Comparing groups and statistical decision-making

Hematology/Oncology Lecture Series

Elizabeth G. Hill, PhD

Associate Professor of Biostatistics

14 November 2011

Didactic example: Overall response

From Ohtsu et al., JCO 2011, page 4:

“Overall response rate was improved significantly with the addition of bevacizumab (46.0% v 37.4% in the placebo group; P = .0315.)”

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 bevacizumab - p_B
- True (*but unknown*) ORR for the population of patients represented by those in the study receiving placebo - 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:

$$H_0 : p_B = p_P.$$

This can also be written

$$H_0 : p_B - p_P = 0.$$

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:

$$H_A : p_B > p_P.$$

This can also be written

$$H_A : p_B - p_P > 0.$$

This is called a *one-tailed* or *one-sided test*. Note that we don’t actually specify a value for $p_B - p_P$ under the alternative. We simply state the direction of the effect - is the ORR for B greater than, less than, or simply different from the ORR for P?

3. Collect data

- How to collect data?
- How much data to collect?
- Not trivial - should utilize resources of statistician
- We'll talk about this a little more in next lecture

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 $\hat{p}_B$ 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,

$$\hat{p}_B = 46.0\% \text{ and } \hat{p}_P = 37.4\%,$$

or equivalently

$$\hat{p}_B - \hat{p}_P = 8.6\%.$$

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 $\hat{p}_B - \hat{p}_P$. 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 $\hat{p}_B - \hat{p}_P$.

The sampling distribution

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

Example based on ORRs

$\hat{p}_B =$ 46.0% and $\hat{p}_P =$ 37.4%. The null hypothesis states that $p_B = p_P$. If the null hypothesis is true, then our best estimate of the true underlying ORR (for either arm) is a pooled estimate. From Table 2, we obtain the frequencies of overall response in each arm. Then,

$$\hat{p}_B = 143/311 \doteq 0.460$$

and

$$\hat{p}_P = 111/297 \doteq 0.374.$$

Then

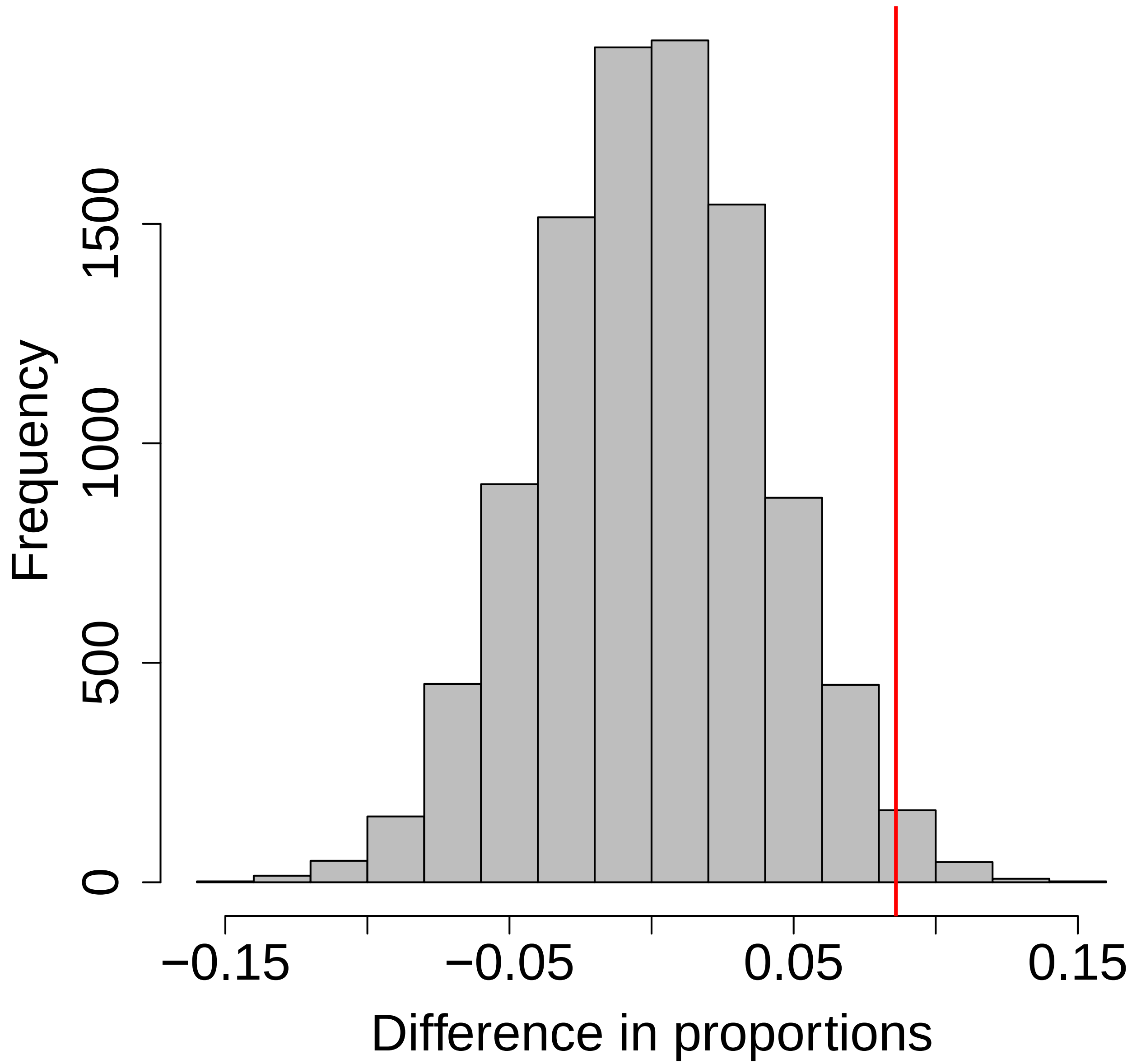
$$\hat{p}_{pooled} = (143 + 111)/(311 + 297) \doteq 0.418.$$

Repeated sampling under H_0

Simulate 10,000 data sets with sample sizes of 311 and 297, and assume $p_{PB} = p_P =$0.418. For each sample, compute $\hat{p}_B$ and $\hat{p}_P$ and construct their difference.

Sample no.	$\hat{p}_B$	$\hat{p}_P$	$\hat{p}_B - \hat{p}_P$
	(n = 311)	(n = 297)	
1	0.447	0.364	0.083
2	0.463	0.421	0.042
3	0.414	0.428	-0.014
$\vdots$	$\vdots$	$\vdots$	$\vdots$
9,999	0.415	0.431	-0.016
10,000	0.412	0.414	-0.002

Histogram of resulting differences

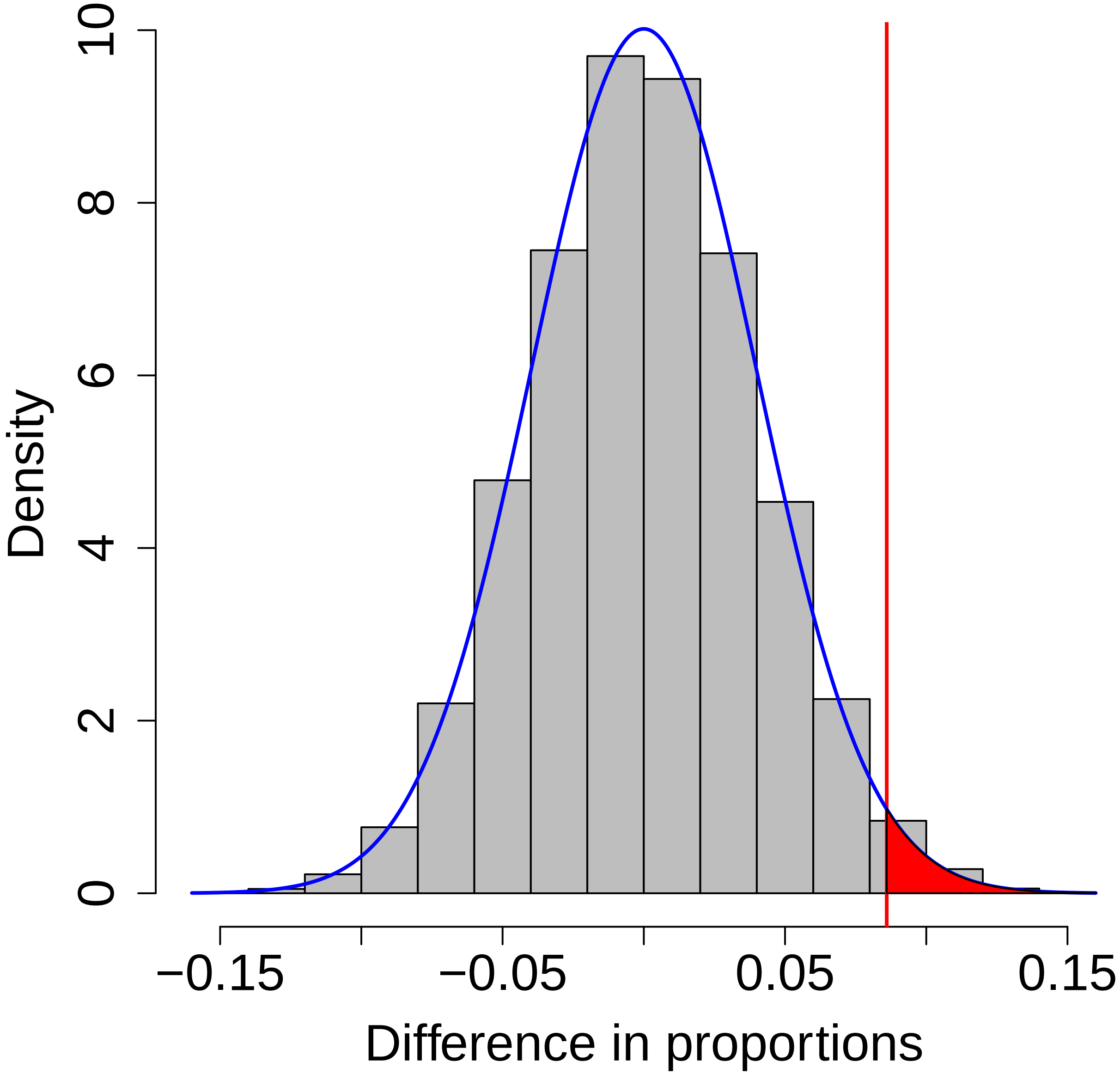


What is a P value?

Question: How can we decide how unusual the observed difference is? That is to say, if the histogram on slide 13 shows us what differences in proportions should look like when they are coming from the same underlying distribution, how do we quantify the ‘extremeness’ of the observed difference?

Answer: We estimate the probability of observing a difference in ORRs as or more extreme than the one we observed in our study. This probability is called a *P value*.

P value for our example



P value for our example (cont.)

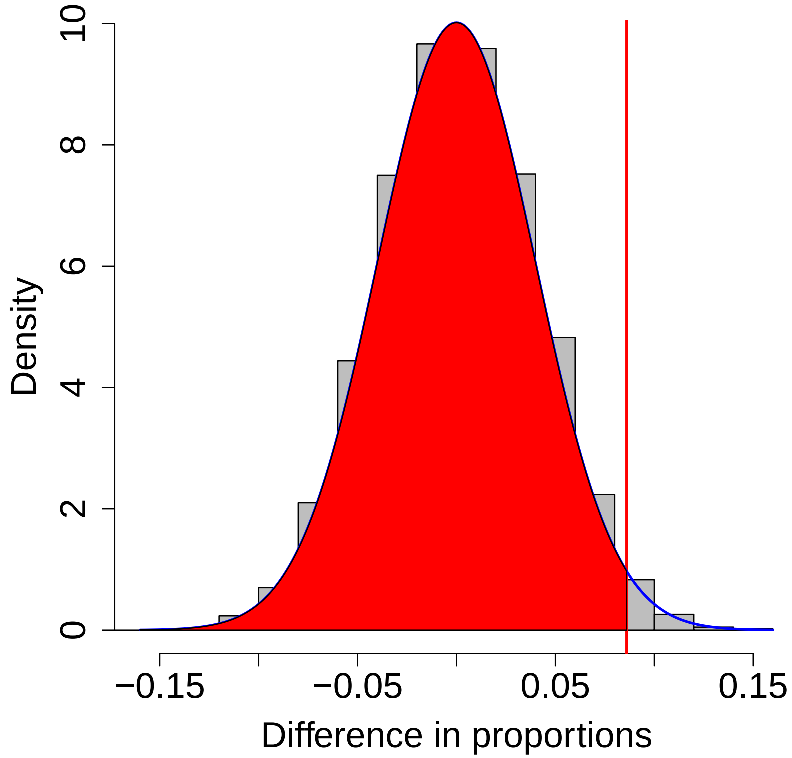
We approximate the theoretical distribution of the differences (blue line overlaying the histogram) using a normal distribution. The p-value is the red-shaded area in the plot. In this example, the p-value is approximately 0.016. (*This is different than what was reported in the paper. We'll explain why in a minute.*) This means that the likelihood of observing a difference of 8.6% (or one even larger) by chance alone is less than 2%.

Other alternatives

Another one-tailed test would be:

$$H_A : p_B < p_P$$

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

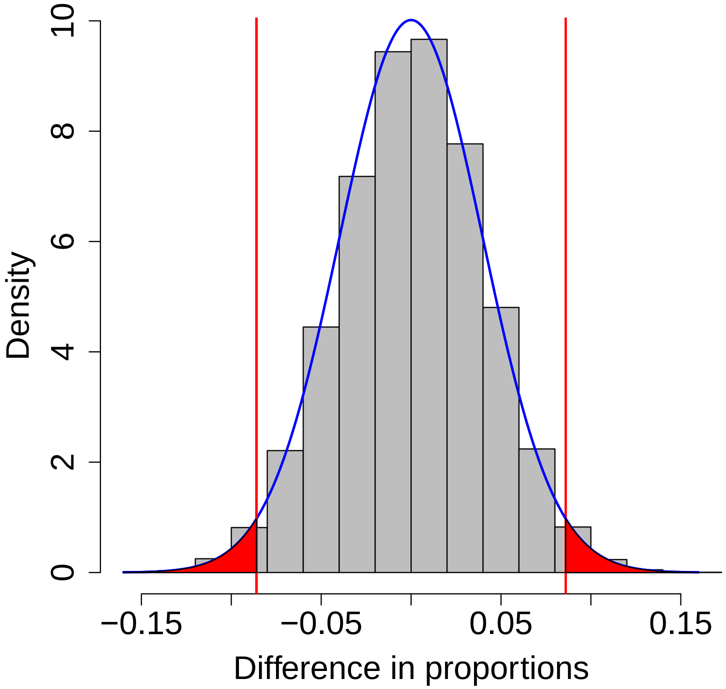


P value in the paper

A *two-tailed* or *two-sided test* is given by the alternative hypothesis

$$H_A : p_B \neq p_P$$

The p-value is approximated by the sum of the relative areas to the left of -0.086 and to the right of 0.086. The authors of the Ohtsu paper performed a two-sided test and report a p-value of 0.0315.



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.

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*. When we fail to reject the null hypothesis, we are unable to rule out random chance as a plausible explanation for any observed difference.

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 a sample resulting in extreme differences. You reject the null, but you really shouldn't.
2. If the null hypothesis is true, then 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.

Power

	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).

Common endpoints to compare groups

- Difference in means
  - Compares a continuous variable between two groups
  - Null value is 0
  - two-sample t-test or Wilcoxon rank sum test (independent groups)
  - paired t-test or Wilcoxon signed rank test (paired groups)
- Difference in proportions
  - Compares a categorical variable between two groups
  - Null value is 0
  - Chi-square test or Fisher’s exact test (independent groups)
  - McNemar’s test (paired groups)

Common endpoints to compare groups (cont.)

- Odds ratio (OR) = odds of being a “1” for subjects in group A/odds in of being a “1” for subjects in group B
    - Compares a binary variable between groups A and B (e.g. 1 = disease present, 0 = disease absent)
    - OR > 0
    - Null value is 1
    - E.g. OR = 1.3 means there is a 30% increase in the odds of being diseased comparing subjects in group A to those in group B
    - E.g. OR = 0.8 means there is a 20% reduction in the odds of being diseased comparing subjects in group A to those in group B
    - Chi-square test or logistic regression

Common endpoints to compare groups (cont.)

- Hazard ratio (HR) = hazard of death in group A/hazard of death in group B
  - ○ Compares a time-to-event variable between groups A and B
  - ○ HR > 0
  - ○ Null value is 1
  - ○ E.g. HR = 1.3 means there is a 30% increase in the risk of death comparing subjects in group A to those in group B
  - ○ E.g. HR = 0.8 means there is a 20% reduction in the risk of death comparing subjects in group A to those in group B
  - ○ Log rank test or Cox proportional hazards regression

Putting some of this together ...

From page 3 of Ohtsu et al., JCO 2011:

“*On the basis of a systematic literature review, it was assumed that median OS in the placebo group would be 10 months. The study was powered to test the hypothesis that the addition of bevacizumab would improve median OS to 12.8 months, equivalent to a hazard ratio (HR) of 0.78 between study groups, assuming an exponential distribution for the time-to-death variable. ... To detect an HR of 0.78, 509 deaths were necessary to ensure 80% power for a two-sided log-rank test at a significance level of 0.05.*”

Confidence intervals

A *confidence interval* is an estimated range of values that provides a measure of uncertainty associated with an estimated parameter. The wider the interval the greater the uncertainty in the estimate. The more narrow the interval the greater our faith in the estimate. Typically, a 95% CI is reported. The “95%”-part of the interval means that if you were able to replicate the exact same study an infinite number of times, 95% of the resulting CIs would contain the true parameter of interest. Of course, no one is ever going to actually repeat the study over and over again.

Here is an example from the Ohtsu paper, page 4.

“*PFS was prolonged significantly in the bevacizumab group compared with the placebo group (HR, 0.80; 95% CI, 0.67 to 0.93; P = .0037).*”

Confidence intervals (cont.)

How do we interpret this information? Our best estimate of the true HR is 0.80, but our data are consistent with a HR ranging from 0.67 to 0.93.

Notice that the CI does not contain 1.0, the null value for a HR, and the p-value of the test indicates significance. In general, let S be the statistic used to perform a test that compares two groups. For example, S could be a difference in sample means, a difference in sample proportions, an estimated OR, or an estimated HR. If the test is significant at α-level 0.05, then the 95% CI constructed with S will not contain the relevant null value.

Confidence intervals (cont.)

A final word about confidence intervals.

A CI is an estimated range of values that indicates the uncertainty in *any* estimated parameter, whether or not that parameter is used to compare two groups.

For example, in Table 2 of Ohtsu, the estimated 1-year survival rate for the bevacizumab group is 50.2% with a corresponding 95% CI of 45.1% to 55.3%. Based on our data, we estimate that slightly more than half of bevacizumab-treated subjects will survive to one year, but our data are consistent with one-year survival rates as low as 45% and as high as 55%.