FIT2086 Lecture 5 Summary
Hypothesis Testing

Dr. Daniel F. Schmidt

September 20, 2017

1 Part I: Hypothesis Testing

Hypothesis tests. 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.

Testing normal mean, known variance. We start with the most basic hypothesis testing setting, which can be used to demonstrate the basic 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

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

As we know, the ML estimate $\hat{\mu} \equiv \bar{Y}$ will not equal μ_0 for any random sample, due to the variability of sampling and 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 Lectures 3 and 4) 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

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

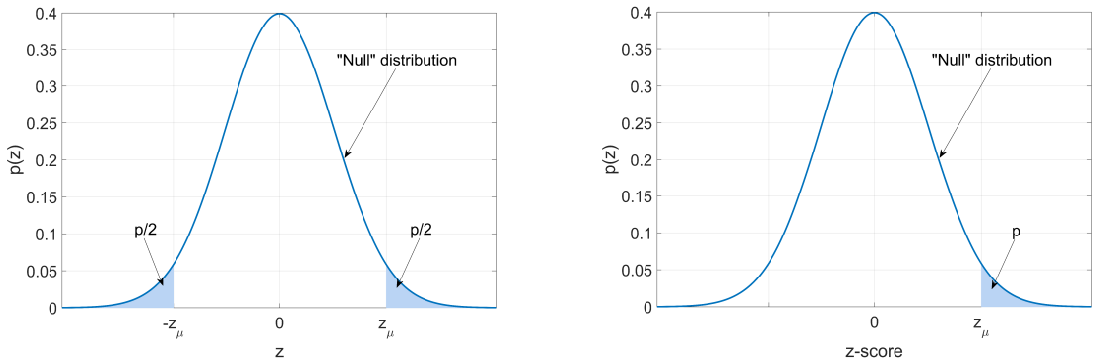
Recalling the distribution of the sample mean $\bar{Y}$ when the population is normally distributed (see Lectures 3 and 4), we that the sampling distribution of $\hat{\mu}$ under our null hypothesis (1) is

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

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. 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 normal random variables can always be standardised to an RV that follows a unit normal $N(0, 1)$ distribution, 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 "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,



(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 $H_A : \mu > \mu_0$.

$\sigma/\sqrt{n}$, 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 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 from our population would lead to a difference of $|\mu - \hat{\mu}|$, or equivalently, a standardised difference of $|z_{\hat{\mu}}|$ or greater in either direction (negative or positive), just by chance.

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. Informally, we can grade the p-value; for

- 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 of the 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 $(Z < -|z_\mu|)$ and $(Z > |z_\mu|)$ 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.

One-sided tests. The test of $\mu = \mu_0$ vs $\mu \neq \mu_0$ 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 absolute 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 $Z \sim N(0,1)$.

Summary: testing mean of normal population 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 $Z \sim N(0,1)$

Example: Testing mean with known variance. For US women aged between 20 to 34 years of age, the population body mass index (BMI) been estimated to have a mean of $26.8 kg/m^2$ and standard deviation of $4.5 kg/m^2$ *(Source: Center for Disease Control)*. Imagine we have measured BMI on a sample of women aged 20-34 from the Pima ethnic group who do not have diabetes:

$$\mathbf{y} = (46.8, 27.8, 32.5, 39.5, 32.8, 31.0, 26.2, 20.8)$$

Using this data, can we say whether women aged 20-34 in this Pima cohort have the same average BMI as the general US population? We can formulate this by testing the hypothesis that:

$$\begin{aligned} H_0 &: \mu = 26.8 \\ &\text{vs} \\ H_A &: \mu \neq 26.8 \end{aligned}$$

where μ is the population mean BMI of Pima women aged 20-34. The estimated mean $\hat{\mu}$ from our sample is

$$\hat{\mu} = 32.175$$

We can then form our z-statistic for the difference from $\mu_0 = 26.8$ as:

$$z_{\hat{\mu}} = \frac{32.175 - 26.8}{(4.5/\sqrt{8})} \approx 3.3784,$$

which yields a p-value of

$$\begin{aligned} 1 - \mathbb{P}(-z_{\hat{\mu}} < Z < z_{\hat{\mu}}) &= 2 * \texttt{pnorm}(-\texttt{abs}(3.3784)) \\ &= 7.29 \times 10^{-4}. \end{aligned}$$

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

> If the null was true, i.e., Pima ethnic women aged 20-34 have the same average BMI as the average US woman aged 20-34, 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 less than 1/1371.

So we can conclude that if the Pima ethnic women aged 20-34 had the same population BMI as the average US woman aged 20-34, it would be extremely unlikely to see the sample we have observed. This leads us to conclude that Pima ethnic women aged 20-34 do not have the same population average BMI as the "average US woman".

Testing normal mean, unknown variance. Now let us 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

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

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

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

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 (see Lecture 4). 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 by

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

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

Testing difference of means, known variances. One of the most important estimates we are often interested in is the difference in population means between two populations. As an example, imagine we have a cohort of people in a medical trial for a weight-loss drug. At the start of the trial, all the weights of all participant's are measured and recorded. Call this sample $\mathbf{x}$, and assume it has an unknown population mean of μ_x. The participants are then administered a weight-loss drug for 6 months, and at the end of the trial period, we re-measure the participant's weights; call this sample $\mathbf{y}$, with population mean μ_y. 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., $\mu_x - \mu_y$. If there is no difference at a population level, $\mu_x = \mu_y \Rightarrow \mu_x - \mu_y = 0$.

In Lecture 4 we examined 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., we can test

$$\begin{aligned} H_0 &: \mu_x = \mu_y \\ &\text{vs} \\ H_A &: \mu_x \neq \mu_y \end{aligned}$$

in the case that the population means μ_x, μ_y are unknown and the population variances σ_x^2, σ_y^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 – so large (positive or negative) differences between the means of the two samples can be viewed as evidence against the null distribution that $\mu_x = \mu_y$. We first estimate the sample means of the two samples:

$$\hat{\mu}_x = \frac{1}{n_x}\sum_{i=1}^{n_x} x_i, \quad \hat{\mu}_y = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$$

where n_x and n_y are the sizes of the two samples. Then, if the populations follow the null hypothesis and $\mu_x = \mu_y$, the difference will follows

$$\hat{\mu}_x - \hat{\mu}_y \sim N\left(0, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

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, assuming the if the two populations had the same mean (the null hypothesis), the difference in sample means between a sample from both of these populations would be as great as $|z_{(\hat{\mu}_x - \hat{\mu}_y)}|$, or greater, in either direction, i.e.,

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

which tells us the probability of observing a (standardised) difference between the sample means of $|z_{(\hat{\mu}_x - \hat{\mu}_y)}|$ 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 $\mu_x = \mu_y$. For testing against the one-sided alternative $H_0 : \mu_x \geq \mu_y$ vs $H_A : \mu_x < \mu_y$ we can compute

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

which can also be used to test $\mu_x > \mu_y$ by noting this is the same as $\mu_y < \mu_x$, i.e., to test against the one-sided alternative $H_0 : \mu_x \leq \mu_y$ vs $H_A : \mu_x > \mu_y$

$$p = \mathbb{P}\left(Z > z_{(\hat{\mu}_x - \hat{\mu}_y)}\right)$$

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

Testing difference of means, unknown equal variances. More generally, we do not know the values of the population variances. Unfortunately, when we relax this assumption things become trickier. If we assume that the variances are unknown but equal, i.e., $\sigma_x^2 = \sigma_y^2$, then we can derive a variation of the t-test. To do so, we first estimate the population variances for each sample individually

$$\hat{\sigma}_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \hat{\mu}_x)^2, \quad \hat{\sigma}_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \hat{\mu}_y)^2.$$

Then, the next step is to form a pooled estimate of σ^2:

$$\hat{\sigma}_p^2 = \frac{(n_x - 1)\hat{\sigma}_x^2 + (n_y - 1)\hat{\sigma}_y^2}{n_x + n_y - 2}$$

This estimate will be more accurate than either $\hat{\sigma}_x^2$ or $\hat{\sigma}_y^2$ if the population follows the null distribution. Then, we can form a t-score of the difference in means

$$t_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\hat{\sigma}_p^2(1/n_x + 1/n_y)}} \tag{4}$$

which follows a $\mathrm{T}(n_x + n_y - 2)$ distribution. Our p-value for testing $H_0 : \mu_x = \mu_y$ vs $H_A : \mu_x \neq \mu_y$ (i.e., the two-sided p-value) is then

$$p = 2\,\mathbb{P}\left(T < -|t_{(\hat{\mu}_x - \hat{\mu}_y)}|\right)$$

where $T \sim \mathrm{T}(n_x + n_y - 2)$. In R, we can compute this quantity in the following fashion: if `tdiff` is a variable containing our t-score (4) then the code `p = 2 * pt(-abs(zdiff),df=n_x+n_y-2)` will give us our p-value.

Testing difference of means, unknown and unequal variances. Let us relax our assumptions even further and assume that $\sigma_x^2 \neq \sigma_y^2$, and both are unknown. Unfortunately, in this situation there exists no exact test statistic; however, one way to get an approximate p-value is to use the test for difference of means when variances are known, but replace the unknown population variances with their estimates. This is not exact, but for moderate sample sizes n_x, n_y, gives p-values that are close to the more exact procedures. Our procedure is:

1. First, calculate the estimates of the means of each of the two samples:

$$\hat{\mu}_x = \frac{1}{n_x}\sum_{i=1}^{n_x} x_i, \quad \hat{\mu}_y = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$$

2. Next, we need to calculate the unbiased estimates of the population variances for each of the two samples:

$$\hat{\sigma}_x^2 = \frac{1}{n_x - 1}\sum_{i=1}^{n_x}(x_i - \hat{\mu}_x)^2, \quad \hat{\sigma}_y^2 = \frac{1}{n_y - 1}\sum_{i=1}^{n_y}(y_i - \hat{\mu}_y)^2.$$

3. Using these estimates, we can construct an (approximate) z-score using

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

   which approximately follows a $N(0, 1)$ distribution for large samples (using the central limit theorem, see Lecture 4).

4. From this test-statistic, we can then find approximate p-values for the two- and one-sided tests using

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

While this procedure is not exact, it does provide decent approximate p-values if the samples are moderate in size, and is frequently used in practice (see Ross, Chapter 8, Sec. 8.4.3 for details). More exact – but complicated procedures do exist – and some are implemented in R. The `t.test()` function examined in Studio 5 implements one of these improved approximations.

Example: Testing differences of normal means, known variances. Let us revisit our example; imagine now, that we have also obtained a sample of BMI measurements from Pima ethnic women aged 20-34 who do have diabetes. Further, 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:

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

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 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 different average body mass index.

Example: 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:

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

which leads to a p-value of $2\,\mathbb{P}(Z < -0.068) \approx 0.94$, so our conclusions do not change.

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 are important in industry and research, 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. 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 product and observing a failure rate in that sample, a customer could test to see if the advertised failure rate was achieved.

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

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

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 using the central limit theorem, 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 mean θ_0 and variance $\theta_0(1-\theta_0)/n$, under the assumption that the population follows our null distribution. Using this result, we can use an approximate z-score as our test statistic

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

and from this, we can then 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 $Z \sim N(0,1)$.

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 $\mathbf{x}$ and $\mathbf{y}$ of binary data, we want to test

$$\begin{aligned} H_0 &: \theta_x = \theta_y \\ &\text{vs} \\ H_A &: \theta_x \neq \theta_y \end{aligned}$$

where θ_x, θ_y are the (unknown) population success probabilities. Under the null hypothesis, we assume that $\theta_x = \theta_y = \theta$ (i.e., the two populations have the same unknown success probability θ) so use a pooled estimate of θ

$$\hat{\theta}_p = \frac{m_x + m_y}{n_x + n_y}$$

where m_x, m_y are the number of successes in the two samples, and n_x, n_y is the total number of trials. This estimate is more accurate than either $\hat{\theta}_x = m_x/n_x$ or $\hat{\theta}_y = m_y/n_y$ if the population follows the null hypothesis. We can use the test statistic

$$z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{\hat{\theta}_x - \hat{\theta}_y}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)(1/n_x + 1/n_y)}}$$

which approximately follows an $N(0,1)$ distribution if the population follows the null hypothesis; this follows from the central limit theorem. We can then get approximate p-values using

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

In the case of testing Bernoulli populations, there exist more exact methods for computing p-values (see for example, Ross Chapter 8). These are more complex than our approximate procedures and make use of the fact that the counts m of successes 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.

Example: Testing a Bernoulli population. Let us demonstrate an example of using these methods to test a Bernoulli population. Imagine 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 view someone choosing France as a "success" (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. We might want to ask: is there a real preference amongst people one country over the other, which would imply that $\theta \neq 1/2$, or is this observation just random chance and there is no preference at a population level which would imply that $\theta = 1/2$?

To test this, we use our approximate method described above. 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

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

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, 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 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 any strong evidence against the null distribution that there is no preference ($\theta = 1/2$). 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

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

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.

Interpreting p-values and tests. 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. This means that p-values in the $0.05 - -0.2$ range are sometimes inconclusive. For example, if we have observed a large difference between our sample and the null hypothesis, but the sample size is small and the p-value is in the "gray" zone of $0.05 - -0.2$, then our test is inconclusive.

This is because it is difficult to determine whether the reason we did not see stronger evidence is simply because our sample size was not large enough, or whether there really is no difference at a population level. Recall that when testing normal means we generally create a z-score which is standardised by the standard error ($\sigma/\sqrt{n}$) (or similar). The standard error represents how much our estimate might vary if we drew a new sample from our population and recomputed the estimate. We see that the smaller the sample size n, the larger the standard error and therefore the larger the value our estimated difference is divided by. Thus, for smaller sample sizes even large estimated differences often end up providing weak evidence against the null as the variability in the estimates is so large that it is difficult to nail down the real value of the difference at a population level. In this case, the only real way to resolve the issue is to redo the experiment but with a larger sample size.

2 Part II: Significance and Power

Decision making. So far we have computed p-values as evidence against the null, and assessed how strong this evidence is by how small the p-value is. What if we are asked to make a decision regarding our hypothesis? This type of problem happens frequently in practice – after performing a test, a drug company may want to know whether the drug is worth manufacturing, for example. We could use the evidence provided by the p-value to help guide us in making this decision; we could decide that if the evidence was sufficiently strong, we could reject the null hypothesis.

For example, we could say that if we see a sample that has probability α or less of arising by chance if the null distribution was true, then the evidence is strong enough to reject the null, where $\alpha < 1$ is the called the significance level. Formally, we reject the null hypothesis at a significance level of α if we reject the null when $p < \alpha$; that is, we say that the data is so unlikely under the null hypothesis that we can conclude that the population does not follow the null hypothesis. Sometimes you will see people

		Null hypothesis (H0) is	
		Valid (True)	**Invalid (False)**
Judgment:	**Reject**	Type I error *(False positive)*	Correct *(True positive)*
	Do not reject (accept)	Correct *(True negative)*	Type II error *(False negative)*

Figure 2: Table of possible outcomes that can occur when we choose to reject a null hypothesis if $p < \alpha$, where $\alpha < 1$ is the significance level.

(papers, articles) say that a result is "statistically significant" if the p-value is below some threshold. A common convention in the literature is to take $\alpha = 0.05$ – this is a throwback to the early days of statistical inference when R.A.Fisher, who pioneered much of this work, suggested that probability of the order of 1 in 20 was sufficiently "unlikely". However, while this is convention, there is nothing special about $\alpha = 0.05$ and we can take α to be any small value we like.

Type I errors. An obvious question: is the above procedure any good? One answer is to observe that if we choose to reject the null when $p < \alpha$, we are saying that we will reject the null hypothesis as being true if we observe a sample resulting in a test statistic that has probability of α, or less, of occuring just by chance, if the null was true. If we choose to reject the null hypothesis in the case that the null is true, we are erroneously rejecting the null. This is called a false discovery. The name makes sense if we think of the null as being "no discovery" – for example, that the difference between two populations is the same (e.g., a drug having no effect).

Erroneously rejecting the null distribution is then akin to making a discovery of an effect that is not really there, hence false discovery. In statistics language, a false discovery is called a Type I error. If we reject the null when $p < \alpha$, we will make a false discovery $100\alpha\%$ of the time. This is called controlling the Type I error rate at probability α.

Type II errors. If by rejecting the null only when $p < \alpha$ we are keeping the probability of making a false discovery to probability α, it follows that we could make the probability of a false discovery as small as we want by taking α as small as we want. However, to see why this is not a good idea, let us consider the alternative situation in which the null hypothesis is not true (for example, a drug having an effect on mortality). Now, if we fail to reject the null hypothesis in this situation we have also made an error, sometimes called a false negative, or a Type II error. The smaller we take our threshold α for significance, the more evidence we need against the null hypothesis before we are willing to reject it – which means it becomes less likely we will correctly reject the null in the case that the null is *not true*.

Thus, as we decrease α, we decrease the Type I error rate, (chance of making a false discovery if the null is true) but increase the Type II error rate (chance of incorrectly accepting the null, when the null is false), which we call β. In statistics, it is common to refer to the power of a test – this is just

$1-\beta$; this is the probability we will correctly reject the null hypothesis in the case that the null is false (make a true discovery). So, we can ensure that we are very unlikely to make a false discovery by keeping α small, but we must recognise that the smaller the α we take the less likely we are to make a true discovery if there really is a difference at the population level. Figure 2 summarises the possible outcomes we can have when making a decision by checking if a p-value is less than some level α.

Power. The power (probability of making a true discovery if there is a difference at population level) depends on the population. The larger the difference at the population level, the easier it becomes for a test to determine there is a difference. For illustration purposes, let us consider testing the means of two normal populations μ_x and μ_y from two samples with known variances σ_x^2 and σ_y^2. The test statistic we use is the z-score

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

and the difference at the population level

$$\mu_x - \mu_y$$

is sometimes called the effect size. Under our assumption

$$\hat{\mu}_x \sim N\left(\mu_x, \frac{\sigma^2}{n_x}\right), \quad \hat{\mu}_y \sim N\left(\mu_y, \frac{\sigma^2}{n_y}\right)$$

so that the z-score for the difference in sample means follows

$$z_{(\hat{\mu}_x-\hat{\mu}_y)} \sim N\left(\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}, 1\right)$$

under repeated sampling from our population. The quantity

$$\frac{\mu_x - \mu_y}{\sqrt{\sigma_x^2/n_x + \sigma_y^2/n_y}}$$

can be viewed as the *standardised* effect size – it is the difference in means relative to the variability in the estimate of the difference in means. The absolute value of this standardised effect size increases with (i) increasing population effect size $|\mu_x - \mu_y|$, (ii) increasing sample sizes n_x, n_y and (iii) decreasing population variances σ_x^2, σ_y^2. The greater the standardised effect size, the larger the value of the (absolute) z-score (on average) we would observe if we repeatedly sampled from our populutions and computed $z_{\hat{\mu}_x-\hat{\mu}_y}$. This makes sense – the larger the difference between the two populations, the larger the observed differences will be on average. Recall the p-value is given by

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

where $Z \sim N(0,1)$. Large values of $|z_{(\hat{\mu}_x-\hat{\mu}_y)}|$ are unlikely under the null ($\mu_x = \mu_y$), and they subsequently result in small p-values. This implies that the greater the standardised effect size, the smaller the average p-value we will obtain if we repeatededly sampled from our populuations, and therefore, the greater chance to correctly reject the null (and achieve a true positive!) Therefore, we can conclude that the power increases with increasing effect size. This makes sense – the larger the difference between the populations, the easier it is to identify that there is a difference. The smaller the difference, the more similar the two populations look and the harder it is to tease them apart. These ideas apply to most hypothesis testing situations, though the definition of the "effect size" may vary depending on

the type of test.

Significance and power summary. We can summarise the conclusions regarding significance and power. If we reject the null if $p < \alpha$, then

- If the null is true, the chance of incorrectly rejecting it (false discovery) is α, and decreases with decreasing α;
- If the null is false, the chance of incorrectly accepting (not rejecting) it increases with decreasing α;
- If the null is false, the chance of correctly rejecting it increases with: (i) increasing effect size; (ii) increasing sample size.