Hypothesis testing

Hem/Onc Journal Club

September 13th, 2010

Didactic example: Objective response

- “Secondary end points included objective response rate, toxic effects, overall survival, and quality of life.” (p. 2668)
- “Both the objective response rate and toxic effects were compared with the use of Fisher’s exact test.” (p. 2670)
- “All P values are two-sided ...” (p. 2670)
- “The addition of bevacizumab to paclitaxel significantly improved the objective response rate in all eligible patients (36.9% vs. 21.2%, $P<0.001$) ...” (page 2670)

The hypothesis testing paradigm

1. Assume a null condition for population parameter(s).
2. Formulate a statement in terms of the population parameter(s) that reflects an effect.
3. Collect data
4. Organize the evidence in favor of the null condition. This evidence takes the form of a statistical test that is usually constructed from the estimate of the population parameter based on the data.
5. Under the assumption of no effect, construct a probabilistic statement (called a P value) based on the evidence in the data.
6. Reject the null or fail to reject the null, depending on the probability statement.

What is a population parameter?

Statisticians differentiate between that which is being estimated and the estimate itself. That which is being estimated is called a population parameter and we (typically) assume it has a true but unknown value. In this example, there are two population parameters.

- True (*but unknown*) ORR for the population of patients represented by those in the study receiving paclitaxel plus bevacizumab - p_{PB}
- True (*but unknown*) ORR for the population of patients represented by those in the study receiving paclitaxel alone - p_P

1. Assume a null condition

Statistical testing starts with a statement of the *null hypothesis* - i.e. the hypothesis of no effect. It is a statement about the population parameter(s).

The notation for the null hypothesis is H_0 (read “H naught”). For the ORRs, the null hypothesis is:

$$H_0 : p_{PB} = p_P.$$

This can also be written

$$H_0 : p_{PB} - p_P = 0.$$

2. Formulate a statement of effect

We next state the *a*lternative hypothesis. This is the state of nature we will accept if the evidence in the data suggests that the null is implausible.

The notation for the alternative hypothesis is H_A (read “H A”). For the ORRs, let’s assume the alternative hypothesis is:

$$H_0 : p_{PB} > p_P.$$

This can also be written

$$H_0 : p_{PB} - p_P > 0.$$

Note that we don’t actually specify a value for $p_{PB} - p_P$ under the alternative. We simply state the direction of the effect - is the ORR for P + B greater than, less than or simply different from the ORR for P alone?

3. Collect data

- How to collect data?
- How much data to collect?
- Not trivial - should utilize resources of statistician

4. Construct test based on data

Intuitively, it makes sense that any 'test' for a comparison of ORRs should be constructed from their estimates, namely $\hat{p}_{PB}$ and $\hat{p}_P$. The circumflex (or hat -$\hat{}$) notation indicates the estimate of the population parameter based on sample data. From the results stated in the paper,

$$\hat{p}_{PB} = 36.9\% \text{ and } \hat{p}_P = 21.2\%,$$

or equivalently

$$\hat{p}_{PB} - \hat{p}_P = 15.7\%.$$

Are these estimates meaningfully different, or is the magnitude of their difference a chance occurrence - the luck of the draw?

4. Construct test based on data (cont.)

The test is constructed based on the data, i.e. based on $\hat{p}_{PB} - \hat{p}_P$. Our goal is to determine how unlikely it is to observe a difference in the estimated ORRs at least as large as the one we've observed in our data under the assumption that there is actually no difference in the ORRs. This will allow us to make a probabilistic statement about the likelihood of the observed difference in ORRs being attributable to chance and not to the intervention.

This requires understanding the sampling distribution of $\hat{p}_{PB} - \hat{p}_P$.

The sampling distribution

You can think of a sampling distribution as a histogram of all the possible values of $\hat{p}_{PB} - \hat{p}_P$ you could observe *if in fact the true ORRs,* p_{PB} *and* p_P*, are actually equal*. This is what we mean when we say that the test is constructed based on the null condition.

Example based on ORRs

$\hat{p}_{PB} =$ 36.9% and $\hat{p}_P =$ 21.2%. The null hypothesis states that $p_{PB} = p_P$. If the null hypothesis is true, then our best estimate of the true underlying ORR (for either arm) is a pooled estimate. That is,

$$\hat{p}_{PB} = 128/347 \doteq 0.369$$

and

$$\hat{p}_P = 69/326 \doteq 0.212.$$

Then

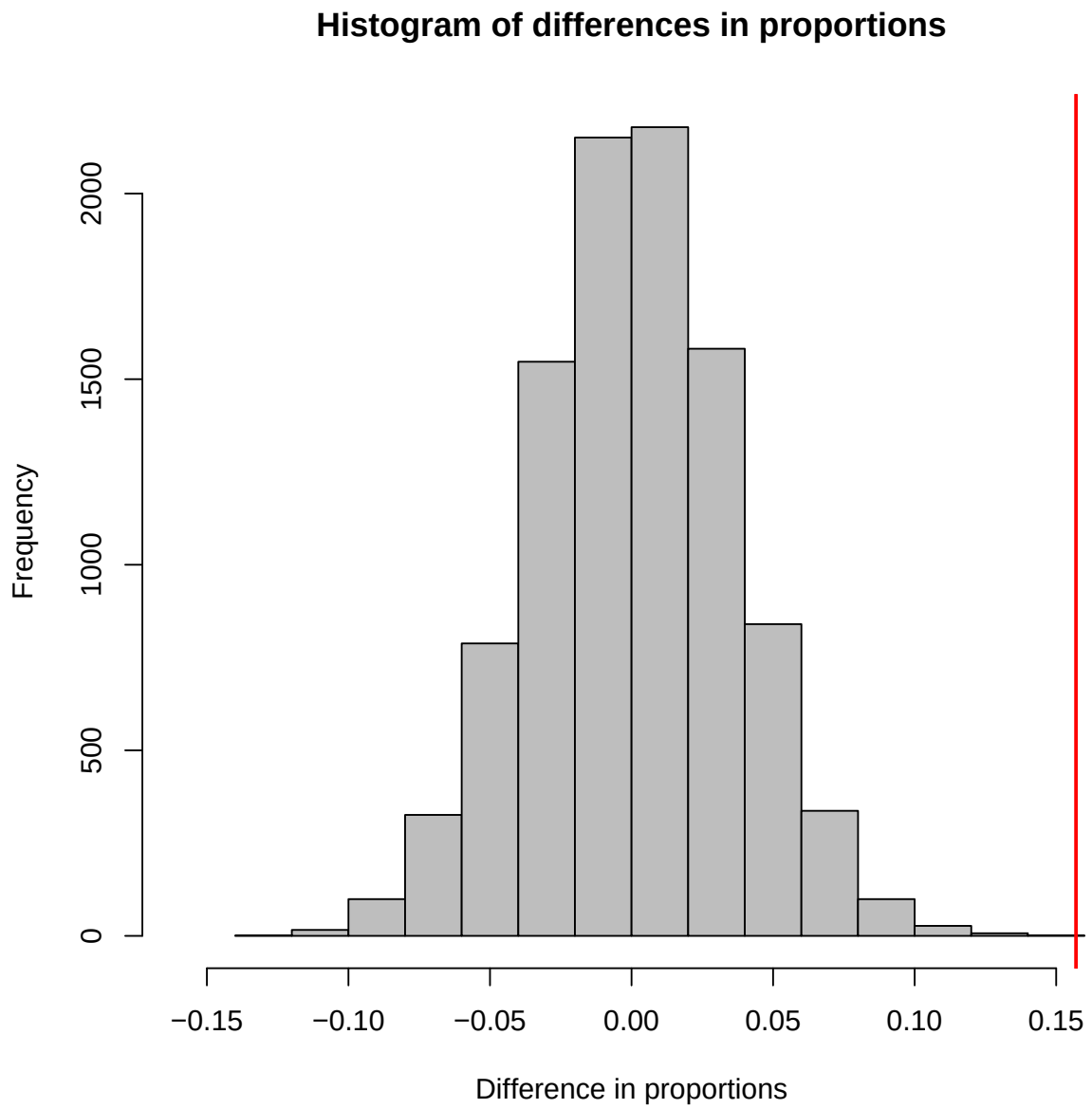
$$\hat{p}_{pooled} = (128 + 69)/(347 + 326) \doteq 0.293.$$

Repeated sampling under H_0

For the purposes of this simulation, assume $p_{PB} = p_P =$0.293.

Sample no.	$\hat{p}_{PB}$ (n = 347)	$\hat{p}_P$ (n = 326)	$\hat{p}_{PB} - \hat{p}_P$
1	0.291	0.35	-0.059
2	0.308	0.282	0.026
3	0.317	0.245	0.072
	⋮	⋮	⋮
9,999	0.349	0.291	0.057
10,000	0.288	0.31	-0.022

Histogram of resulting differences



What is a P value?

How would you use the histogram on the previous slide to estimate the probability of observing a difference in ORRs at least as extreme as the difference observed, assuming that there really is no difference?

What is a P value?

How would you use the histogram on the previous slide to estimate the probability of observing a difference in ORRs at least as extreme as the difference observed, assuming that there really is no difference?

Answer: The probability is approximately equal to the relative area under the histogram to the right of the observed difference. This probability is called a *P value*.

Other alternatives

If the alternative hypothesis had been

$$H_0 : p_{PB} < p_P$$

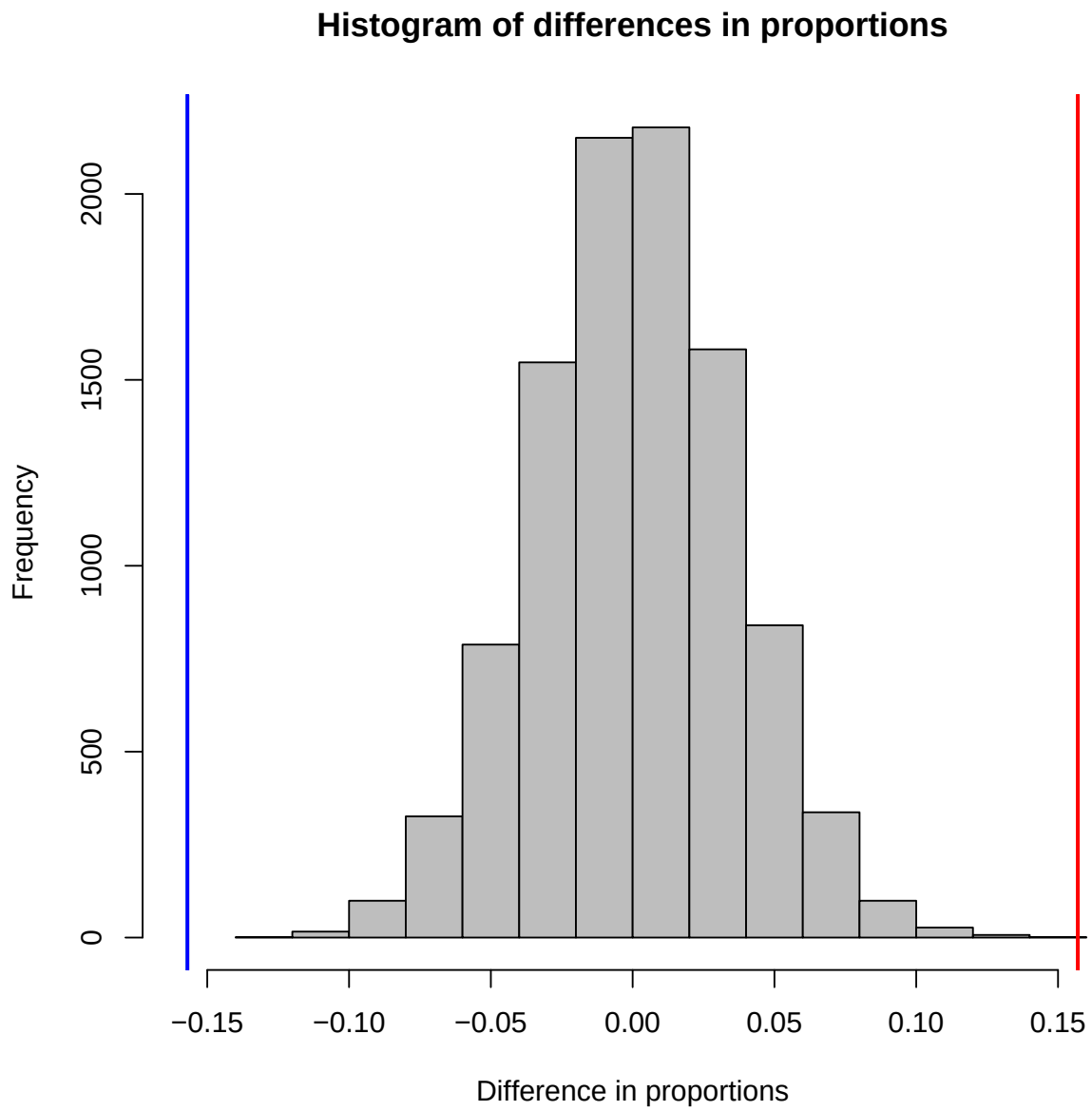
then the P value would be approximated by the relative area to the left of the observed difference.

If the alternative hypothesis had been

$$H_0 : p_{PB} \neq p_P$$

then the P value would be approximated by the sum of the relative areas to the left of the blue line and to the right of the red line (see next slide). The authors of the paper performed a two-sided test.

Histogram of resulting differences



Definition of P value

The P value is the probability of observing a result at least as extreme as the one you observed if the null is true. Another way of thinking about a p-value is this: the p-value is the probability of observing by chance alone a result at least as extreme as yours.

In this paper, the authors report that the test comparing the ORRs for the two treatment arms has a two-sided P value of $P < 0.001$.

5. Make a decision

Once you have your p-value, you make a judgement about the null hypothesis. When the p-value is very small (i.e. observing the outcome or one more extreme due to chance alone is highly improbable), we *reject the null hypothesis in favor of the alternative* and call the finding *statistically significant*. It is common practice to require a P value to be less than 0.05 before we declare a finding to be "... significant at level 0.05." When the p-value is not small (i.e. observing the outcome or one more extreme due to chance alone is probable), we *fail to reject the null*.

Notice that we don't say we *accept the null*. The only actions the statistical model allows us to take are reject or fail to reject the null hypothesis.

Actions and errors

Of course it is always possible to make a mistake. The table below details the two types of errors one can commit in conducting a statistical test.

	H_0 true	H_A true
Reject H_0	Type I error	No error
Fail to reject H_0	No error	Type II error

- A type I error is the probability of rejecting the null given that the null is true.
- We denote the probability of a type I error as α.
- A type II error is the probability of failing to reject the null given that the alternative is true.
- We denote the probability of a type II error as β.

Type I errors and the α-level of a test

When and how frequently does a type I error occur?

1. A type I error occurs when the null hypothesis is true, but you've had the bad luck of drawing an unusual sample so you reject the null when you really shouldn't.
2. We know that an unusual sample will be drawn with probability α (definition of probability of a type I error).
3. To reject the null, the p-value has to be smaller than some threshold, usually set at 0.05.
4. If the threshold is set at, say, 0.05, then we would expect to reject the null by mistake less than 5% of the time - that is, we expect to make a type I error less than 5% of the time.

Type I errors and the α-level of a test (cont.)

Therefore, *when you specify the type I error rate you're willing to allow in your testing scheme, you are also specifying the upper limit of the p-value for which statistical significance is declared.*

The α-level of a test is also called the *significance level* of the test.

Type II errors and power

Condition on the columns ...

	H_0 true	H_A true
Reject H_0	α	1 - β = Power
Fail to reject H_0	1 - α	β

The power of a test is the probability of correctly rejecting the null in favor of the alternative. A well-designed study will strike a balance between acceptable levels of type I error (usually 0.05) and power (often set at a minimum of 0.8).