FIT2086 Lecture 5 Notes
Hypothesis Testing

Dr. Daniel F. Schmidt*

September 9, 2020

Hypothesis testing is the process of deciding whether a particular hypothesis regarding reality is supported by the data we have obtained. In statistical parlance, a hypothesis is usually expressed in terms of a parametric probability distribution; in particular, we are often asking whether the parameters of a probability distribution are equal to a specific value, or whether the parameters of two different distributions are equivalent. Surprisingly, the majority of scientific questions can be expressed in these terms. The framework we will be using involves nominating one model of reality as a null hypothesis, and using the data to see whether there is evidence to suggest that reality is incompatible with this null hypothesis. More formally, we say we are testing

$$\begin{aligned} H_0 &: \text{Null hypothesis} \\ &\text{vs} \\ H_A &: \text{Alternative hypothesis} \end{aligned}$$

using our observed sample $\mathbf{y} = (y_1, \ldots, y_n)$. In this framework we take the null hypothesis to be our default position, and then ask how much evidence the data we have observed carries *against* this null hypothesis. We can never prove the null hypothesis to be true in this setting – only gather sufficient evidence to *disprove* it. For example, imagine we are modelling our population using a normal distribution; this has a mean parameter μ and a variance parameter σ^2. A standard "hypothesis" we might want to test is:

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ &\text{vs} \\ H_A &: \mu \neq \mu_0 \end{aligned}$$

In this case our null hypothesis is that μ is equal to μ_0 at the population level; our alternative hypothesis is that μ is not equal to μ_0 at the population level. To test the assertion that $\mu = \mu_0$, we obtain a sample of data $\mathbf{y}$ from our population, and then ask: "is there sufficient evidence in the data to dismiss the hypothesis that μ is equal to some fixed value μ_0?"

The obvious question is then to ask how we quantify evidence against the null? The approach we use, which is routinely used in industry and research environments, is to: (i) assume that the null hypothesis is true, i.e., assume that the population follows the null hypothesis; (ii) calculate how likely our sample would be to arise just by chance under this assumption; that is, ask, what is the probability of seeing the sample $\mathbf{y}$ we have observed just by chance, if the population followed the null

hypothesis. The smaller this probability, the more incompatible our sample is with our null hypothesis, and therefore the stronger the evidence against our null hypothesis.

Example 1: Testing Ohm's Law

Ohm's law is a famous formula, named in honour of Georg Ohm, that relates the current I, voltage V and resistance R of a circuit via the relationship

$$V = IR.$$

We might want to perform a physics experiment to test the accuracy of this law. Given a specified resistance and current we could measure the resulting voltage in a circuit and compare this with the prediction made by Ohm's law. Imagine that we choose $R = 100\Omega$ and $I = 0.01A$; Ohm's law would then predict that $V = 100 \cdot 0.01 = 1V$. Further, imagine that we went to a laboratory, selected a 100Ω resistor, set the current to 0.01 amps, and used a voltmeter to obtain a voltage reading of 0.956 volts.

Ohm's law predicts $1V$, but of course, due to measurement error and component manufacturing tolerances (up to 20% variation is plausible for cheap resistors) we would not actually expect to see *exactly* $1V$. So the question now becomes: is the difference between our measured voltage and the voltage predicted by Ohm's law large enough to suggest that Ohm's law does not hold? How much evidence against Ohm's law does this difference suggest, and how do we formally quantify this evidence? □

1 Testing a normal distribution with known variance

We start with the most basic hypothesis testing setting which can be used to demonstrate the relevant ideas. Assume our population is normally distributed with an unknown mean μ and known variance σ^2. Then, given a sample $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ we want to test

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ &\text{vs} \\ H_A &: \mu \neq \mu_0 \end{aligned}$$

That is, we want to test the hypothesis that the unknown population mean is equal to some specified value μ_0. As we know, the ML estimate $\hat{\mu} \equiv \bar{Y}$ will not be equal to μ_0 for any random sample, due to the variability of sampling, and the inherent randomness in the sample, even if $\mu = \mu_0$ at the *population* level. So instead, what we ask is: if $\mu = \mu_0$ at the population level, how unlikely would it be to see our estimate $\hat{\mu}$ just by chance? To answer this, we make use of the sampling distribution of $\hat{\mu}$ (see Lecture 3) *under the null hypothesis*. As a quick refresher, this is the distribution of the estimate $\hat{\mu}$ that we would see if we repeated drew samples of size n from our population, and assumed that our population followed our null hypothesis (i.e., $\mu = \mu_0$ at the population level). If our null hypothesis was true, then our sample follows

$$Y_1, \ldots, Y_n \sim N(\mu_0, \sigma^2), \tag{1}$$

i.e., if our null hypothesis was true, the population is normally distributed with a mean of μ_0 and variance σ^2. Our maximum likelihood estimate of the population mean is the sample mean

$$\hat{\mu} \equiv \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Recalling that the distribution of the sample mean $\bar{Y}$ when the population is normally distributed (see Lecture 3), we know that the sampling distribution of $\hat{\mu}$ under our null hypothesis (1) is

$$\hat{\mu} \sim N\left(\mu_0, \frac{\sigma^2}{n}\right). \tag{2}$$

This is the distribution of the sample mean $\hat{\mu}$ if we repeatedly took samples of size n from our population, and our population followed the null hypothesis (i.e., Equation 1). Given this estimate $\hat{\mu}$, we can calculate the difference between our sample mean estimate $\hat{\mu}$ and the hypothesised population mean μ_0 – the larger this difference, the more at odds with our null distribution this sample is. We can quantify exactly how much at odds with our null distribution a sample is by determining how likely it would be see a difference from μ_0 of size $\hat{\mu} - \mu_0$ just by chance, if our population followed the null hypothesis (1). To do this, we note that if the null hypothesis is true, then the sampling distribution of $\hat{\mu}$ follows (2). Recalling that any normally distributed random variable can always be standardised to an RV that follows a unit normal $N(0, 1)$ distribution (see Lecture 2), we can calculate the z-score for our estimate $\hat{\mu}$ under the assumption that the population follows the null distribution (we can also say under the assumption that "the null is true"):

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{\sigma/\sqrt{n}}. \tag{3}$$

The quantity $z_{\hat{\mu}}$ represents a standardised difference between the null μ_0 and our sample estimate $\hat{\mu}$. Recalling that $\sigma/\sqrt{n}$ is the standard error for the estimate $\hat{\mu}$, it tells us how *many* standard errors, $\sigma/\sqrt{n}$, the estimate $\hat{\mu}$ is away from the null $\mu = \mu_0$. If the population follows the null hypothesis, then the z-score satisfies

$$z_{\hat{\mu}} \sim N(0, 1);$$

that is, $z_{\hat{\mu}}$ would follow the standard unit normal distribution. Then, the probability of seeing a standardised difference from μ_0 of $|z_{\hat{\mu}}|$ or greater, in either direction (i.e., negative or positive), would be

$$\begin{aligned} p &= 1 - \mathbb{P}(-|z_{\hat{\mu}}| < Z < |z_{\hat{\mu}}|) \\ &= \mathbb{P}(Z < -|z_{\hat{\mu}}|) + \mathbb{P}(Z > |z_{\hat{\mu}}|) \end{aligned}$$

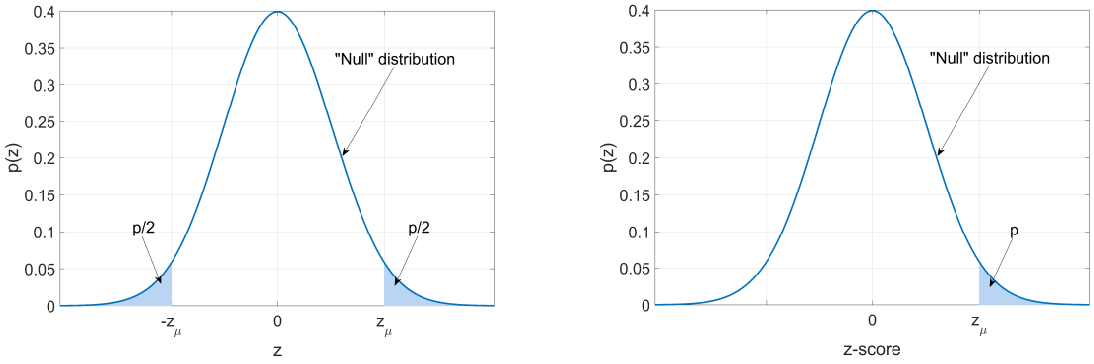
where $Z \sim N(0, 1)$. We ignore the sign, as a big difference in either direction (i.e., big positive or big negative difference from μ_0) is equally strong evidence against the hypothesis that $\mu = \mu_0$. Using the symmetry of the normal distribution, we can re-write the above probability statement as

$$p = 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|).$$

The quantity p is called a "p-value", and can be calculated in R using

```
pval = 2*pnorm(-abs(z)).
```

In this particular setting, the p-value is the probability, *if our population followed the null distribution* and $\mu_0 = \mu$, that a random sample drawn from our population would have a sample mean that was at least $|\mu_0 - \hat{\mu}|$ far away from μ_0; or equivalently, a standardised difference of $|z_{\hat{\mu}}|$ *or greater* in either direction (negative or positive), just by chance. It measures how compatible our data is with the proposed null hypothesis (1). Under the null, the mean of any data sample is μ_0, and from the properties of the sample mean we know that $\mathbb{E}[\hat{\mu}] = \mu_0$. Therefore, we expect that if our null hypothesis is true, our data sample should have a sample mean close to μ_0, relative to the amount of variability in the population. A small standardised difference between the sample mean and μ_0 is therefore suggestive that the data is compatible with the proposed null hypothesis, while a large standardised difference is suggestive that the data is incompatible with our proposed null hypothesis.



(a) Two-sided test

(b) One-sided test

Figure 1: Computation of p-values for the two-sided test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$, and for the one-sided test $H_0 : \mu \leq \mu_0$ vs $H_A : \mu > \mu_0$.

1.0.1 Grading p-values and "significance"

The larger the standardised difference from the null, $|z_{\hat{\mu}}|$, the smaller the resulting p-value will be. The smaller the p-value, the more unlikely or improbable the sample we have obtained would be if our population followed the null distribution; therefore, the smaller the p-value the stronger the evidence against the null distribution being true. But how much evidence does a particular p-value provide? A conventional threshold for "significance" is $p = 0.05$. If a p-value is smaller than 0.05 it is sometimes deemed to be indicative of a difference that is "statistically significant".

This unit recommends strongly against rote use of arbitrary thresholds such as 0.05. To understand why, consider what it means if the p-value you obtain after performing the above statistical test is 0.05. This means that there is a 5% chance that after drawing a random sample from our population, we would see as big, or bigger, a standardised difference as the one we observed, even if the null was true, just by chance. In other words, one in twenty random samples we would draw would have a sample mean that differed from the null hypothesis mean μ_0 by this much just by chance. A one in twenty event is not particularly rare – so in fact a p-value of 0.05 does not provide particularly strong evidence against the null hypothesis being true. Further, there is nothing special about the value 0.05; one could just have easily chosen the threshold to be 0.04, or 0.01. The choice of 0.05 is purely convention. Rather than utilise arbitrary thresholds to label findings to be be "significant", we recommend interpreting the p-value purely as a measure of incompatibility of our data with our null hypothesis. In this sense it measures the amount of evidence the data (for example, some experiment), offers *against* the null hypothesis. Our choice of whether to reject the null hypothesis should then be based on how strong we feel the evidence against the null needs to be, after weighing in factors such as the result of an incorrect decision (for example, incorrectly deciding that a drug with potentially harm-side effects is worthwhile using to treat cancer). *Informally*, we can grade the p-value into some rough categories:

- for $p > 0.1$ we have very weak/no evidence against the null;
- for $0.05 < p < 0.1$ we have marginal/borderline evidence against the null;
- for $0.01 < p < 0.05$ we have moderate evidence against the null;
- for $p < 0.01$ we have strong evidence against the null.

The quantity that we use to compute our p-value – in this case the z-score $z_{\hat{\mu}}$ – is generally called our test statistic. Figure 1(a) shows the idea underlying a p-value. In this plot we see the probability density function for our test statistic under the null hypothesis; the shaded areas represent the probability for $(Z < -|z_{\hat{\mu}}|)$ and $(Z > |z_{\hat{\mu}}|)$ respectively. Given the symmetry, the overall probability of $Z < -|z_{\hat{\mu}}|$ or $Z > |z_{\hat{\mu}}|$ is equal to twice the probability in either of those tails; for this reason, this type of test ($\mu = \mu_0$ vs $\mu \neq \mu_0$) is called a two-sided test. See Studio 5 for more details. The difficulty in constructing tests of this type for a particular statistical model usually lies in finding an appropriate test statistic.

1.1 One-sided tests

The test of $\mu = \mu_0$ versus $\mu \neq \mu_0$ is called a two-tailed test. This is because we treat either large negative or positive deviations in the sample mean from μ_0 as strong evidence against the null hypothesis. Imagine instead we assume our population is normally distributed with unknown mean and known variance and want to test

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ &\text{vs} \\ H_A &: \mu > \mu_0 \end{aligned}$$

That is, we are interested in testing whether the unknown population mean is *less than* some value μ_0. This is similar to the above problem, but differs in that now only sample means greater than μ_0 will offer any real evidence against our null position, which is that the population mean μ is less than or equal to μ_0. Therefore, this type of test is called a one-sided test. Once again, for this problem, a suitable test statistic is the standardised difference of the sample mean $\hat{\mu}$ from the hypothesised upper bound of μ_0, i.e.,

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\sigma/\sqrt{n})}.$$

However, unlike in our previous two-sided test, where either large negative or positive deviations suggested the sample was incompatible with the null, now only large positive deviations will be unlikely if our null is true. This is because our null now says that $\mu \leq \mu_0$, so sample means smaller than μ_0 are not in conflict with this statement. Therefore, the p-value is the probability of seeing a z-score *at least as large*, or larger, than our observed standardised difference $z_{\hat{\mu}}$, i.e.,

$$p = \mathbb{P}(Z > z_{\hat{\mu}}) = 1 - \mathbb{P}(Z < z_{\hat{\mu}})$$

where $Z \sim N(0, 1)$. Note we do not take the absolute value of $z_{\hat{\mu}}$ in this formula. Figure 1(b) demonstrates the idea behind the one-sided test; if we compare this to Figure 1(a), we see that in this case we are only interested in large positive deviations; therefore we calculate our p-value as the probability that our test-statistic would exceed the observed standardised difference of our sample $z_{\hat{\mu}}$ just by chance; that is, the p-value is the probability of seeing a deviation from μ_0 of size $\hat{\mu} - \mu_0$ just by chance, if the population followed the null hypothesis. Similary, we can test

$$\begin{aligned} H_0 &: \mu \geq \mu_0 \\ &\text{vs} \\ H_A &: \mu < \mu_0 \end{aligned}$$

that is, test whether the unknown population mean μ is greater than some value μ_0. This time, differences between the sample mean $\hat{\mu}$ and μ_0 that are *large and negative* are treated as evidence against the null; the greater the difference becomes in a negative direction, the more unlikely the

sample would be to arise from population if the null hypothesis was true and $\mu \geq \mu_0$. Therefore, the p-value is the probability of seeing a difference as small as $\hat{\mu} - \mu_0$, or smaller, just by chance if the null hypothesis were true, i.e.,

$$p = \mathbb{P}(Z < z_{\hat{\mu}})$$

where $Z \sim N(0, 1)$.

1.2 Summary: testing normal populations with known variance

We now summarise the above two and one-sided tests. First, we assume that the population follows a normal distribution with unknown mean and known variance σ^2. Then to test the inequality of μ:

1. First we calculate the ML estimate of the mean (equivalent to the sample mean):

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

2. Then, we calculate the test-statistic (or z-score) which is the standardised difference of the sample mean $\hat{\mu}$ from the null distribution reference point μ_0:

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\sigma/\sqrt{n})}.$$

3. Finally, we can calculate our p-value using:

$$p = \begin{cases} 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|)) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(Z < z_{\hat{\mu}}) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases},$$

   where $Z \sim N(0, 1)$

We can then assess how much evidence our data provides against our chosen null distribution by assessing the p-value as discussed above.

Example 2: Testing mean with known variance

Let us revisit Example 1 concerning the testing of Ohm's law. Given our experimental setting, with $R = 100\Omega$ and $I = 0.01A$ the voltage predicted by Ohm's law is $V = 1V$. If we assume that a normal distribution is suitable to model the experimental error then we can formulate our experiment as testing the hypothesis

$$\begin{aligned} H_0 &: \mu = 1 \\ &\text{vs} \\ H_A &: \mu \neq 1 \end{aligned}$$

where μ is the voltage (population mean voltage over all experiments) across our circuit. After factoring component tolerances (around 20%) and measurement accuracy specifications (around 5%) we might estimate the error to be $+/- 0.25V$. In terms of a normal distribution this can be well represented by setting $\sigma = 0.125$ (recall that 95% of values fall within $+/- 1.96\sigma$ of the mean). Imagine that we decide to repeat our experiment $n = 6$ times, and we obtain the following readings of voltage:

$$\mathbf{y} = (0.956, 1.364, 1.103, 1.172, 0.868, 0.966)V.$$

Using this data, and the procedure in Section 1.2, let us test to see whether Ohm's law appears to hold. We first compute the sample mean

$$\hat{\mu} = 1.0715.$$

We can then form our z-statistic for the difference of the sample mean from the predicted voltage $\mu_0 = 1$ as:

$$z_{\hat{\mu}} = \frac{1.0715 - 1}{(0.125/\sqrt{6})} \approx 1.401,$$

which yields a p-value of

$$\begin{aligned} 1 - \mathbb{P}(-z_{\hat{\mu}} < Z < z_{\hat{\mu}}) &= \texttt{2*pnorm(-abs(1.401))}, \\ &= 0.1612. \end{aligned}$$

So, what do we make of this p-value? We can interpret it in the following way:

> *"If the null was true, i.e., the voltage for this experiment is predicted by Ohm's law, then the chance of observing a sample with as an extreme, or more extreme, difference from the null as the one that we saw would be around 1/6.2."*

That is, we expect, if Ohm's law is true, about 1 in 6 experiments of $n = 6$ readings would yield a sample mean $\hat{\mu}$ that differed from $\mu_0 = 1$ by $0.0715V$ or greater, in either direction, just by chance. It is evident that 1 in 6 events are pretty common, so this offers very weak evidence against the null. We can conclude that it appears our experimental data is not incompatible with Ohm's law – at least for these particular values of the experimental setup. □

1.3 Testing a normal distribution with unknown variance

Now let us now relax our assumptions and assume that our population is normally distributed with unknown mean and unknown variance. This is more realistic than the previous situation in which we assumed the variance was known. We now want to test inequality of the mean population μ – either a one-sided or two-sided test. As we do not know the population variance we must estimate it from our sample; we can use the unbiased estimate of variance

$$\hat{\sigma}^2 = \left(\frac{1}{n-1}\right) \sum_{i=1}^{n} (y_i - \hat{\mu})^2.$$

If we calculate the standardised difference of $\hat{\mu}$ from our reference point μ_0 using (3), with our estimate $\hat{\sigma}$ in place of the unknown population standard deviation σ, then we have a t-score (see Lecture 4)

$$t_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\hat{\sigma}/\sqrt{n})}.$$

If the null hypothesis is true, then our t-score will follow

$$t_{\hat{\mu}} \sim \mathrm{T}(n-1),$$

where $\mathrm{T}(d)$ denotes a standard Student-t distribution with d degrees-of-freedom. Due to the symmetry and self-similarity of the t distribution, we can apply the same arguments we used previously to calculate the one- and two-sided p-values in the case that the variance was known, and calculate the

p-value of our test using

$$p = \begin{cases} 2\,\mathbb{P}(T < -|t_{\hat{\mu}}|)) & \text{if } H_0 : \mu = \mu_0 \text{ vs } H_A : \mu \neq \mu_0 \\ 1 - \mathbb{P}(T < t_{\hat{\mu}}) & \text{if } H_0 : \mu \leq \mu_0 \text{ vs } H_A : \mu > \mu_0 \\ \mathbb{P}(T < t_{\hat{\mu}}) & \text{if } H_0 : \mu \geq \mu_0 \text{ vs } H_A : \mu < \mu_0 \end{cases},$$

where $T \sim \mathrm{T}(n-1)$.

1.4 Testing the difference of means with known variances

As previously discussed in Lecture 4, one of the most important estimates we are often interested in is the difference in population means between two populations. Recall our example: imagine we have a cohort of people in a medical trial for a weight-loss drug. At the start of the trial, the weights of all participants are measured and recorded; call this sample $\mathbf{y}_A$, and assume it has an unknown population mean of μ_A. The participants are then administered a weight-loss drug for six months, and at the end of the trial period, the participants weights are remeasured; call this sample $\mathbf{y}_B$, with population mean μ_B. To see if the drug had any real effect on weight-loss we can try to estimate the population mean difference in the weights pre- and post-trial, i.e., $\mu_A - \mu_B$. If there is no difference at a population level, $\mu_A = \mu_B \Rightarrow \mu_A - \mu_B = 0$.

In Lecture 4 we examined the derivation of confidence intervals for the difference of population means to analyse such data. Now we will look at using hypothesis testing to formally test the hypothesis that population means are the same, i.e., to test

$$\begin{aligned} H_0 &: \mu_A = \mu_B \\ &\text{vs} \\ H_A &: \mu_A \neq \mu_B \end{aligned}$$

in the case that the population means μ_A, μ_B are *unknown* and the population variances σ_A^2, σ_B^2 are known. The key idea is to note that if two populations have the same mean, then their difference will have a mean of zero at the population level – therefore, large (positive or negative) differences between the means of the two samples can be viewed as evidence *against* the null distribution that $\mu_A = \mu_B$. The first step to derive a p-value for testing the above hypothesis is to calculate the sample means of the two samples; call these $\hat{\mu}_A$ and $\hat{\mu}_B$, respectively. Then, if the populations follow the null hypothesis and $\mu_A = \mu_B$, the difference will follows

$$\hat{\mu}_A - \hat{\mu}_B \sim N\left(0, \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}\right)$$

where n_A and n_B are the sizes of the two samples. We take the z-score for the difference in means as our test statistic:

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}.$$

The p-value is then the probability that, under the assumption that the two populations have the same mean (the null hypothesis), the difference in sample means between a sample drawn from both of these populations would be as great as $|z_{(\hat{\mu}_A - \hat{\mu}_B)}|$, or greater, in *either direction*, i.e.,

$$p = 2\,\mathbb{P}\left(Z < -|z_{(\hat{\mu}_A - \hat{\mu}_B)}|\right),$$

which is the probability of observing a (standardised) difference between the sample means of magnitude $|z_{(\hat{\mu}_A - \hat{\mu}_B)}|$ or greater in either direction, if the *null was true*. Thus the larger the difference

between the means of the two samples, the greater the evidence against the null hypothesis that $\mu_A = \mu_B$. For testing against the one-sided alternative $H_0 : \mu_A \geq \mu_B$ vs $H_A : \mu_A < \mu_B$ we can compute

$$p = \mathbb{P}\left(Z < z_{(\hat{\mu}_A - \hat{\mu}_B)}\right)$$

which can also be used to test $\mu_A \leq \mu_B$ by noting this is the same as $\mu_B \geq \mu_A$. To test against the one-sided alternative $H_0 : \mu_A \leq \mu_B$ vs $H_A : \mu_A > \mu_B$ we can essentially do the same procedure as above but reverse the role of sample "A" and sample "B", i.e., we can use

$$p = \mathbb{P}\left(Z < z_{(\hat{\mu}_B - \hat{\mu}_A)}\right)$$

where again, $Z \sim N(0, 1)$.

Example 3: Testing differences of normal means, known variances

Imagine we have measured BMI on a sample of women aged 20-34 from the Pima ethnic group who do not have diabetes and a sample of BMI measurements from Pima ethnic women aged 20-34 who do have diabetes. Let us assume that the population standard deviations of BMI for Pima ethnic people with and without diabetes have been estimated from another, larger study, and are known to be $\sigma_n = 6.79$ and $\sigma_d = 6.69$ for non-diabetics and diabetics, respectively. The two samples are:

$$\begin{aligned} \mathbf{y}_n &= (46.8, 27.8, 32.5, 39.5, 32.8, 31.0, 26.2, 20.8) \\ \mathbf{y}_d &= (33.6, 23.3, 43.1, 31.0, 30.5, 38.0, 30.1, 25.8) \end{aligned}$$

Researchers want to know if there is a difference, at the population level, in BMIs between Pima ethnic women aged 20-34 with and without diabetes, i.e., we want to test:

$$\begin{aligned} H_0 &: \mu_n = \mu_d \\ &\text{vs} \\ H_A &: \mu_n \neq \mu_d. \end{aligned}$$

The two sample means are $\hat{\mu}_n = 32.175$ and $\hat{\mu}_d = 31.925$, respectively. The z-score for the difference, $32.175 - 31.925 = 0.25 kg/m^2$ is

$$z_{\hat{\mu}_n - \hat{\mu}_d} = \frac{0.25}{\sqrt{\frac{6.79^2}{8} + \frac{6.69^2}{8}}} \approx 0.074,$$

which leads to a p-value of

$$2\,\mathbb{P}(Z < -0.074) \approx 0.94$$

The p-value says that if there was no difference at the population level in BMI between Pima ethnic women aged 20-34 with and without diabetes, we would expect to see a difference as large as, or larger than, the one we have observed 94% of the time if we drew two samples of size $n = 8$ from these populations. So the p-value suggests that we have no evidence to reject the null that diabetics and non-diabetics in the female Pima population, aged 20-34, have the same average body mass index. □

1.5 Testing the difference of means with unknown variances

More generally, we do not know the values of the population variances. Unfortunately, when we relax this assumption things become trickier. Let us assume that $\sigma_A^2 \neq \sigma_B^2$, and both are unknown.

Unfortunately, in this situation there exists no simple exact test statistic; however, in a similar fashion to the confidence interval for the difference (see Lecture 4) of the population means with unknown variances, we can obtain an approximate p-value by using a variation of the test for difference of means when the variances are known and replacing the unknown population variances with their estimates. This approach is not exact, but for moderate sample sizes n_A and n_B, it yields p-values that are close to the more exact procedures, while remaining simple to implement. Our procedure is:

1. First, calculate the estimates of the means, $\hat{\mu}_A$ and $\hat{\mu}_B$, and variances, $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$, of the two samples, using the sample mean and unbiased estimate of variance, respectively.
2. Using these estimates, we can construct an (approximate) z-score using
$$z_{(\hat{\mu}_A - \hat{\mu}_B)} = \frac{\hat{\mu}_A - \hat{\mu}_B}{\sqrt{\frac{\hat{\sigma}_A^2}{n_A} + \frac{\hat{\sigma}_B^2}{n_B}}}$$
3. From this test-statistic, we can then find approximate p-values for the two- and one-sided tests using
$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\mu}_A - \hat{\mu}_B)}|) & \text{if } H_0 : \mu_A = \mu_B \text{ vs } H_A : \mu_A \neq \mu_B \\ 1 - \mathbb{P}(Z < z_{(\hat{\mu}_A - \hat{\mu}_B)}) & \text{if } H_0 : \mu_A \leq \mu_B \text{ vs } H_A : \mu_A > \mu_B \\ \mathbb{P}(Z < z_{(\hat{\mu}_A - \hat{\mu}_B)}) & \text{if } H_0 : \mu_A \geq \mu_B \text{ vs } H_A : \mu_A < \mu_B \end{cases}.$$
where $Z \sim N(0, 1)$.

As in the case of the approximate confidence interval for the difference of normal means with unknown variances discussed in Lecture 4, we acknowledge that the above procedure is not exact. However, just as with the approximate confidence intervals, the approximation is quite reasonable for even moderate samples sizes n_A and n_B, and it is frequently used in practice.

Example 4: Testing differences of normal means, unknown variances

Let us now imagine that we were not told the population standard deviations, and instead needed to estimate them from our samples. Let us assume that the population variances of BMI in diabetics and non-diabetics are different, and use the approximate method for testing differences of means. We now need to estimate the population variances from our samples, using our unbiased estimates of variance from our samples; these are

$$\hat{\sigma}_n^2 = 64.80, \quad \text{and } \hat{\sigma}_d^2 = 40.38$$

respectively. We can now compute the approximate z-score using these estimates:

$$z_{\hat{\mu}_n - \hat{\mu}_d} = \frac{0.25}{\sqrt{\frac{64.8}{8} + \frac{40.38}{8}}} \approx 0.068,$$

Using this quantities results in a p-value of $2\,\mathbb{P}(Z < -0.068) \approx 0.945$, so our conclusions do not change. □

2 Testing a Bernoulli Population

A particularly important application of hypothesis testing is in conjunction with binary data arising from Bernoulli distributions. Hypothesis tests of Bernoulli populations play an important role in both

industry and research settings as they can be used to test if rates of events occuring have changed due to some intervention, or if rates of failure of products do not exceed a certain level, and so on. For example, consider a production line making certain electronic components. If the manufacturer guarantees that the failure rate of components is less than some amount θ_0 – a requirement potentially needed to supply to customers like the military – then after obtaining a sample of the products in question, and observing the failure rate in that sample, a customer could test to see if the advertised failure rate was achieved.

2.1 Tests for a Single Bernoulli Population

Assume that our population is Bernoulli distributed, with a success rate of θ, i.e., $Y_1, \ldots, Y_n \sim \text{Be}(\theta)$. Then, given a sample, we want to test

$$\begin{aligned} H_0 &: \quad \theta = \theta_0 \\ &\text{vs} \\ H_A &: \quad \theta \neq \theta_0 \end{aligned}$$

or an appropriate one-sided test. To obtain an approximate p-value we can use the central limit theorem to derive an approximate null distribution for our test statistic. Recall that the maximum likelihood estimate for the success probability in a sample of Bernoulli data $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{m}{n}$$

where m is the number of successes in our sample. This is equivalent to the sample mean of our data $\mathbf{y}$, and so by using the central limit theorem (see Lecture 4), we know that

$$\hat{\theta} - \theta_0 \xrightarrow{d} N\left(0, \frac{\theta_0(1-\theta_0)}{n}\right)$$

as our sample size $n \to \infty$; in words, this says that as our sample gets larger, the difference between the estimated success probability and our reference θ_0 is approximately normally distributed with a mean of zero and variance $\theta_0(1-\theta_0)/n$, under the assumption that the population follows our null distribution. Using this result, we can obtain an approximate z-score to use as our test statistic

$$z_{\hat{\theta}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}},$$

and from this, we can calculate two or one-sided approximate p-values using

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{\hat{\theta}}|) & \text{if } H_0 : \theta = \theta_0 \text{ vs } H_A : \theta \neq \theta_0 \\ 1 - \mathbb{P}(Z < z_{\hat{\theta}}) & \text{if } H_0 : \theta \leq \theta_0 \text{ vs } H_A : \theta > \theta_0 \\ \mathbb{P}(Z < z_{\hat{\theta}}) & \text{if } H_0 : \theta \geq \theta_0 \text{ vs } H_A : \theta < \theta_0 \end{cases}.$$

where $Z \sim N(0,1)$.

2.2 Testing two Bernoulli populations

We now consider the problem of testing if two Bernoulli populations have the same probability of success. This type of test occurs frequently in practice. For example, we may have two groups of people, both with a particular disease. We administered one group with a drug designed to reduce the

negative effects of the disease, and the other with a placebo (a substance or treatment known to have no theraputic value). In this case, surviving for a period of time is a "success", and to see if our drug had any effect, we can compare the probability of success in the two groups. If they are different, then the drug has had an effect – otherwise it has had no effect. Given two samples $\mathbf{y}_A$ and $\mathbf{y}_B$ of binary data, we want to test

$$\begin{aligned} H_0 &: \theta_A = \theta_B \\ &\text{vs} \\ H_A &: \theta_A \neq \theta_B \end{aligned}$$

where θ_A and θ_B are the (unknown) population success probabilities of the two populations. Under the null hypothesis, we assume that $\theta_A = \theta_B = \theta$ (i.e., the two populations have the same unknown success probability θ), and so can use a pooled estimate of θ

$$\hat{\theta}_p = \frac{m_A + m_B}{n_A + n_B}$$

where m_A and m_B are the number of successes in the two samples, and n_A and n_B are the total number of trials. This estimate is more accurate than either $\hat{\theta}_A = m_A/n_A$ or $\hat{\theta}_B = m_B/n_B$ if the population *follows the null hypothesis* that $\theta_A = \theta_B = \theta$. Using $\hat{\theta}_p$ and the properties of the Bernoulli distribution we can approximate the variance of the estimates for the two samples using

$$\mathbb{V}\left[\hat{\theta}_A\right] \approx \frac{\hat{\theta}_p(1-\hat{\theta}_p)}{n_A} \text{ and } \mathbb{V}\left[\hat{\theta}_B\right] \approx \frac{\hat{\theta}_p(1-\hat{\theta}_p)}{n_B}$$

under the assumption that the population follows the null distribution. Then, using these variances and the central limit theorem we can define the test statistic

$$z_{(\hat{\theta}_A - \hat{\theta}_B)} = \frac{\hat{\theta}_A - \hat{\theta}_B}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)(1/n_A + 1/n_B)}}$$

which approximately follows an $N(0,1)$ distribution if the population follows the null hypothesis. From this we can then get approximate p-values using

$$p \approx \begin{cases} 2\,\mathbb{P}(Z < -|z_{(\hat{\theta}_A - \hat{\theta}_B)}|) & \text{if } H_0 : \theta_A = \theta_B \text{ vs } H_A : \theta_A \neq \theta_B \\ 1 - \mathbb{P}(Z < z_{(\hat{\theta}_A - \hat{\theta}_B)}) & \text{if } H_0 : \theta_A \leq \theta_B \text{ vs } H_A : \theta_A > \theta_B \\ \mathbb{P}(Z < z_{(\hat{\theta}_A - \hat{\theta}_B)}) & \text{if } H_0 : \theta_A \geq \theta_B \text{ vs } H_A : \theta_A < \theta_B \end{cases}.$$

In the case of testing Bernoulli populations, there exist more exact methods for computing p-values. These are more complex than our approximate procedures and make use of the fact that the counts m of successes are known exactly to follow a binomial distribution. Some of these are implemented in R; for example:

- `binom.test()` can be used to test a single Bernoulli sample;
- `prop.test()` can be used to test difference in Bernoulli samples.

The derivation of these more exact tests is beyond the scope of this unit, but we highlight these as an example of the vast range of procedures that exist for computing p-values for most statistical problems. The specific technical details of how a package or procedure computes the p-value are to a degree less important than having an understanding of the general concepts underlying hypothesis testing and the *meaning* of the p-value.

Example 5: Testing a Bernoulli population

Imagine that we have run a survey for a travel agent asking $n = 60$ people who recently travelled to Europe whether they prefer to holiday in France or Spain. We can nominate someone choosing France as a "success" (of course, it makes no difference to the conclusions which country we choose as a success). Imagine that $m = 37$ out of $n = 60$ people surveyed preferred France to Spain; we then have $\hat{\theta} = m/n = 37/60 \approx 0.6166$. So in our sample, 61% of people preferred France to Spain. A standard question our client might ask is: is there a real preference amongst people for one country over the other (which would imply that $\theta \neq 1/2$), or is this observed preference just due to random chance and there is no preference at a population level (which would imply that $\theta = 1/2$)? To provide an answer to a question like this, we use our approximate method for testing Bernoulli populations described in Section 2.1. First, we calculate our approximate z-score:

$$z_{\hat{\theta}} = \frac{(37/60) - 1/2}{\sqrt{(1/2)(1-1/2)/60}} \approx 1.807$$

which yields an approximate p-value of

$$2\,\mathbb{P}(Z < -1.807) = \texttt{2*pnorm(-1.807)} \approx 0.0707.$$

So, using this technique we can say that *if there was no preference* amongst people for Spain over France, or vice versa at a population level (i.e., if $\theta = 1/2$), then the chance of seeing 37 out of 60 people, or greater, preferring one of those countries over the other is around 7%, or 1 in 14. That is, we would expect around 1 in 14 times that we surveyed $n = 60$ people from this population we would see a preference of 37, or more, out of 60 people in favour of either of the two countries, *even if there was no preference* at the population level, just by random chance. This is not particularly unlikely, and therefore does not offer much evidence against the null distribution that there is no preference (i.e., that $\theta = 1/2$). We can also use R to compute the exact p-value in this situation with the `binom.test()` function; this yields a p-value of

$$\texttt{binom.test(x=37,n=60,p=0.5)} = 0.0924$$

which says that the chance of seeing a preference for one country over the other as strong as we have seen, or stronger, just by chance if there was no preference at a population level, is around 9%. This is even weaker evidence than suggested by our approximate test; from this, we would conclude that it is unlikely there is a preference in our group for either country over the other. □

3 Can we prove the null hypothesis?

A common misconception is that a large p-value proves the null hypothesis is true. This is not the case: in fact, the p-value represents evidence against the null only. That is, we can only gather evidence to disprove the null hypothesis, never in favour of the null hypothesis. Essentially, under this model of scientific enquiry, we look for counterexamples to try and falsify a proposed hypothesis. The more data we examine without finding a counter-example, the more compelling the hypothesis becomes – but we acknowledge that just because we have not yet found data incompatible with our hypothesis does not mean there does not exist any counter-examples.

For example, in the context of our Ohm's law experiment, we tried single experimental setting ($R = 100\Omega$, $I = 0.01A$) and found that our experimental data, consisting of $n = 6$ observations, was not incompatible with Ohm's law's prediction of the voltage for this setting. Of course, the fact that

such a small sample of $n = 6$ observations was not incompatible is not particularly compelling evidence in terms of *accepting* the null hypothesis. But what if we repeated the experiment, with a very large number of observations, say $n = 10^6$, and still obtained a large p-value. We have gathered a lot of data, and still found it to be not incompatible with Ohm's law. Shouldn't this be strong evidence to prove that Ohm's law is true?

The answer is still no. To see why, consider the following alternative law relating voltage to current and resistance:

$$V = \frac{R^2\sqrt{I}}{1000}. \tag{4}$$

For the settings of our experiment, $R = 100\Omega$ and $I = 0.01A$, this law (which is clearly a wrong model of reality!) also predicts the voltage to be $1V$, just as Ohm's law does. Therefore, if we experimentally measured the voltage for a large number of repetitions with these settings, the data would also not be incompatible with the law (4). However, if we varied the settings of our experiment, and took $R = 200\Omega$, Ohm's law would predict a voltage of $2V$, while law (4) would predict a voltage of $4V$. If we performed a number of measurements in our laboratory for these new settings, we would find the data would be incompatible with (4), yielding a small p-value, but not incompatible with Ohm's law. Therefore, we could move to dismiss law (4) as a model of reality, but we could still not accept Ohm's law because there could *conceivably* be an experimental setting for which it failed to be compatible with the experimental data. A prime example of this phenomenon is the situation of Newton's classical laws of physics, and the more modern theory of relativity. Initially it seemed that Newton's these laws provided predictions compatible with all experimental data. It was not until many years later, when the predictions made by these laws were compared against data collected on very large, or very fast objects, the experimental data began to show incompatibilities with Newton's laws, throwing doubt on their universal nature. The laws were not wrong, *per se*, in that they offer an excellent explanation for physical processes in specific settings, but they are not universally correct. This is why simply failing to finding data that is incompatible with a null hypothesis is insufficient to "prove" the null hypothesis correct. In the words of the famous Nobel prize winning physicist Richard Feynmann: "we can never be sure we are right, we can only ever be sure we are wrong".