



COLUMBIA
BUSINESS
SCHOOL

Hypothesis Testing

Fall 2001
B6014: Managerial Statistics

Professor Paul Glasserman
403 Uris Hall

General Ideas of Hypothesis Testing

1. *Hypothesis testing* refers to a general class of procedures for weighing the strength of statistical evidence — more specifically, for determining whether the evidence supporting one **hypothesis** over another is sufficiently strong.
2. A crucial feature of hypothesis testing is that the two competing hypotheses are not treated in the same way: one is given the benefit of the doubt, the other has the burden of proof. The one that gets the benefit of the doubt is called the **null hypothesis** and is denoted H_0 . The other is called the **alternative hypothesis** and is denoted H_1 (or H_a).
3. By definition, the default is the null H_0 . When we carry out a test, we are asking whether the available data is significant evidence in favor of the alternative H_1 . We are not testing whether H_1 is true; rather, we are testing whether the evidence supporting H_1 is **statistically significant**.
4. The conclusion of a hypothesis test is that we either **reject** the null hypothesis (and accept the alternative) or we **fail to reject** the null hypothesis. Failing to reject the null H_0 does not quite mean that the evidence supports H_0 ; rather, it means that the evidence does not strongly favor the alternative H_1 . Again, H_0 gets the benefit of the doubt.
5. Some preliminary examples:
 - Suppose we want to determine if stocks picked by experts generally perform better than stocks picked by darts. We might conduct a hypothesis test to determine if the available data should persuade us that the experts do better. In this case, we would have

H_0 : experts not better than darts

H_1 : experts better than darts

- Suppose we are skeptical about the effectiveness of a new product in promoting dense hair growth. We might conduct a test to determine if the data shows that the new product stimulates hair growth. This suggests

$$\begin{aligned}H_0 : & \quad \text{New product does not promote hair growth} \\H_1 : & \quad \text{New product does promote hair growth}\end{aligned}$$

Choosing the hypotheses this way puts the onus on the new product; unless there is strong evidence in favor of H_1 , we stick with H_0 .

- Suppose we are considering changing the packaging of a product in the hope of boosting sales. Switching to a new package is costly, so we will only undertake the switch if there is significant evidence that sales will increase. We might test-market the change in one or two cities and then evaluate the results using a hypothesis test. Since the burden of proof is on the new package, we should set the hypotheses as follows:

$$\begin{aligned}H_0 : & \quad \text{New package does not increase sales} \\H_1 : & \quad \text{New package does increase sales}\end{aligned}$$

6. Every hypothesis test concludes by rejecting or not rejecting H_0 . Hence, there are two ways of drawing incorrect conclusions in a hypothesis test. These go by the following names:

- Type I error: rejecting H_0 when H_0 is true.
- Type II error: sticking with H_0 when H_1 is true.

A Type I error is a false positive; a Type II error is a false negative.

7. Ideally, we would like the probability of both types of errors to be small. In practice, we can only make one smaller at the expense of making the other larger. (If we make it harder to reject H_0 , then we are less likely to commit a Type I error but more likely to commit a Type II error.)
8. Since we can't ensure that both error probabilities will be small, we design the test so that the Type I error probability is **at most** α , where α is a small number we pick. Typical values of α are .10, .05, .01.
9. α is called the **significance level** of the test. Its role here is very similar to that in confidence intervals.

Hypothesis Test on a Population Mean

1. We begin with the simplest case of a test. Suppose we are inclined to believe that some (unknown) population mean μ has the value μ_0 , where μ_0 is some (known) number. We have samples X_1, \dots, X_n from the underlying population and we want to test our hypothesis that $\mu = \mu_0$. Thus, we have

$$H_0 : \quad \mu = \mu_0$$

$$H_1 : \quad \mu \neq \mu_0$$

What sort of evidence would lead us to reject H_0 in favor of H_1 ? Naturally, a sample mean far from μ_0 would support H_1 while one close to μ_0 would not. Hypothesis testing makes this intuition precise.

2. This is a **two-sided** or **two-tailed** test because sample means that are very large **or** very small count as evidence against H_0 . In a **one-sided** test, only values in one direction are evidence against the null hypothesis. We treat that case later.
3. Example: Suppose we have reason to believe that the true average monthly return on stocks selected by darts is 1.5%. (See *Dart Investment Fund* in the casebook for background and data.) We want to choose between

$$H_0 : \quad \mu = 1.5$$

$$H_1 : \quad \mu \neq 1.5,$$

where μ is the true mean monthly return.

4. We need to select a significance level α . Let's pick $\alpha = .05$. This means that there is at most a 5% chance that we will mistakenly reject H_0 if in fact H_0 is true (Type I error). It says nothing about the chances that we will mistakenly stick with H_0 if in fact H_1 is true (Type II error).
5. **Large sample hypothesis test.** Let's suppose we have samples X_1, \dots, X_n with $n > 30$. The first step in choosing between our hypotheses is computing the following **test statistic**:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}.$$

I am temporarily assuming that we know σ .

6. Remember that we know μ_0 (it's part of the null hypothesis we've formulated), even though we don't know μ .
7. What counts as strong evidence in favor of H_1 ? A value of \bar{X} far from μ_0 , which is equivalent to a value of Z far from 0. Thus, we will reject the null if Z is sufficiently far from 0.
8. If the null hypothesis is true, then the test statistic Z has approximately a standard normal distribution. That's how we determine how far is far enough. At a significance level α , "far" means more than $\pm z_{\alpha/2}$ away from 0.

9. Now we carry out the test: If $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$ we **reject** the null hypothesis; otherwise, we stick with the null hypothesis. (Recall that $z_{\alpha/2}$ is defined by the requirement that the area to the right of $z_{\alpha/2}$ under $N(0, 1)$ is $\alpha/2$. Thus, with $\alpha = .05$, the cutoff is 1.96.)
10. Another way to express this is to say that we reject if $|Z| > z_{\alpha/2}$; i.e., we reject if the test statistic Z lands in the set of points having absolute value greater than $z_{\alpha/2}$. This set is called the **rejection region** for the test.
11. Every hypothesis test has this general form: we compute a test statistic from data, then check if the test statistic lands inside or outside the rejection region. The rejection region depends on α but not on the data.
12. Notice that saying

$$-z_{\alpha/2} < \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

is equivalent to saying

$$\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu_0 < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

So, here is another way to think of the test we just did. We found a confidence interval for the mean μ , namely

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}},$$

and checked to see if μ_0 lands in that interval. If μ_0 lands inside, we don't reject H_0 ; if μ_0 lands outside, we do reject H_0 .

13. This supports our intuition that we should reject H_0 if \bar{X} is far from μ_0 .
14. As usual, if we don't know σ we replace it with the sample standard deviation s .
15. **T-test for normal population.** Suppose now that we don't necessarily have a large sample but we do have a normal population. Consider the same hypotheses as before. Now our **test statistic** becomes

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

16. Under the null hypothesis, the test statistic t has a t -distribution with $n - 1$ degrees of freedom.
17. Now we carry out the test. Reject if

$$t < -t_{n-1, \alpha/2} \quad \text{or} \quad t > t_{n-1, \alpha/2};$$

otherwise, do not reject.

18. As before, rejecting based on this rule is equivalent to rejecting whenever μ_0 falls outside the confidence interval for μ .

19. Example: Let's continue with the hypothesis test for the mean return on darts. As above, $\mu_0 = 1.5$ and $\alpha = .05$. Suppose we have $n = 20$ observations (the 1-month contests in the casebook) with a sample mean of -1.0 and a sample standard deviation of 7.2 . Our test statistic is therefore

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = \frac{-1.0 - 1.5}{7.2/\sqrt{20}} = -1.55$$

The threshold for rejection is $t_{19,.025} = 2.093$. Since our test statistic t has an absolute value smaller than the cutoff, we cannot reject the null hypothesis. In other words, based on a significance level of $.05$ the evidence does not significantly support the view that $\mu \neq 1.5$.

20. Keep in mind that the deck is stacked in favor of H_0 ; unless the evidence is very compelling, we stick with the null hypothesis. The smaller we make α , the harder it is to reject H_0 .
21. If a test leads us to reject H_0 , we say that **the results are significant at level α** .

P-Values

1. There is something rather arbitrary about the choice of α . Why should we use $\alpha = .05$ rather than $.01$, $.10$ or some other value? What if we would have rejected H_0 at $\alpha = .10$ but fail to reject it because we chose $\alpha = .05$? Should we change our choice of α ?
2. Changing α after a hypothesis test is "cheating" in a precise sense. Recall that, by the definition of α , the probability of a Type I error is at most α . Thus, fixing α gives us a guarantee on the effectiveness of the test. If we change α , we lose this guarantee.
3. Nevertheless, there is an acceptable way to report what would have happened had we chosen a different significance level. This is based on something called the **p-value** of a test.
4. The **p-value** is the smallest significance level (i.e., the smallest α) at which H_0 would be rejected, for a given test statistic. It is therefore a measure of how significant the evidence in favor of H_1 is: the smaller the p -value, the more compelling the evidence.
5. Example: Consider the test of mean returns on stocks picked by darts, as above. To simplify the present discussion, let's suppose we have 30 data points, rather than 20. (See *Dart Investment Fund* in the casebook for background and data.) As before the hypotheses are

$$H_0 : \quad \mu = 1.5$$

$$H_1 : \quad \mu \neq 1.5$$

Let's suppose that our sample mean \bar{X} (based on 30 observations) is -0.8 and the sample standard deviation is 6.1 . Since we are assuming a large sample, our test statistic is

$$Z = \frac{\bar{X} - 1.5}{s/\sqrt{n}} = \frac{-0.8 - 1.5}{6.1/\sqrt{30}} = -2.06$$

With a significance level of $\alpha = .05$, we get $z_{\alpha/2} = 1.96$, and our rejection region would be

$$Z < -1.96 \text{ or } Z > 1.96.$$

So, in this case, $Z = -2.06$ would be significant: it is sufficiently far from zero to cause us to reject H_0 .

We now ask, what is the smallest α at which we would reject H_0 , based on $Z = -2.06$. We are asking for the smallest α such that $-2.06 < -z_{\alpha/2}$; i.e., the smallest α such that $z_{\alpha/2} < 2.06$. To find this value, we look up 2.06 in the normal table. This gives us .9803. Now we subtract this from 1 to get the area to the right of 2.06. This gives us $1 - .9803 = .0197$. This is the the smallest $\alpha/2$ at which we would reject H_0 . To find the smallest α , we double this to get .0394. This is the p -value for the test. It tells us that we would have rejected the null hypothesis using any α greater than .0394.

6. To verify that this is correct, let's work backwards. Suppose we had chosen $\alpha = .0394$ in the first place. Our rejection cut-off would then have been $z_{\alpha/2} = z_{.0197}$. To find this value, we look up $1 - .0197 = .9803$ in the body of the normal table; we find that $z_{.0197} = 2.06$. So, we reject if $Z < -2.06$ or $Z > 2.06$. Since our test statistic was $Z = 2.06$, we conclude that the p -value .0394 is indeed the significance level at which our test statistic just equals the cutoff.
7. **Summary of steps to find p -value** in a two-sided, large-sample hypothesis test on a population mean:
 - (a) Compute the test statistic Z .
 - (b) Look up the value of $|Z|$ in the normal table.
 - (c) Subtract the number in the table from 1.
 - (d) Multiply by 2; that's the p -value.
8. In principle, to find a p -value based on a t -test we would follow the same steps; however, our t -table does not give us all the information we have in the normal table, so we cannot get the p -value exactly. We can only choose from the α values available on the table or interpolate between them. (The Excel function TDIST gives the area under a t distribution to the right of a specified value and can therefore be used to find a p -value.)
9. Here is another interpretation of the p -value: it is the probability of observing the results actually observed if the null hypothesis were true. Thus, a small p -value implies that it would be very unusual to observe the results actually observed if the null hypothesis were true. This leads us to reject the null hypothesis.
10. Most statistical packages (including spreadsheets) automatically report a p -value when they carry out a hypothesis test. You can then determine, based on this p -value, whether the evidence in favor of H_1 is statistically significant for your purposes.

One-Sided Tests on a Population Mean

1. In the setting considered so far, the null hypothesis is $\mu = \mu_0$, with μ_0 a fixed value. Very large values of \bar{X} and very small values both count as evidence against H_0 . We now consider cases (which are actually more common) in which only values in one direction support the alternative hypothesis. The general setting is this:

$$\begin{aligned}H_0 : & \quad \mu \leq \mu_0 \\H_1 : & \quad \mu > \mu_0\end{aligned}$$

2. Example: Consider the effect of a packaging change on sales of a product. Let μ be the (unknown) mean increase in sales due to the change. We have data available from a test-marketing study. We will not undertake the change unless there is strong evidence in favor of **increased** sales. We should therefore set up the test like this:

$$\begin{aligned}H_0 : & \quad \mu \leq 0 \\H_1 : & \quad \mu > 0\end{aligned}$$

This formulation implies that a large \bar{X} (i.e., large increases in sales in a test market) will support H_1 (i.e., cause us to switch to the new package) but negative values of \bar{X} (reflecting decreased sales) support H_0 .

3. The mechanics of this test are quite similar to those of the two-sided test. The first thing we do is compute a test statistic. Assuming a large sample, we compute

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

If our significance level is α , we find the corresponding value z_α (not $z_{\alpha/2}$!). We reject H_0 if

$$Z > z_\alpha.$$

4. Why are we now using z_α rather than $z_{\alpha/2}$? The short answer is that we are doing a one-sided rather than a two-sided test; so, we need all the area α in one tail rather than split between two tails.
5. A better answer is this: By definition, α controls the probability of a Type I error. Recall that Type I means rejecting H_0 when H_0 is true. If z is our cutoff, then a Type I error means observing $Z > z$ even though H_0 is true. So, we want

$$P(Z > z) = \alpha \text{ when } H_0 \text{ is true.}$$

Recall that when H_0 is true, Z has a standard normal distribution; the value of z that makes $P(Z > z) = \alpha$ is precisely z_α .

6. If you understand the explanation just given, then you have appreciated the fundamental principles of hypothesis testing. If not, just remember that in a two-sided test you use $z_{\alpha/2}$ and in a one-sided test use z_α .

7. We return to the packaging example: Suppose that based on test-marketing in 36 stores we observe a sample mean increase in sales of 13.6 units per week with a sample standard deviation of 42. Is the observed increase significant at level $\alpha = .05$? To answer this, we compute the test statistic

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} = (13.6 - 0)/(42/6) = 1.80.$$

Our cutoff is $z_\alpha = z_{.05} = 1.645$. Since $Z > z_\alpha$, the increase is significant.

8. The t -distribution modification follows the usual pattern. Assuming a normal population, the test statistic

$$t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

has a t -distribution with $n - 1$ degrees of freedom. We reject the null hypothesis if $t > t_{n-1, \alpha}$.

9. The definition of a p -value is the same as before: it is the smallest α at which the results would be rejected. The only difference comes from the fact that our cutoff is now z_α rather than $z_{\alpha/2}$.
10. **Finding p -value** in a **one-sided**, large-sample hypothesis test on a population mean:
- (a) Compute the test statistic Z .
 - (b) Look up the value of Z in the normal table.
 - (c) Subtract the number in the table from 1; that's the p -value.

We no longer multiply by 2.

11. If we reverse the inequalities in the hypotheses to get

$$\begin{aligned} H_0 : & \quad \mu \geq \mu_0 \\ H_1 : & \quad \mu < \mu_0 \end{aligned}$$

the steps are exactly the same as before, except that now we reject H_0 if

$$Z < -z_\alpha \quad \text{or} \quad t < -t_{n-1, \alpha};$$

in other words, when we reverse the inequalities, large negative test statistics support H_1 .

12. To compute the p -value in this case, look up $-Z$ rather than Z . The other steps are unchanged.

Hypothesis Test on a Proportion

1. We now turn to a hypothesis test for a proportion, always assuming a large sample. The mechanics of this case are essentially the same as those for a mean.

2. The general two-sided test is

$$\begin{aligned} H_0 : & \quad p = p_0 \\ H_1 : & \quad p \neq p_0 \end{aligned}$$

and the general one-sided test is

$$\begin{aligned} H_0 : & \quad p \leq p_0 \\ H_1 : & \quad p > p_0 \end{aligned}$$

In all cases, p is an unknown population parameter while p_0 is a number we pick in formulating our hypotheses and is thus known.

3. Example: Let's test whether at least 50% of men who use Rogaine can be expected to show minimal to dense growth. This makes $p_0 = .50$ and p the true (unknown) proportion. Our test is

$$\begin{aligned} H_0 & \quad p \leq 0.5 \\ H_1 & \quad p > 0.5 \end{aligned}$$

The burden of proof is on Rogaine to show that the proportion is greater than one-half.

4. To test the hypotheses, we compute the test statistic

$$Z = \frac{\hat{p} - p_0}{\sigma_{p_0}},$$

where

$$\sigma_{p_0} = \sqrt{\frac{p_0(1 - p_0)}{n}};$$

in other words, σ_{p_0} is what the standard error of \hat{p} would be if p were actually p_0 .

5. In an Upjohn study, 419 men out of 714 had at least minimal growth, so $\hat{p} = .59$. This is more than .50, but does it exceed .50 by a statistically significant amount? Since $p_0 = 0.5$, we have

$$\sigma_{p_0} = \sqrt{\frac{0.5(1 - 0.5)}{714}} = .0183$$

Thus, our test statistic is

$$Z = (.59 - .50)/.0183 = 4.81.$$

We now reject the null hypothesis if $Z > z_\alpha$. Clearly, 4.81 is larger than z_α for any reasonable choice of α , so the results are very significant.

6. The procedure for finding a p -value in this setting is exactly the same as in the test of a mean. In the example just carried out, 4.81 is off the normal table. Since our normal table goes up to 0.9998, we know that the p -value is less than 0.0002; in fact, we could report it as 0.000 since it is 0 to three decimal places. This means that the evidence is so overwhelming that there would be virtually no chance of observing the results in the study if the null hypothesis were true.

Patient Evaluation of Hair Growth at 12 Months					
Total		No Growth	Minimal Growth	Moderate Growth	Dense Growth
619		98	224	246	50
100%		16%	36%	40%	8%
Patient Evaluation of Hair Growth at 4 Months					
		No Growth	Minimal Growth	Moderate Growth	Dense Growth
ROGAINE	714	295	232	180	5
	100%	41%	32%	25%	0.7%
Placebo	717	413	225	75	3
	100%	58%	31%	10%	0.4%

Table 1: Results of Rogaine study published by Upjohn Corporation

7. In the case of a two-sided test, we use $z_{\alpha/2}$ for a cut-off rather than z_{α} . If we reverse the inequalities in H_0 and H_1 , the rejection condition becomes $Z < -z_{\alpha}$ rather than $Z > z_{\alpha}$. (You don't need to memorize this; just think about which direction supports H_1 and which supports H_0 .)

Tests on Differences of Population Means

- Hypothesis testing is frequently used to determine whether observed differences between two populations are significant:
 - Is the observed difference in performance between experts and darts significant?
 - Is the mean midterm score among foreign students really greater than that among non-foreign students, or can the observed difference be attributed to chance?
 - Is a new drug treatment more effective than an existing one?
 - Do full-page ads significantly increase sales compared to half-page ads?

In each case, we are comparing two population means. Assuming we have samples from both populations, we can test hypotheses about the difference between the means.

- We refer to one population using X quantities and the other using Y quantities; the two means are μ_X and μ_Y . The general two-sided test for the difference of two means has the form

$$\begin{aligned} H_0 : \quad & \mu_X - \mu_Y = D_0 \\ H_1 : \quad & \mu_X - \mu_Y \neq D_0 \end{aligned}$$

where D_0 is a specified (known) value we are testing. The general one-sided test is

$$\begin{aligned} H_0 : \quad & \mu_X - \mu_Y \leq D_0 \\ H_1 : \quad & \mu_X - \mu_Y > D_0 \end{aligned}$$

or else the same thing with the inequalities reversed.

3. We concentrate on the case $D_0 = 0$; that is, testing whether there is any difference between the two means. This is the most interesting and important case.
4. As with confidence intervals, here we distinguish a **matched pairs** setting and an **independent** setting. We begin with matched pairs.
5. We assume that we have samples X_1, \dots, X_n and Y_1, \dots, Y_n from the two populations. Each X_i and Y_i are matched — they need not be independent. However, we do assume that (X_1, Y_1) are independent of (X_2, Y_2) , etc. We compute the average difference

$$\bar{d} = \bar{X} - \bar{Y}$$

and the sample standard deviation of the differences

$$s_d = \sqrt{\frac{1}{n-1} \sum_{i=1}^n [(X_i - Y_i) - (\bar{X} - \bar{Y})]^2}.$$

From these we get the test statistic

$$Z = \frac{\bar{d} - D_0}{s/\sqrt{n}}$$

or, in case of samples from normal populations,

$$t = \frac{\bar{d} - D_0}{s/\sqrt{n}}.$$

In the one-sided case, we compare the test statistic with z_α or $t_{n-1, \alpha}$ accordingly; in a two-sided test we use $\pm z_{\alpha/2}$ and $\pm t_{n-1, \alpha/2}$.

6. Example: Let's compare experts and darts. The burden is on the experts to prove they are better than darts, so we have

$$\begin{aligned} H_0 : \quad & \mu_X - \mu_Y \leq 0 \\ H_1 : \quad & \mu_X - \mu_Y > 0 \end{aligned}$$

the X 's refer to expert returns, the Y 's to dart returns. From the 20 1-month contests, we get $\bar{d} = 5.3$ and $s = 6.8$. Our test statistic is

$$t = \frac{5.3 - 0}{6.8/\sqrt{20}} = 3.48$$

Using $\alpha = 0.05$, our cutoff is $t_{19, 0.05} = 1.729$. Since $t > 1.729$ we reject the null hypothesis: the evidence in favor of the experts is significant.

7. Caveat: "Significant" here only refers to the outcome of this test. It does not imply that the conclusion itself is sound. For example, we may have reason to believe that the experts were unduly aided by the publication of their picks. This interferes with the sampling mechanism that generated our 20 data points. The test above is premised on the data coming from a random sample. The test says nothing about the quality of the data itself.

8. Now we consider the case of independent samples. We have samples X_1, \dots, X_{n_X} and Y_1, \dots, Y_{n_Y} from two populations. We assume these are all independent of each other. Consider the test

$$\begin{aligned} H_0 : \quad & \mu_X - \mu_Y \leq D_0 \\ H_1 : \quad & \mu_X - \mu_Y > D_0 \end{aligned}$$

From the sample means \bar{X} and \bar{Y} and the sample standard deviations s_X and s_Y , we compute the test statistic

$$Z = \frac{\bar{X} - \bar{Y} - D_0}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}},$$

assuming that the sample sizes n_X and n_Y are large. We reject H_0 if $Z > z_\alpha$. In a two-sided test, we reject if $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$.

9. Calculating p -values, whether for matched pairs or independent samples works exactly the same way as before. Remember to multiply by 2 in the two-sided case.
10. The case of t -based comparisons is similar but a bit more complicated. If the variances of the two populations are assumed equal, we use a pooled estimate of the standard deviation and a t statistic with $n_X + n_Y - 2$ degrees of freedom. We will not cover this case. It is handled automatically by Excel using **t-Test: Two-Sample Assuming Equal Variances** under **Tools/Data Analysis**.

Testing a Difference of Proportions

1. We now test the difference between two unknown population proportions, p_X and p_Y .
 - (a) Is Rogaine more effective in promoting hair growth than a placebo?
 - (b) Based on a random sample of 100 shoppers, can we conclude that the market share of Colgate toothpaste differs from that of Crest?
 - (c) Is the proportion of women being promoted in an organization significantly smaller than the proportion of men promoted?
2. Let's formulate these comparisons in more detail:
 - (a) With p_X the proportion for Rogaine and p_Y the proportion for the placebo, we want to test

$$\begin{aligned} H_0 : \quad & p_X - p_Y \leq 0 \\ H_1 : \quad & p_X - p_Y > 0 \end{aligned}$$

- (b) With p_X and p_Y the proportion of Colgate and Crest buyers, we are testing

$$\begin{aligned} H_0 : \quad & p_X - p_Y = 0 \\ H_1 : \quad & p_X - p_Y \neq 0 \end{aligned}$$

In other words, we are testing if the market shares are the same or different.

- (c) With p_X the chances a woman is promoted and p_Y the chances a man is promoted, we are testing

$$\begin{aligned} H_0 : & \quad p_X - p_Y \geq 0 \\ H_1 : & \quad p_X - p_Y < 0 \end{aligned}$$

In this formulation, the burden of proof is on the claim of discrimination against women.

3. In each case above, the null hypothesis can equivalently be taken to state that $p_X = p_Y$. We compute a test statistic under the assumption that the null hypothesis is true. If the two proportions are equal, the following is a pooled estimate of the common proportion:

$$\hat{p}_0 = \frac{n_X \hat{p}_X + n_Y \hat{p}_Y}{n_X + n_Y}.$$

The corresponding estimate of the standard error is

$$s_{\hat{p}_0} = \sqrt{\hat{p}_0(1 - \hat{p}_0) \left(\frac{n_X + n_Y}{n_X n_Y} \right)}.$$

Our test statistic is

$$Z = \frac{\hat{p}_X - \hat{p}_Y}{s_{\hat{p}_0}}.$$

For the three tests described above, we have

- (a) Reject H_0 if $Z > z_\alpha$
- (b) Reject H_0 if $Z < -z_{\alpha/2}$ or $Z > z_{\alpha/2}$
- (c) Reject H_0 if $Z < z_\alpha$