Introduction to hypothesis testing: hypothesis testing for proportions

Chrysafis Vogiatzis

Lecture 24-25

Learning objectives

After lectures 24–25, we will be able to:

- Formulate statistical hypotheses for testing.
  - Carefully define null and alternative hypotheses.
  - Define what are the errors of Type I and Type II in hypothesis testing.
- Accept or reject hypotheses for proportions.
- Define the P-value and use it to accept or reject a hypothesis.

Motivation: True or False?

1. 94% of UIUC's College of Engineering graduates secure employment or go to graduate school within a year of graduation.
2. The average starting salary for these Engineering graduates is $78,159.
3. Electrical Engineering or Construction Management? Electrical engineers earn more in the start of their careers.
4. Electrical Engineering or Construction Management? The top 10% construction management professionals earn more than the top 10% electrical engineering professionals.
5. The majority of customers prefers Coke to Pepsi.
6. People with a dog in the house live longer.

What do all the above have in common and what are their differences? How can we *test these claims*? This is what hypothesis testing is all about!

Motivation: Grainger College of Engineering internships

The University of Illinois is interested in finding how many of their Engineering students already have internships lined up for next

summer. The University believes that the proportion is 50%: that is, roughly half the students have secured internships.

The Universite sent out a survey that 140 students filled out with 84 of them stating they have an internship offer at their hands. Is the true percentage 50%? Or is it different than that?

Hypothesis testing

Once more, let us go back to the last weeks of lectures. We have seen **point estimation**, **confidence intervals**, and we are now moving to **hypothesis testing**. A quick review:

How do we estimate an unknown parameter/quantity?

1. **Point estimation**: provides us with a *single estimate* for some unknown parameter of a population.
   - Example: 63% prefer Coke to Pepsi.
2. **Interval estimation**: provides us with a *range/interval containing believable values* for some unknown parameter of a population.
   - Example: The percentage of people preferring Coke to Pepsi lies somewhere between 55% and 71%.
3. **Hypothesis testing**. We form a *hypothesis* or a *claim* for some unknown parameter of a population.
   - Example: Our claim is that more than half of the population prefers Coke to Pepsi.
   - We now need to somehow accept that claim; or reject it, based on **observations**.

Before we formally define hypothesis testing, we ask ourselves a series of motivating questions. Namely, we want to address the following:

1. How do we **formally state a hypothesis**? How do we put it in the proper mathematical terms?
2. When do we **accept** and when do we **reject a hypothesis**?
   - What does "accepting" mean in this mathematical context?
   - What does "rejecting" mean in this mathematical context?
3. What is the **likelihood of reaching the wrong conclusion**? That could mean that..
   - either we accept something that is false.

- or we reject something that is true.

We are now ready to formally define hypothesis testing. We have the following definitions.

Definition 1 (Statistical hypothesis) *With the term* ***statistical hypothesis*** *we mean a claim about some unknown parameters or the unknown distributions of a population. Some examples include:*

- *The mean grade of a student in a class is a B+.*
- *The proportion of students that end up with an A in a class is 25%.*
- *The grade of a student in a class is normally distributed.*

A statistical hypothesis is divided in two parts. The first one is referred to as a **null hypothesis**, H_0, which is the hypothesis/claim that is being tested. As an example, our null hypothesis could be that the mean grade is a B+, or that the true proportion of students with an A is 25%.

The second one is the **alternative hypothesis**, H_1, which is either the opposite of or simply an alternative to the null hypothesis/claim. For example, the alternative hypothesis could be that the mean grade is **not** a B+, or that the true proportion of students with an A is smaller than 25%. We proceed with some examples of formulating statistical hypotheses.

Average grades

Let's assume our claim is the following:

"The average grade of a student in a class is 84&."

Define μ as the average score of a student in a class. Based on that we formulate the statistical hypothesis as:

$$H_0 : \mu = 84\%.$$
$$H_1 : \mu \neq 84\%.$$

Coke vs. Pepsi

Say, our claim is now that:

> "More than half of the population prefers Coke to Pepsi."

Let p be the proportion of people preferring Coke to Pepsi. Then, we can formulate this hypothesis as

$$H_0 : p = 0.5$$
$$H_1 : p < 0.5$$

Note that the null hypothesis is always an equality. The alternate hypothesis though changes depending on our original claim.

Eating greens

What about the following claim?

> "There is no life expectancy change by eating vegetables."

First, we assume we have two populations: one that eats vegetables and one that does not. Let μ_i be the true mean life expectancy of each group. Then, we formulate our hypothesis as:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0.$$

Eating greens: reformulated

This will look very similar. Pay attention to the detail that changes!

> "There is no life expectancy increase by eating vegetables."

Again, we assume the existence of two (eating vs. non-eating vegetables) populations. However, our hypothesis changes slightly now to:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 > 0.$$

Before we head to the next definitions, we summarize some finer details of formulating a hypothesis.

- The null hypothesis is always an equality.
- The alternative hypothesis can be one- or two-sided, depending on the claim we are trying to prove/disprove.
- The hypothesis can deal with a single population; or with the comparison between two populations.

Let us get to the fundamental part of this lecture. **How do we perform hypothesis tests?** How do we decide whether we have enough information to accept or reject a hypothesis?

Definition 2 (Hypothesis test) *A **hypothesis test** is a statistical procedure to collect information based on a random sample, which can lead to making a decision about the null hypothesis.*

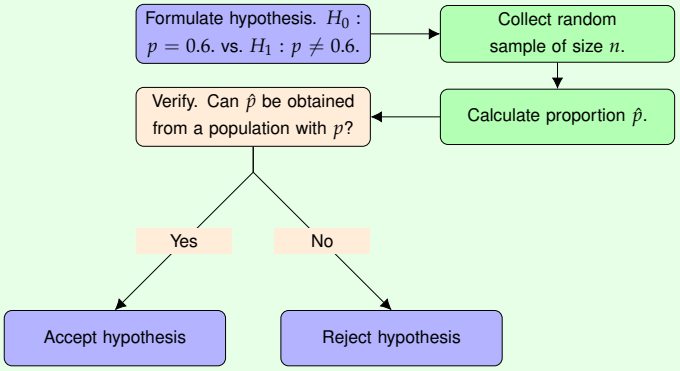
Let's see this with an example.

Coke vs. Pepsi

Assume you want to check whether 60% of the people prefer Coke to Pepsi. We can do the following operations.

1. First, formulate the statistical hypothesis as $H_0 : p = 0.6$ vs. $H_1 : p \neq 0.6$. We could have formulated a one-sided alternative hypothesis as $H_1 : p > 0.6$ or $H_1 : p < 0.6$ if we had more information about the original claim, but now $H_1 : p \neq 0.6$ will do.
2. Secondly, collect a random sample. Use it to estimate the proportion of people observed that prefer Coke to Pepsi.
3. Thirdly, try to verify. If your original hypothesis/claim is true, could you have gotten the observed proportion in the sample? If so, accept; if not, reject.

Visually:



We proceed to discuss how we may accept or reject a hypothesis. First of all, let us get one thing out of the way. While "accepting" and "rejecting" are universally used for hypothesis testing, it is more correct to think of them as "failing to reject" and "rejecting".

Think of the following parallel: say you are the jury at a trial. The hypothesis is that the defendant is innocent, no? The attorneys present data (observations) and it is up to you to decide whether it is enough to "reject innocence" or "fail to reject innocence". Note that failing to reject innocence is not the same as being innocent! It merely implies that there was not enough evidence to persuade you.

So, how does that translate to hypothesis testing? Let us study this using proportions.

Hypothesis testing for proportions

Assume we have a population X that has some unknown proportion p. We collect a sample of size n from that population. Recall that if we have $np \geq 5, n(1-p) \geq 5$, then we may make the claim that the observed proportion out of a sample of size n (defined as $\hat{p} = \frac{x}{n}$) is distributed as $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$. [1]

[1] This comes straight from our discussion about confidence intervals on proportions. See Lectures 20-23.

For the sake of the example, assume that we have

$$H_0 : p = 0.6.$$
$$H_1 : p \neq 0.6.$$

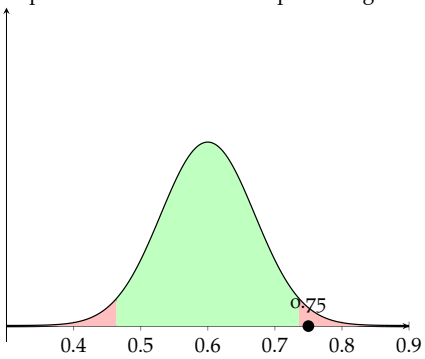
Say we have collected a sample of size $n = 50$. Then, **if the null hypothesis is true**, we'd expect a distribution of $\mathcal{N}(0.6, 0.0048)$. Visually, we get a normal distribution as the one presented in Figure 1. Now, say we select a confidence level of 95%. That means, visually, we'd expect 95% of the potential sample averages to fall in the green area; not the red. Finally, let's say that our sample average (for this $n = 50$) amounts to $\hat{p} = 0.75$. We also mark that in the figure.

We claim that the above figure essentially captures hypothesis testing. The question becomes: does the observed proportion $\hat{p} = 0.75$ fall in the range of believable values (the critical regions of **accepting**)? Or does it fall outside them (in the critical regions of **rejecting** the hypothesis)?

Based on this, let us revisit the terms we use. **Accepting** and **rejecting** a hypothesis are not the most appropriate terms for the outcomes of a hypothesis test. Instead, from now on, we will write that we:

- **Reject** the hypothesis, when we have sufficient observations to claim that the null hypothesis is not true.
  - This is a s **strong conclusion**.

Figure 1: A figure showing the distribution of the population **if the null hypothesis is true**, green and red areas marking the critical acceptance and rejection regions, and a point at $\hat{p} = 0.75$ that represents the obtained sample average.



  - It implies the existence of sufficient evidence agains the hypothesis.
  - In the end of this, we are quite certain that H_0 is wrong.

- **Fail to reject** the hypothesis, when we are not sure about the validity of the null hypothesis.

  - Consequently, it is a **weak conclusion**.
  - It merely implies the lack of sufficient evidence agains the hypothesis.
  - It does not mean that H_0 is true! It only implies that we are uncertain about either H_0 or H_1 being true.

Reaching the wrong conclusions

How many types of errors do you foresee appearing with this way of testing a hypothesis? Let's see this in tabular form:

Decision	H_0 is true	H_0 is false
Reject H_0	incorrect decision	correct decision
Fail to reject H_0	correct decision	incorrect decision

We will then need to formally define these two types of errors. Their names are "uninspired". These two are defined as **Type I** or α error [2] and **Type II** or β error.

[2] α? Is it.. the same as α in confidence intervals? Oh, yes. Yes it is.

Definition 3 (Type I errors) *The Type I or α error happens when we reject H_0 even though H_0 is valid. It is quantified as*

$$\alpha = P(\textit{reject } H_0 \textit{ when } H_0 \textit{ is true}) = P(\textit{reject } H_0 | H_0).$$

Definition 4 (Type II errors) *The Type II or β error happens when we fail to reject H_0 even though H_0 is not true. It is quantified as*

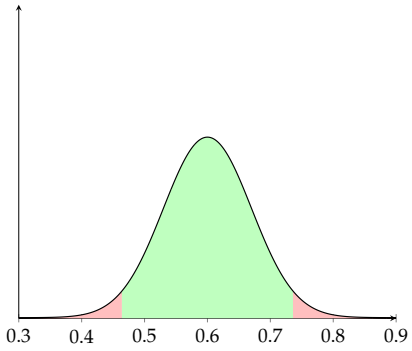
$$\beta = P(\textit{fail to reject } H_0 \textit{ when } H_0 \textit{ is false}) = P(\textit{fail to reject } H_0 | \overline{H}_0).$$

The courthouse parallel

Take a minute and think of the parallels to the jury trial example from before. What is α and β in a trial setting?

Let us focus on α. We have been using it all along!

$$\boxed{\alpha = P(\text{Type I error}) = P(\text{reject } H_0 \text{ when } H_0 \text{ is true}).}$$



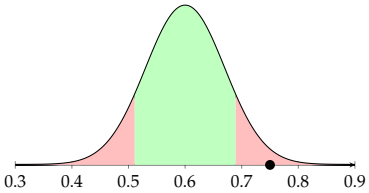
- $1 - \alpha$ is called the *significance* or the *size* of the test.
- It is equivalent to the red shaded areas.
- It can be improved by selecting stricter confidence levels.

The courthouse parallel (cont'd)

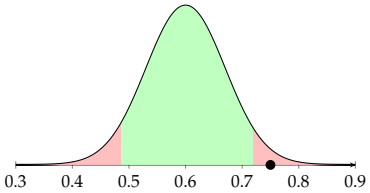
You can improve α in a trial by asking for more and more evidence. For example, "I will not find anyone guilty unless you present video evidence that they have done it" increases α signficantly, doesn't it? I wonder what happens to β, though...

Let us see another example for the effect of α. In our motivating example, we asked $n = 50$ people to check the hypothesis that $p = 0.6$. Assume out of that sample we get $\hat{p} = 0.75$. Then, the following would be the visual results for different significance levels (values for α):

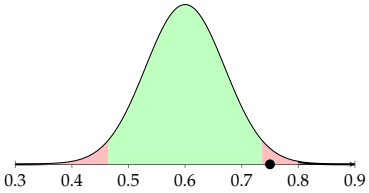
$\alpha = 20\%$



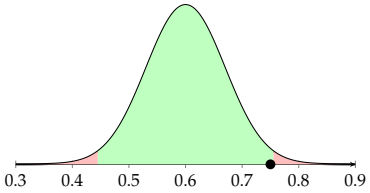
$\alpha = 10\%$



$\alpha = 5\%$



$\alpha = 2.5\%$



Hence, the bigger the α we are willing to accept, the tougher it becomes to reject a hypothesis. When $\alpha = 0$ (no error accepted!), then we no longer can reject a hypothesis.

We now move to β.

$$\boxed{\beta = P(\text{Type II error}) = P(\text{fail to reject } H_0 \text{ when } H_0 \text{ is false}).}$$

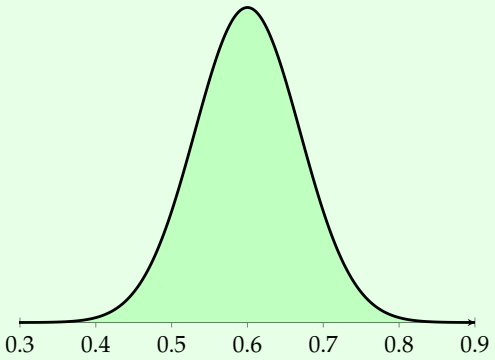
- It is related to $1 - \beta$, the *power* of the test.
- To formulate, it requires a **specific alternative hypothesis**.
- It decreases as the difference between the hypothesized and the true value of the hypothesis increases.

What this tells us is that β is not universal, given a hypothesis test. Instead, it depends on what we are comparing H_0 to. We show this in practice in the next pages.

Type II errors

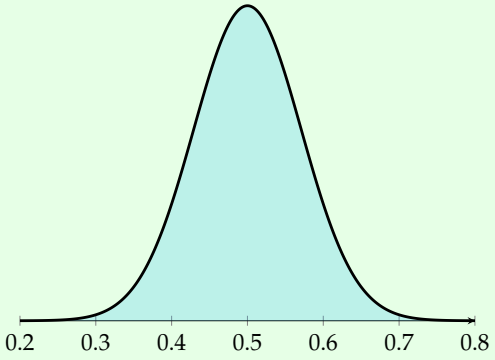
We have a company that offers a service that needs to be at 60% or above. The company is in trouble when the service quality lowers at 50%. To avoid this, they have an inspection mechanism in place. From time to time, they collect a sample of $n = 50$ services and make sure that average lies in the acceptance region! How often are they wrong and they *believe* they are good when they are not? What is the probability they accept the hypothesis that $p = 0.6$ when in fact the true p has lowered to 0.5?

The sampling distribution for $n = 50$ **if** the null hypothesis that $p = 0.6$ is true. Recall this is $\mathcal{N}(0.6, 0.0048)$.



In a similar manner, we can represent the sampling distribution for $n = 50$ when $p = 0.5$ instead! It would be $\mathcal{N}(0.5, 0.05)$.

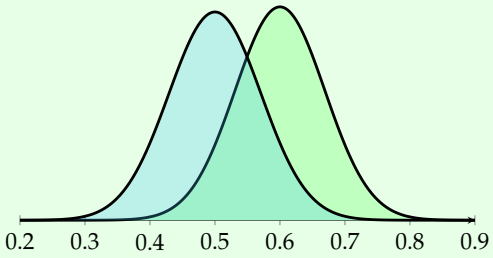
The sampling distribution for $n = 50$ **if** $p = 0.5$ is true.



Let us try to plot these two together!

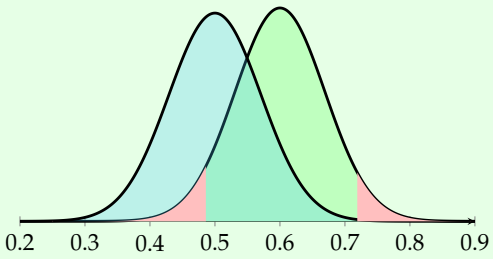
Type II errors

Plotting the two together reveals quite the overlap.



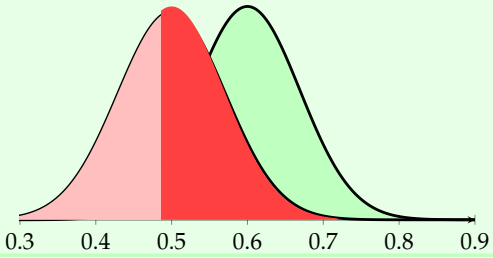
This means that it is quite possible that a value in the overlap may correspond to a "reality" of $p = 0.6$ or one of $p = 0.5$. But, remember! We only accept part of the first curve, depending on our α! Let's add this to the plot!

When we add the regions where we'd reject the original null hypothesis $H_0 : p = 0.6$. Here we use $\alpha = 10\%$.



Still, though. Observe the area in dark red below.

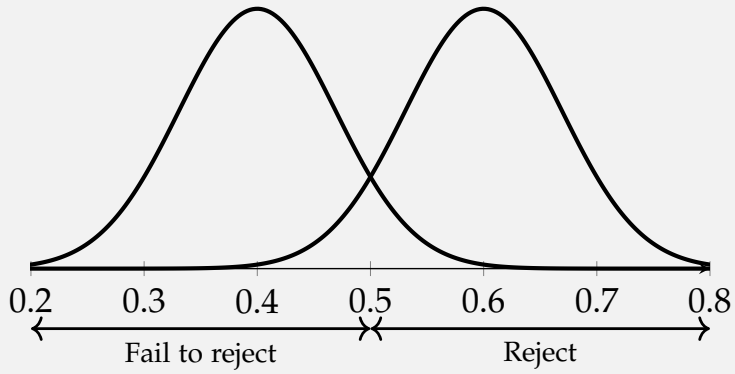
The dark area represents the β error! The lighter red area shows the power of the test $(1 - \beta)$.



As practice, paint the following.

Practice with the visuals

Here, we assume that $H_0 : p = 0.4$ (the null hypothesis) and say the alternative is $p = 0.6$. We have already provided where you would reject the null hypothesis and where you would not. Mark the following areas: (i) the region where you reject the null hypothesis, (ii) the region where you have the β error, (iii) the region of the power of the test $(1 - \beta)$.



How could we mathematically calculate the β error? Let us see the red area we need to be covering. That would be between $0.6 - z_{\alpha/2} \cdot 0.0693$ (recall that we have $\mathcal{N}(0.6, 0.048)$, so $\sqrt{0.048} = 0.0693$) and $0.6 + z_{\alpha/2} \cdot 0.0693$. For the sake of the example let us use $\alpha = 10\%$, which leads to $z_{0.05} = 1.645 \implies 0.6 - z_{\alpha/2} \cdot 0.0693 = 0.486$ and $0.6 + z_{\alpha/2} \cdot 0.0693 = 0.714$. Hence, we have, assuming that $p = 0.5$ is right and hence distributed with $\mathcal{N}(0.5, 0.005)$:

$$\beta = P(0.486 \leq p \leq 0.714) = P(p \leq 0.714) - P(p \leq 0.486) =$$
$$= \Phi\left(\frac{0.714 - 0.5}{\sqrt{0.005}}\right) - \Phi\left(\frac{0.486 - 0.5}{\sqrt{0.005}}\right) = \Phi(3.03) - \Phi(-0.20) = \Phi(3.03) - 1 + \Phi(0.20) =$$
$$= 0.9988 - 1 + 0.5793 = 57.81\%.$$

Before we finish this discussion, we provide a couple of observations about the Type I and Type II errors:

- Observation #1: assuming a fixed sample size, then decreasing one error will result in an increase of the other error.
  - Decreasing α will imply an increase in β.

  - Decreasing β will imply an increase in α.
- Observation #2: both errors can be reduced by increasing the sample size.

Finishing the proportion hypothesis testing procedure

We are *finally* ready to finish the discussion on hypothesis testing for proportions! We separate our discussion in three cases, depending on the hypothesis testing format (two-sided or one-sided).

Two-sided hypothesis testing

1. **Preliminaries.**
   - Select the desired α (significance $1-\alpha$).
   - Set up your hypothesis test as:

$$H_0 : p = p_0$$
$$H_1 : p \neq p_0.$$

2. **Compute test statistic** based on sample of size n.
   - $\hat{p}$

     or
   - $Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.
3. **Check.**
   - Is $\hat{p}$ below $p_0 - z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$ or above $p_0 + z_{\alpha/2}\sqrt{\frac{p_0(1-p_0)}{n}}$?
   - Equivalently, is Z_0 below $-z_{\alpha/2}$ or above $z_{\alpha/2}$?
4. **Decide.**
   - If the check is true, reject the hypothesis.
   - Otherwise, fail to reject it.

To calculate the power of the test (or β), first identify the alternative you are investigating, say $p = p_1$. Then, assume that your sample is distributed such that $p \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$. Finally, calculate:

$$\beta = P\left(p_0 - z_{\alpha/2}\sqrt{\frac{p_0\,(1-p_0)}{n}} \leq p \leq p_0 + z_{\alpha/2}\sqrt{\frac{p_0\,(1-p_0)}{n}}\right).$$

One-sided hypothesis testing Assume now that we are looking for the upper alternative hypothesis ($p > p_0$).

1. **Preliminaries.**
   - Select the desired α (significance $1 - \alpha$).
   - Set up your hypothesis test as:

$$H_0 : p = p_0$$
$$H_1 : p > p_0.$$

2. **Compute test statistic** based on sample of size n.
   - $\hat{p}$

     or
   - $Z_0 = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.
3. **Check.**
   - Is $\hat{p}$ above $p_0 + z_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$?
   - Equivalently, is Z_0 above z_α?
4. **Decide.**
   - If the check is true, reject the hypothesis.
   - Otherwise, fail to reject it.

To calculate the power of the test (or β), again identify the alternative you are investigating, say $p = p_1$. Then, assume that your sample is distributed such that $p \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$. Finally, calculate:

$$\beta = P\left(p \leq p_0 + z_\alpha \sqrt{\frac{p_0\,(1-p_0)}{n}}\right).$$

For the lower alternative hypothesis ($p < p_0$), we take very similar steps.

1. **Preliminaries.**
   - Select the desired α (significance $1 - \alpha$).
   - Set up your hypothesis test as:

$$H_0 : p = p_0$$
$$H_1 : p < p_0.$$

2. **Compute test statistic** based on sample of size n.
   - $\hat{p}$

     or
   - $Z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$.
3. **Check.**
   - Is $\hat{p}$ below $p_0 - z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}$?
   - Equivalently, is Z_0 below $-z_\alpha$?
4. **Decide.**
   - If the check is true, reject the hypothesis.
   - Otherwise, fail to reject it.

To calculate the power of the test (or β), first identify the alternative you are investigating, say $p = p_1$. Then, assume that your sample is distributed such that $p \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$. Finally, calculate:

$$\beta = P\left(p_0 - z_\alpha\sqrt{\frac{p_0\left(1-p_0\right)}{n}} \leq p\right).$$

A comprehensive example

We claim that the percentage of people in favor of a law is 0.5. A sample of 50 people gave $\hat{p} = 0.62$. Our hypothesis then is that $H_0 : p = 0.5$.

1. We would like the limits of our hypothesis test to be between 0.45 and 0.55. What is α?
2. What is the acceptance region for $\alpha = 0.05$ and a two-sided test? Can we reject the null hypothesis in favor of the alternative $p \neq 0.5$?
3. What is the acceptance region for $\alpha = 0.05$ and a one-sided test (alternative is $H_1 : p > 0.5$)? Can we reject the null hypothesis in favor of the alternative?
4. What is β if the true percentage in favor of the law is 0.70? Assume we are interested in a one-sided (upper) hypothesis test.

A comprehensive example

Recall that $\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$ and if $p = 0.5$ then $\hat{p} \sim \mathcal{N}(0.5, 0.005)$. We have

$$\begin{aligned} 1-\alpha &= P(0.45 \leq \hat{p} \leq 0.55) \implies \\ \implies \alpha &= P(\hat{p} < 0.45) + P(\hat{p} > 0.55) = \\ &= 2 - 2\Phi(0.71) = 2 - 1.5222 = 0.4778. \end{aligned}$$

Hence $\alpha = 0.5222 = 52.22\%$.

For the second part, this is easier: for $\alpha = 0.05$, we have $z_{\alpha/2} = z_{0.025} = 1.96$. Hence, the acceptance region would be between $0.5 - 1.96\sqrt{0.005} = 0.361$ and $0.5 + 1.96\sqrt{0.005} = 0.639$. We fail to reject the hypothesis, and hence we do not have enough evidence to disagree with $p = 50\%$.

For the third part, the only difference is that we are only focused on the alternative hypothesis of $H_1 : p > p_0$. Hence, we could only reject on that side. For $\alpha = 0.05$, we now use $z_\alpha = z_{0.05} = 1.645$ and we get: $0.5 + 1.645\sqrt{0.005} = 0.616$. The acceptance region is between o and o.616. This means that we do have enough evidence to reject the null hypothesis now! We have enough evidence to disagree with $p = 50\%$ in favor of $p > 50\%$.

Finally, for the power of the test against $p = 0.7$: we already have that the upper limit is equal to 0.616. Hence, we would reject the hypothesis for any $\hat{p}$ above this. We are then looking at

$$\begin{aligned} 1-\beta &= P(\hat{p} > 0.616) = 1 - \Phi\left(\frac{0.616 - 0.7}{\sqrt{0.7 \cdot 0.3/50}}\right) = \\ &= 1 - \Phi(-1.30) = \Phi(1.30) = 0.9032. \end{aligned}$$

This is a pretty powerful test, even with a small sample size (comparatively) at $n = 50$.