosorbide dinitrate.

The next question is whether the estimated 5.4% risk reduction in Table 3 is important. If we answer yes, then how can we assert that there is no important difference in Table 2 when the estimated risk reduction is identical?

The issue is not that the 5.4% is unimportant in Table 2. <u>The issue is that, with the sample size in Table 2, we are not sure whether the 5.4% is "real"</u>. More precisely: with the sample size in Table 2, the probability under the null hypothesis of obtaining an estimated risk difference of magnitude at least 5.4% — i.e., 0.109 — is large enough that observing an estimated 5.4% risk reduction does not permit us to reject the null hypothesis.

Yet, if real, the 5.4% is important and its failure to trigger rejection of the null hypothesis in Table 2 is what we call a Type II error (failing to reject a false null hypothesis).

Reducing the probability of a Type II error — or, equivalently, increasing the power — is why investigators exercise so much care in determining the sample sizes for their studies. In particular, an investigator usually chooses the sample size so that the Type II error probability on the primary endpoint is 20% — i.e., the power is 80% — if the difference between the two groups is some specified amount that the investigator considers important (e.g., a risk reduction of 3%).

Exploring the third misunderstanding: "When comparing two groups, a p-value less than 0.05 *means that there is an important difference between the groups."* Central to this misunderstanding is the failure to separate "statistical significance" from "clinical significance". Consider the following fictional data.

Table 4:

	Death	No Death	Row Total
Enalapril	264	542	806
Hydralazine	264	540	804
Column Total	528	1082	1610

The death rate in the enalapril group was 32.754%, while the death rate in the hydralazine-isosorbide dinitrate group was 32.836%. The estimated risk difference of −0.082% is accompanied by a p-value of 0.972. Based on the data in Table 4, we cannot conclude that enalapril is more effective than hydralazine-isosorbide dinitrate.

Now suppose that we had the following data.

Table 5:

	Death	No Death	Row Total
Enalapril	2640000	5420000	8060000
Hydralazine	2640000	5400000	8040000
Column Total	5280000	10820000	16100000

All I did was multiply the number in each cell by 10000, so the estimated risk difference remains −0.082%. But now the p-value is less than 0.001, and we can conclude that enalapril is more effective than hydralazine-isosorbide dinitrate.

Is the 0.082% estimated risk reduction in Table 4 important? If we answer no, then how can we assert that there is an important difference in Table 5 when the estimated risk reduction is identical?

Well, perhaps there isn't an important difference in Table 5. The small p-value tells us not to believe that the risk reduction is 0. But the small p-value does not promise us that the risk reduction is large enough to be important. Why not?

Conceptually, the alternative hypothesis allows for infinitely many possibilities: the risk reduction may be 8.2%, 0.82%, 0.082%, 0.0082%, or any number between. A p-value only measures the implausibility of the null hypothesis, namely that the risk reduction is 0. A p-value does not measure the implausibility of any possibility contained within the alternative hypothesis.

In other words, <u>we cannot tell just by looking at a p-value whether the state of nature is far away from or close to the null hypothesis</u>.

Far away from or close to the truth? Let's face the facts: we're not really interested in whether the null hypothesis is true, because in the real world the null hypothesis for a two-sided hypothesis test is always false — if only by a small amount. Rather, we're interested in whether the state of nature is far away from the null hypothesis, in which case one treatment has a large enough advantage that our clinical practice should change, or whether the state of nature is close to the null hypothesis, in which case neither treatment has a large advantage.

Whenever we conduct a hypothesis test, we attempt to use statistical significance as a proxy for clinical significance. Sometimes this does not lead us astray. But we have seen in Table 5 that a <u>clinically insignificant result can be statistically significant with an immense sample size</u>. On the other hand, Table 2 showed that a <u>clinically significant result can be statistically insignificant with a small sample size</u>.

So now what? Given the limitations of p-values, we need: (i) a way to assess whether a statistically insignificant result is also clinically insignificant; and, (ii) a way to assess whether a statistically significant result is also clinically significant.

We can rephrase (i) by saying that we need a way to see whether a negative trial is definitively negative, and we can rephrase (ii) by saying that we need a way to see whether a positive trial is definitively positive.

These ideas will be pursued in Lecture 10B.

Lecture 10B: Confidence intervals and clinical trials

Preface. Lecture 10B roughly follows pages 339 through 349 of the *Users' Guides to the Medical Literature* (2005) edited by Guyatt and Rennie. Sometimes I refer to *New England Journal of Medicine* papers published in 1991, one by Cohn et al titled "A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure" and one by the SOLVD investigators titled "Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure".

Potential problems with hypothesis testing. Suppose we are testing the null hypothesis that $p_1 = p_2$, where p_1 and p_2 are the risks of some unfavorable event under a new treatment and under an existing treatment (or placebo).

In the Cohn et al paper, p_1 can be the risk of death during follow-up among male heart failure patients on digoxin and diuretic therapy who receive enalapril, while p_2 can be the risk among such patients who receive hydralazine and nitrates.

In the paper by the SOLVD investigators, p_1 can be the risk of death or hospitalization for worsening heart failure during follow-up among heart failure patients on conventional treatment who receive enalapril, while p_2 can be the risk among such patients who receive placebo.

Two potential problems with hypothesis testing are as follows:

1. Suppose we accept the null hypothesis that $p_1 = p_2$. This means that we cannot rule out the possibility of equal risks. However, we have not ruled out the possibility of grossly unequal risks. Perhaps the risk difference $p_1 - p_2$ is 0%, but perhaps the risk difference is -5%. (A negative risk difference just means that the risk is lower with the first treatment.) In the former case, we may not prefer the first treatment;

perhaps the first treatment is more expensive or has debilitating side effects. In the latter case, there seems to be very good reason to prefer the first treatment. Acceptance of the null hypothesis doesn't tell us which case we're in, so without further information the implications for clinical practice are unclear.

2. Suppose we reject the null hypothesis that $p_1 = p_2$. This means that we have ruled out the possibility of equal risks. However, we have not ruled out the possibility of almost equal risks. Perhaps the risk difference $p_1 - p_2$ is -5%, but perhaps the risk difference is -1%. In the former case, there seems to be very good reason to prefer the first treatment. In the latter case, we may not prefer the first treatment. Rejection of the null hypothesis doesn't tell us which case we're in, so without further information the implications for clinical practice are unclear.

Point and interval estimates. Besides testing the null hypothesis that $p_1 = p_2$, we can estimate the risk difference $p_1 - p_2$. There are two kinds of estimates, point estimates and interval estimates.

A point estimate is the best single number guess that we can make for $p_1 - p_2$ based on the data. If we apply the principle that the best guess for $p_1 - p_2$ should be the best guess for p_1 minus the best guess for p_2, then a point estimate for $p_1 - p_2$ can be calculated very easily.

In the Cohn et al paper, the point estimate is

$$\frac{132}{403} - \frac{153}{401} = 32.75\% - 38.15\% = -5.40\%.$$

In the paper by the SOLVD investigators, the point estimate is

$$\frac{613}{1285} - \frac{736}{1284} = 47.70\% - 57.32\% = -9.62\%.$$

An interval estimate — commonly called a "confidence interval" — is a range of reasonable guesses for $p_1 - p_2$ based on the data. A 95% confidence interval for $p_1 - p_2$ has the form

$$\text{point estimate } \pm 1.96 \times \text{ standard error.}$$

In the Cohn et al paper, the 95% confidence interval is

$$-5.40\% \pm 1.96 \times 3.37\% = -12.01\% \text{ to } 1.21\%.$$

In the paper by the SOLVD investigators, the 95% confidence interval is

$$-9.62\% \pm 1.96 \times 1.96\% = -13.46\% \text{ to } -5.78\%.$$

Standard error and sample size. You may recall the mathematical formula for the standard error,

$$\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2},$$

where $\hat{p}_1, \hat{p}_2$ denote the point estimates of p_1, p_2 and n_1, n_2 denote the numbers of subjects in the two treatment groups.

An important feature of the standard error is that, to a close approximation, the square of the standard error is inversely proportional to the total sample size (i.e., $n_1 + n_2$). Thus, the standard error will be halved if you quadruple the total sample size.

So, if there had been 3200 patients in the Cohn et al study rather than 804 patients, the standard error would have been 1.7%.

Confidence interval length and sample size. Since the length of a confidence interval is directly proportional to the standard error, we can halve the length of a confidence interval by quadrupling the sample size. Thus, we can make a confidence interval as narrow as we like by taking the sample size large enough.

This is closely related to the idea of "powering a study". In fact, powering a study is really just choosing the sample size large enough so that the confidence interval will be narrow enough to exclude 0%. (Excluding 0% from the confidence interval is logically the same as rejecting the null hypothesis.)

Interpreting confidence intervals. Unfortunately, confidence intervals are often misinterpreted; even the authors of the *Users' Guides* succumb to misinterpretation!

A correct interpretation is that, if we could repeat the clinical trial an unlimited number of times and construct a new 95% confidence interval each time, then 95% of the clinical trials would yield confidence intervals that contained $p_1 - p_2$. (Keep in mind that $p_1 - p_2$ is unknown and can only be estimated, because a clinical trial enrolls a sample rather than a population of patients, notwithstanding the regrettable but ubiquitous references to "study populations" in published manuscripts.)

Stating that a given confidence interval contains $p_1 - p_2$ with 95% probability is nonsensical, because neither $p_1 - p_2$ nor the confidence interval is random. This is closely related to the error of interpreting a p-value as the probability that the null hypothesis is true. In both instances, there is an underlying confusion between what is truly random and what is simply unknown.

Likewise, stating that some values in a confidence interval are more likely than others is nonsensical. However, some values in a confidence interval are more consistent with the observed data than others. Thus, to the extent that we define plausibility as consistency with the observed data, some values in a confidence interval are more plausible than others.

Clinical trial results: when is a study definitively negative? Recall problem 1 on page 9 of Lecture 10B. Suppose we have accepted the null hypothesis that $p_1 = p_2$, which is to say we have not been able to rule out that $p_1 - p_2 = 0\%$. On the other hand, so far we have not been able to rule out that $p_1 - p_2 = -5\%$. Here is where the confidence interval comes into play. Let us consider two examples.

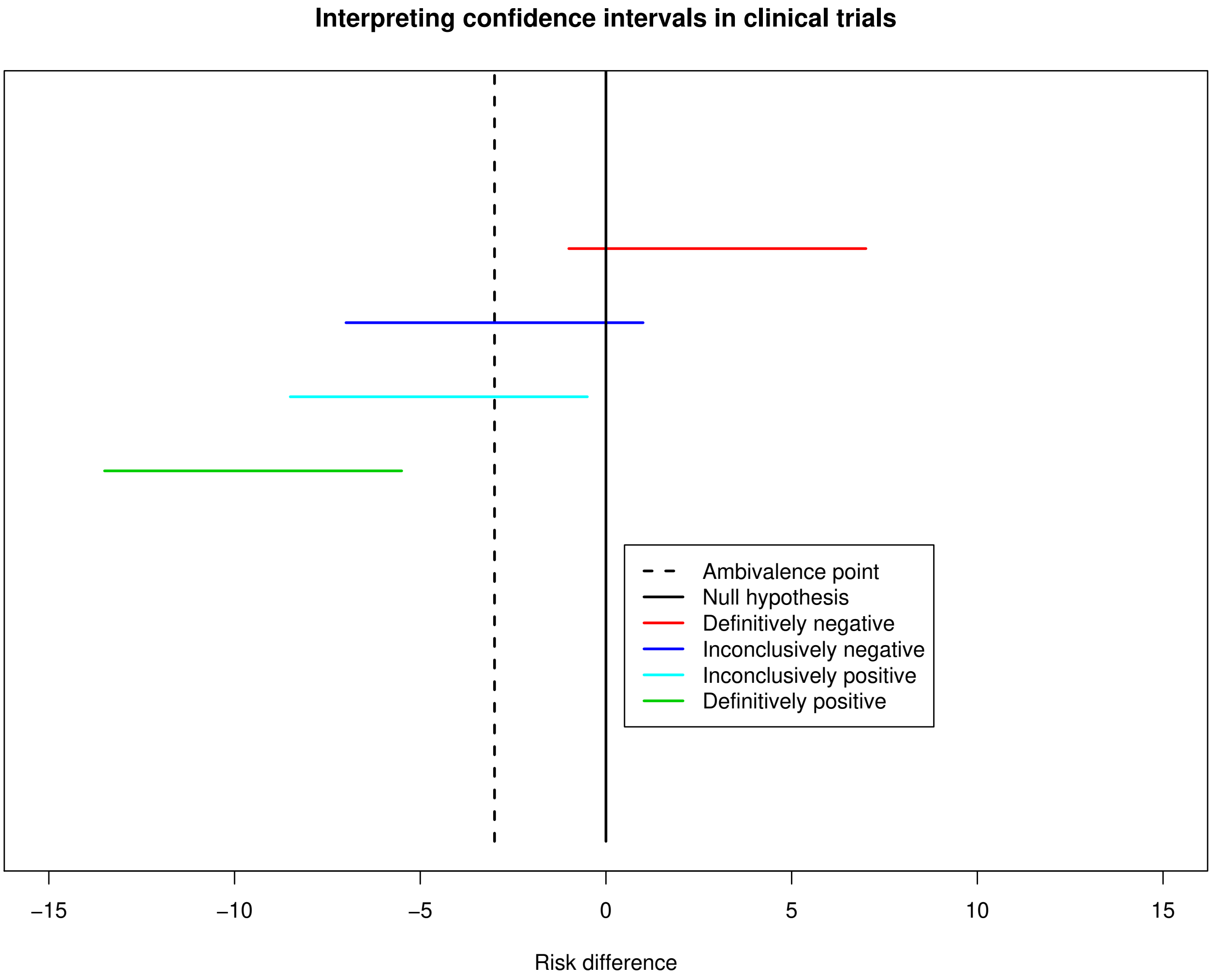
Example 1. Suppose that the confidence interval runs from -1% to $+7\%$. Then, since the plausible values for $p_1 - p_2$ are contained in the confidence interval, -5% is not a plausible value and can be ruled out. Assuming that we are ambivalent between the two treatments if $p_1 - p_2 = -3\%$, the study is definitively negative; all plausible values for the risk difference favor the second treatment.

Example 2. Suppose that the confidence interval runs from -7% to $+1\%$. Then -5% is a plausible value and cannot be ruled out. Assuming that we are ambivalent between the two treatments if $p_1 - p_2 = -3\%$, the study is inconclusive; some plausible values for the risk difference favor the second treatment, but some favor the first treatment.

The two examples suggest the following general strategy for determining whether an apparently negative study is definitively negative or inconclusive:

Look at the left endpoint of the confidence interval. If it falls to the right of the ambivalence point, then the study is definitively negative. Otherwise, the study is inconclusive.

The red and blue line segments in the figure on the next page illustrate this strategy.



Addressing problem 1 in the Cohn et al paper. With respect to death during follow-up, we cannot reject the null hypothesis that $p_1 = p_2$ using the data from the Cohn et al study. (We know this because 0% is contained in the confidence interval.)

So, is the Cohn et al study definitively negative, or is it inconclusive? The left endpoint of the confidence interval is -12.01%, which is to the left of any reasonable ambivalence point, so the study is inconclusive — at least with respect to death during follow-up.

Clinical trial results: when is a study definitively positive? Recall problem 2 on page 10 of Lecture 10B. Suppose we have rejected the null hypothesis that $p_1 = p_2$, which is to say we have been able to rule out that $p_1 - p_2 = 0\%$. On the other hand, so far we have not been able to rule out that $p_1 - p_2 = -1\%$. Again, the confidence interval will come to the rescue. Let us consider two more examples.

Example 3. Suppose that the confidence interval runs from -13.5% to -5.5%. Then, since the plausible values for $p_1 - p_2$ are contained in the confidence interval, -1% is implausible and can be ruled out. Assuming that we are ambivalent between the two treatments if $p_1 - p_2 = -3\%$, the study is definitively positive; all plausible values for the risk difference favor the first treatment.

Example 4. Suppose that the confidence interval runs from -8.5% to -0.5%. Then -1% is a plausible value and cannot be ruled out. Assuming that we are ambivalent between the two treatments if $p_1 - p_2 = -3\%$, the study is inconclusive; some plausible values for the risk difference favor the first treatment, but some favor the second treatment.

The two examples suggest the following general strategy for determining whether an apparently positive study is definitively positive or inconclusive:

Look at the right endpoint of the confidence interval. If it falls to the left of the ambivalence point, then the study is definitively positive. Otherwise, the study is inconclusive.

The green and turquoise line segments in the figure on the preceding page illustrate this strategy.

Addressing problem 2 in the SOLVD investigators paper. With respect to death or hospitalization for worsening heart failure during follow-up, we can reject the null hypothesis that $p_1 = p_2$ using the data from the study by the SOLVD investigators. (We know this because 0% is not contained in the confidence interval.)

So, is the study by the SOLVD investigators definitively positive, or is it inconclusive? The right endpoint of the confidence interval is -5.78%, which is to the left of any reasonable ambivalence point, so the study is definitively positive.

Discussion Questions

1. What would happen if the ambivalence point were 0%?
2. Explain whether you agree or disagree with the following statement: Most researchers who do not explicitly identify an ambivalence point have implictly taken 0% as the ambivalence point.
3. How can the above ideas be extended if we are interested in, for example, a hazard ratio rather than a risk difference?