OpenIntro online supplement

This and other free online supplements to OpenIntro textbooks may be found at

openintro.org/books

Each of our textbooks is available for free in PDF and for $20 or less in print, offering an affordable option at a time where introductory statistics textbook prices regularly reach over $200 from big publishers.

1 Small sample hypothesis testing for a proportion

In this section we develop inferential methods for a single proportion that are appropriate when the sample size is too small to apply the normal model to $\hat{p}$. Just like the methods related to the t-distribution, these methods can also be applied to large samples.

1.1 When the success-failure condition is not met

People providing an organ for donation sometimes seek the help of a special "medical consultant". These consultants assist the patient in all aspects of the surgery, with the goal of reducing the possibility of complications during the medical procedure and recovery. Patients might choose a consultant based in part on the historical complication rate of the consultant's clients. One consultant tried to attract patients by noting the average complication rate for liver donor surgeries in the US is about 10%, but her clients have only had 3 complications in the 62 liver donor surgeries she has facilitated. She claims this is strong evidence that her work meaningfully contributes to reducing complications (and therefore she should be hired!).

GUIDED PRACTICE 0.1

(G) We will let p represent the true complication rate for liver donors working with this consultant. Estimate p using the data, and label this value $\hat{p}$.[1]

EXAMPLE 0.2

Is it possible to assess the consultant's claim with the data provided?

(E) No. The claim is that there is a causal connection, but the data are observational. Patients who hire this medical consultant may have lower complication rates for other reasons.

While it is not possible to assess this causal claim, it is still possible to test for an association using these data. For this question we ask, could the low complication rate of $\hat{p} = 0.048$ be due to chance?

GUIDED PRACTICE 0.3

(G) Write out hypotheses in both plain and statistical language to test for the association between the consultant's work and the true complication rate, p, for this consultant's clients.[2]

EXAMPLE 0.4

In the examples based on large sample theory, we modeled $\hat{p}$ using the normal distribution. Why is this not appropriate here?

(E) The independence assumption may be reasonable if each of the surgeries is from a different surgical team. However, the success-failure condition is not satisfied. Under the null hypothesis, we would anticipate seeing $62 \times 0.10 = 6.2$ complications, not the 10 required for the normal approximation.

The uncertainty associated with the sample proportion should not be modeled using the normal distribution. However, we would still like to assess the hypotheses from Guided Practice 3 in absence of the normal framework. To do so, we need to evaluate the possibility of a sample value ($\hat{p}$) this far below the null value, $p_0 = 0.10$. This possibility is usually measured with a p-value.

[1] The sample proportion: $\hat{p} = 3/62 = 0.048$

[2] H_0: There is no association between the consultant's contributions and the clients' complication rate. In statistical language, $p = 0.10$. H_A: Patients who work with the consultant tend to have a complication rate lower than 10%, i.e. $p < 0.10$.

The p-value is computed based on the null distribution, which is the distribution of the test statistic if the null hypothesis is true. Supposing the null hypothesis is true, we can compute the p-value by identifying the chance of observing a test statistic that favors the alternative hypothesis at least as strongly as the observed test statistic. This can be done using simulation.

1.2 Generating the null distribution and p-value by simulation

We want to identify the sampling distribution of the test statistic ($\hat{p}$) if the null hypothesis was true. In other words, we want to see how the sample proportion changes due to chance alone. Then we plan to use this information to decide whether there is enough evidence to reject the null hypothesis.

Under the null hypothesis, 10% of liver donors have complications during or after surgery. Suppose this rate was really no different for the consultant's clients. If this was the case, we could *simulate* 62 clients to get a sample proportion for the complication rate from the null distribution.

Each client can be simulated using a deck of cards. Take one red card, nine black cards, and mix them up. Then drawing a card is one way of simulating the chance a patient has a complication *if the true complication rate is 10%* for the data. If we do this 62 times and compute the proportion of patients with complications in the simulation, $\hat{p}_{sim}$, then this sample proportion is exactly a sample from the null distribution.

An undergraduate student was paid $2 to complete this simulation. There were 5 simulated cases with a complication and 57 simulated cases without a complication, i.e. $\hat{p}_{sim} = 5/62 = 0.081$.

EXAMPLE 0.5

Is this one simulation enough to determine whether or not we should reject the null hypothesis from Guided Practice 3? Explain.

(E)

No. To assess the hypotheses, we need to see a distribution of many $\hat{p}_{sim}$, not just a *single* draw from this sampling distribution.

One simulation isn't enough to get a sense of the null distribution; many simulation studies are needed. Roughly 10,000 seems sufficient. However, paying someone to simulate 10,000 studies by hand is a waste of time and money. Instead, simulations are typically programmed into a computer, which is much more efficient.

Figure 1 shows the results of 10,000 simulated studies. The proportions that are equal to or less than $\hat{p} = 0.048$ are shaded. The shaded areas represent sample proportions under the null distribution that provide at least as much evidence as $\hat{p}$ favoring the alternative hypothesis. There were 1222 simulated sample proportions with $\hat{p}_{sim} \leq 0.048$. We use these to construct the null distribution's left-tail area and find the p-value:

$$\text{left tail} = \frac{\text{Number of observed simulations with } \hat{p}_{sim} \leq 0.048}{10000} \tag{6}$$

Of the 10,000 simulated $\hat{p}_{sim}$, 1222 were equal to or smaller than $\hat{p}$. Since the hypothesis test is one-sided, the estimated p-value is equal to this tail area: 0.1222.

GUIDED PRACTICE 0.7

(G) Because the estimated p-value is 0.1222, which is larger than the significance level 0.05, we do not reject the null hypothesis. Explain what this means in plain language in the context of the problem.[3]

GUIDED PRACTICE 0.8

(G) Does the conclusion in Guided Practice 7 imply there is no real association between the surgical consultant's work and the risk of complications? Explain.[4]

[3] There isn't sufficiently strong evidence to support an association between the consultant's work and fewer surgery complications.

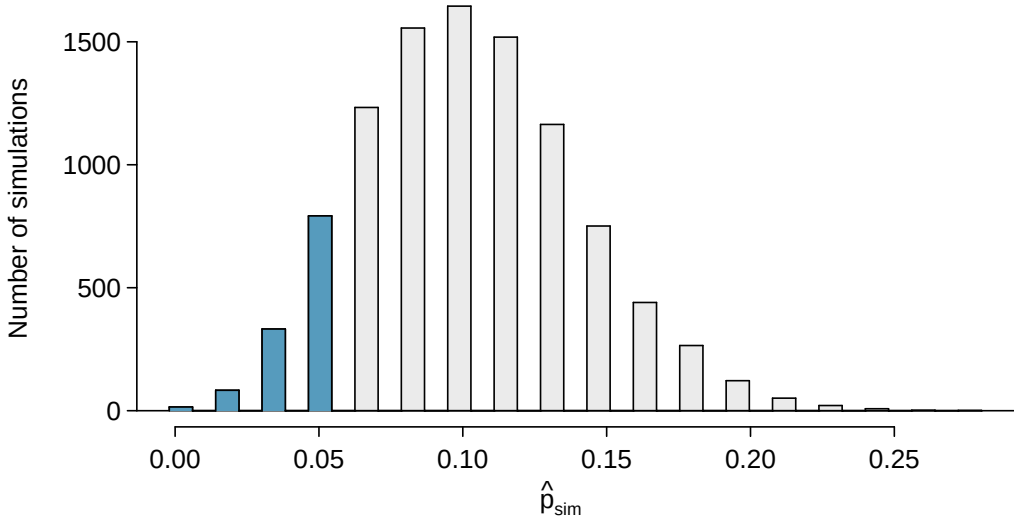

Figure 1: The null distribution for $\hat{p}$, created from 10,000 simulated studies. The left tail, representing the p-value for the hypothesis test, contains 12.22% of the simulations.

ONE-SIDED HYPOTHESIS TEST FOR p WITH A SMALL SAMPLE

The p-value is always derived by analyzing the null distribution of the test statistic. The normal model poorly approximates the null distribution for $\hat{p}$ when the success-failure condition is not satisfied. As a substitute, we can generate the null distribution using simulated sample proportions ($\hat{p}_{sim}$) and use this distribution to compute the tail area, i.e. the p-value.

We continue to use the same rule as before when computing the p-value for a two-sided test: double the single tail area, which remains a reasonable approach even when the sampling distribution is asymmetric. However, this can result in p-values larger than 1 when the point estimate is very near the mean in the null distribution; in such cases, we write that the p-value is 1. Also, very large p-values computed in this way (e.g. 0.85), may also be slightly inflated.

Guided Practice 7 said the p-value is *estimated*. It is not exact because the simulated null distribution itself is not exact, only a close approximation. However, we can generate an exact null distribution and p-value using the binomial model.

1.3 Generating the exact null distribution and p-value

The number of successes in n independent cases can be described using the binomial model. Recall that the probability of observing exactly k successes is given by

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \tag{9}$$

where p is the true probability of success. The expression $\binom{n}{k}$ is read as n *choose* k, and the exclamation points represent factorials. For instance, 3! is equal to $3 \times 2 \times 1 = 6$, 4! is equal to $4 \times 3 \times 2 \times 1 = 24$, and so on.

The tail area of the null distribution is computed by adding up the probability in Equation (9) for each k that provides at least as strong of evidence favoring the alternative hypothesis as the data. If the hypothesis test is one-sided, then the p-value is represented by a single tail area. If the test is two-sided, compute the single tail area and double it to get the p-value, just as we have done in the past.

[4]No. It might be that the consultant's work is associated with a reduction but that there isn't enough data to convincingly show this connection.

EXAMPLE 0.10

Compute the exact p-value to check the consultant's claim that her clients' complication rate is below 10%.

Exactly $k = 3$ complications were observed in the $n = 62$ cases cited by the consultant. Since we are testing against the 10% national average, our null hypothesis is $p = 0.10$. We can compute the p-value by adding up the cases where there are 3 or fewer complications:

$$\begin{aligned}
\text{p-value} &= \sum_{j=0}^{3} \binom{n}{j} p^j (1-p)^{n-j} \\
&= \sum_{j=0}^{3} \binom{62}{j} 0.1^j (1-0.1)^{62-j} \\
&= \binom{62}{0} 0.1^0 (1-0.1)^{62-0} + \binom{62}{1} 0.1^1 (1-0.1)^{62-1} \\
&\quad + \binom{62}{2} 0.1^2 (1-0.1)^{62-2} + \binom{62}{3} 0.1^3 (1-0.1)^{62-3} \\
&= 0.0015 + 0.0100 + 0.0340 + 0.0755 \\
&= 0.1210
\end{aligned}$$

This exact p-value is very close to the p-value based on the simulations (0.1222), and we come to the same conclusion. We do not reject the null hypothesis, and there is not statistically significant evidence to support the association.

If it were plotted, the exact null distribution would look almost identical to the simulated null distribution shown in Figure 1 on page 4.

1.4 Using simulation for goodness of fit tests

Simulation methods may also be used to test goodness of fit. In short, we simulate a new sample based on the purported bin probabilities, then compute a chi-square test statistic X^2_{sim}. We do this many times (e.g. 10,000 times), and then examine the distribution of these simulated chi-square test statistics. This distribution will be a very precise null distribution for the test statistic X^2 if the probabilities are accurate, and we can find the upper tail of this null distribution, using a cutoff of the observed test statistic, to calculate the p-value.

EXAMPLE 0.11

Let's consider an example evaluating whether jurors are racially representative of the population. Would our findings differ if we used a simulation technique? This example was considered with large sample methods in *OpenIntro Statistics* textbook, and we'll compare results to using small sample methods.

Since the minimum bin count condition was satisfied, the chi-square distribution is an excellent approximation of the null distribution, meaning the results should be very similar. Figure 2 shows the simulated null distribution using 100,000 simulated X^2_{sim} values with an overlaid curve of the chi-square distribution. The distributions are almost identical, and the p-values are essentially indistinguishable: 0.115 for the simulated null distribution and 0.117 for the theoretical null distribution.

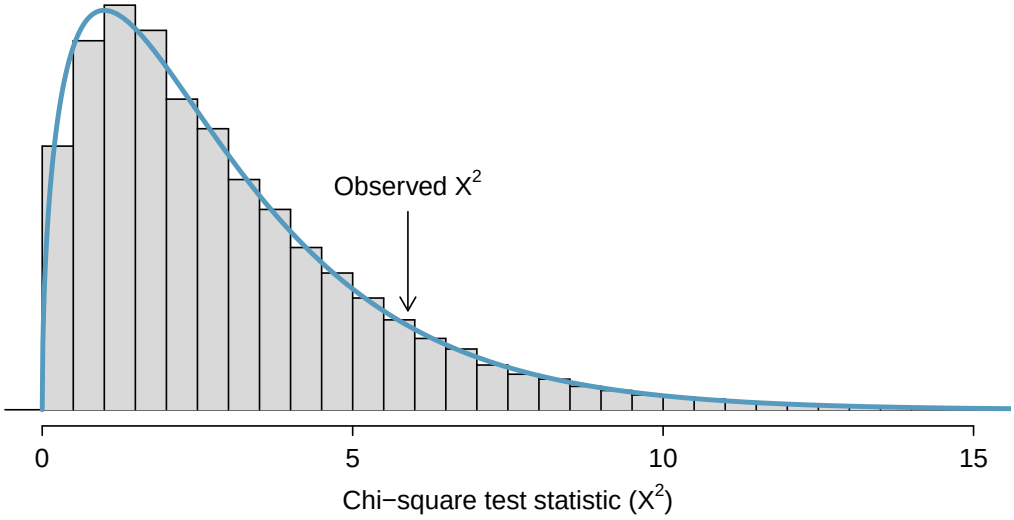

Figure 2: The precise null distribution for the juror example is shown as a histogram of simulated X^2_{sim} statistics, and the theoretical chi-square distribution is also shown.

2 Randomization test

Cardiopulmonary resuscitation (CPR) is a procedure commonly used on individuals suffering a heart attack when other emergency resources are not available. This procedure is helpful in maintaining some blood circulation, but the chest compressions involved can also cause internal injuries. Internal bleeding and other injuries complicate additional treatment efforts following arrival at a hospital. For example, blood thinners may be used to release a clot responsible for a heart attack. However, the blood thinner would negatively affect internal bleeding.

We consider an experiment for patients who underwent CPR for a heart attack and were subsequently admitted to a hospital.[5] These patients were randomly divided into a treatment group where they received a blood thinner or the control group where they did not receive a blood thinner. The outcome variable of interest was whether the patients survived for at least 24 hours.

EXAMPLE 0.12

What is an appropriate set of hypotheses for this study? Let p_c represent the true survival rate of people who do not receive a blood thinner (corresponding to the control group) and p_t represent the survival rate for people receiving a blood thinner (corresponding to the treatment group).

(E) We are interested in whether the blood thinners are helpful or harmful, so a two-sided test is appropriate.

H_0: Blood thinners do not have an overall survival effect, i.e. the survival proportions are the same in each group. $p_t - p_c = 0$.

H_A: Blood thinners do have an impact on survival. $p_t - p_c \neq 0$.

2.1 Large sample framework for a difference in two proportions

There were 50 patients in the experiment who did not receive the blood thinner and 40 patients who did. The study results are shown in Figure 3.

	Survived	Died	Total
Control	11	39	50
Treatment	14	26	40
Total	25	65	90

Figure 3: Results for the CPR study. Patients in the treatment group were given a blood thinner, and patients in the control group were not.

GUIDED PRACTICE 0.13

(G) What is the observed survival rate in the control group? And in the treatment group? Also, provide a point estimate of the difference in survival proportions of the two groups: $\hat{p}_t - \hat{p}_c$.[6]

According to the point estimate, for patients who have undergone CPR outside of the hospital, an additional 13% survive when they are treated with blood thinners. However, this difference might be explainable by chance. We'd like to investigate this using a large sample framework, but we first need to check the conditions for such an approach.

[5] *Efficacy and safety of thrombolytic therapy after initially unsuccessful cardiopulmonary resuscitation: a prospective clinical trial*, by Böttiger et al., The Lancet, 2001.

[6] Observed control survival rate: $p_c = \frac{11}{50} = 0.22$. Treatment survival rate: $p_t = \frac{14}{40} = 0.35$. Observed difference: $\hat{p}_t - \hat{p}_c = 0.35 - 0.22 = 0.13$.

EXAMPLE 0.14

Can the point estimate of the difference in survival proportions be adequately modeled using a normal distribution?

(E)

We will assume the patients are independent, which is probably reasonable. The success-failure condition is also satisfied. Since the proportions are equal under the null, we can compute the pooled proportion, $\hat{p}_{pooled} = (11 + 14)/(50 + 40) = 0.278$, for checking conditions. We find the expected number of successes (13.9, 11.1) and failures (36.1, 28.9) are above 10. The normal model is reasonable.

While we can apply a normal framework as an approximation to find a p-value, we might keep in mind that the expected number of successes is only 13.9 in one group and 11.1 in the other. Below we conduct an analysis relying on the large sample normal theory. We will follow up with a small sample analysis and compare the results.

EXAMPLE 0.15

Assess the hypotheses presented in Example 12 using a large sample framework. Use a significance level of $\alpha = 0.05$.

(E)

We suppose the null distribution of the sample difference follows a normal distribution with mean 0 (the null value) and a standard deviation equal to the standard error of the estimate. The null hypothesis in this case would be that the two proportions are the same, so we compute the standard error using the pooled proportion:

$$SE = \sqrt{\frac{p(1-p)}{n_t} + \frac{p(1-p)}{n_c}} \approx \sqrt{\frac{0.278(1-0.278)}{40} + \frac{0.278(1-0.278)}{50}} = 0.095$$

where we have used the pooled estimate $\left(\hat{p}_{pooled} = \frac{11+14}{50+40} = 0.278\right)$ in place of the true proportion, p.

The null distribution with mean zero and standard deviation 0.095 is shown in Figure 4. We compute the tail areas to identify the p-value. To do so, we use the Z-score of the point estimate:

$$Z = \frac{(\hat{p}_t - \hat{p}_c) - \text{null value}}{SE} = \frac{0.13 - 0}{0.095} = 1.37$$

We can use software or a normal probability table to find the right tail area: 0.0853. The p-value is twice the single tail area: 0.1706. This p-value does not provide convincing evidence that the blood thinner helps. Thus, there is insufficient evidence to conclude whether or not the blood thinner helps or hurts. (Remember, we never "accept" the null hypothesis – we can only reject or fail to reject.)

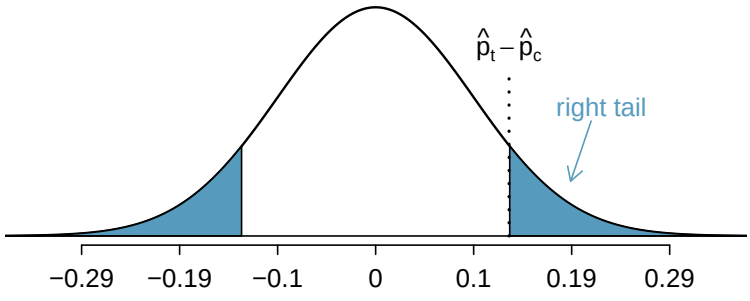

Figure 4: The null distribution of the point estimate $\hat{p}_t - \hat{p}_c$ under the large sample framework is a normal distribution with mean 0 and standard deviation equal to the standard error, in this case $SE = 0.095$. The p-value is represented by the shaded areas.

The p-value 0.176 relies on the normal approximation. We know that when the samples sizes are large, this approximation is quite good. However, when the sample sizes are relatively small as in this example, the approximation may only be adequate. Next we develop a simulation technique, apply it to these data, and compare our results. In general, the small sample method we develop may be used for any size sample, small or large, and should be considered as more accurate than the corresponding large sample technique.

2.2 Simulating a difference under the null distribution

The ideas in this section were first introduced in a section of *OpenIntro Statistics* considering a malaria treatment. For the interested reader, this earlier section provides a more in-depth discussion.

Suppose the null hypothesis is true. Then the blood thinner has no impact on survival and the 13% difference was due to chance. In this case, we can simulate *null* differences that are due to chance using a *randomization technique*.[7] By randomly assigning "fake treatment" and "fake control" stickers to the patients' files, we could get a new grouping – one that is completely due to chance. The expected difference between the two proportions under this simulation is zero.

We run this simulation by taking 40 `treatment_fake` and 50 `control_fake` labels and randomly assigning them to the patients. The label counts of 40 and 50 correspond to the number of treatment and control assignments in the actual study. We use a computer program to randomly assign these labels to the patients, and we organize the simulation results into Figure 5.

	Survived	Died	Total
`control_fake`	15	35	50
`treatment_fake`	10	30	40
Total	25	65	90

Figure 5: Simulated results for the CPR study under the null hypothesis. The labels were randomly assigned and are independent of the outcome of the patient.

GUIDED PRACTICE 0.16

G What is the difference in survival rates between the two fake groups in Figure 5? How does this compare to the observed 13% in the real groups?[8]

The difference computed in Guided Practice 16 represents a draw from the null distribution of the sample differences. Next we generate many more simulated experiments to build up the null distribution, much like we did in Section 1.2 to build a null distribution for a one sample proportion.

Caution: Simulation in the two proportion case requires that the null difference is zero

The technique described here to simulate a difference from the null distribution relies on an important condition in the null hypothesis: there is no connection between the two variables considered. In some special cases, the null difference might not be zero, and more advanced methods (or a large sample approximation, if appropriate) would be necessary.

2.3 Null distribution for the difference in two proportions

We build up an approximation to the null distribution by repeatedly creating tables like the one shown in Figure 5 and computing the sample differences. The null distribution from 10,000 simulations is shown in Figure 6.

[7] The test procedure we employ in this section is formally called a **permutation test**.

[8] The difference is $\hat{p}_{t,fake} - \hat{p}_{c,fake} = \frac{10}{40} - \frac{15}{50} = -0.05$, which is closer to the null value $p_0 = 0$ than what we observed.

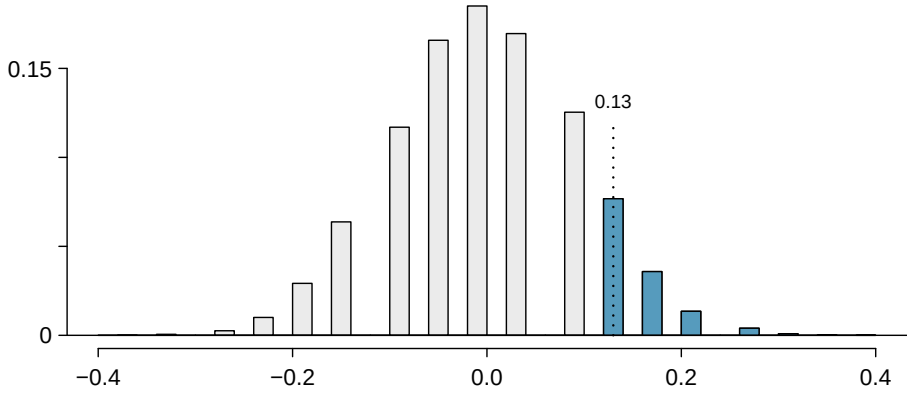

Figure 6: An approximation of the null distribution of the point estimate, $\hat{p}_t - \hat{p}_c$. The p-value is twice the right tail area.

EXAMPLE 0.17

Compare Figures 4 and 6. How are they similar? How are they different?

(E)

The shapes are similar, but the simulated results show that the continuous approximation of the normal distribution is not very good. We might wonder, how close are the p-values?

GUIDED PRACTICE 0.18

(G) The right tail area is about 0.13. (It is only a coincidence that we also have $\hat{p}_t - \hat{p}_c = 0.13$.) The p-value is computed by doubling the right tail area: 0.26. How does this value compare with the large sample approximation for the p-value?[9]

In general, small sample methods produce more accurate results since they rely on fewer assumptions. However, they often require some extra work or simulations. For this reason, many statisticians use small sample methods only when conditions for large sample methods are not satisfied.

2.4 Randomization for two-way tables and chi-square

Randomization methods may also be used for the contingency tables. In short, we create a randomized contingency table, then compute a chi-square test statistic X^2_{sim}. We repeat this many times using a computer, and then we examine the distribution of these simulated test statistics. This randomization approach is valid for any sized sample, and it will be more accurate for cases where one or more expected bin counts do not meet the minimum threshold of 5. When the minimum threshold is met, the simulated null distribution will very closely resemble the chi-square distribution. As before, we use the upper tail of the null distribution to calculate the p-value.

[9] The approximation in this case is fairly poor (p-values: 0.174 vs. 0.26), though we come to the same conclusion. The data do not provide convincing evidence showing the blood thinner helps or hurts patients.