FIT2086 Studio 5
Hypothesis Testing

Daniel F. Schmidt

August 30, 2019

Contents

1 Introduction

This Studio session will introduce you the ideas of hypothesis testing, and how to use hypothesis tests to explore and analyze data. During your Studio session, your demonstrator will go through the answers with you, both on the board and on the projector as appropriate. Any questions you do not complete during the session should be completed out of class before the next Studio. Complete solutions will be released on the Friday after your Studio.

2 Simple hypothesis testing of means

Let us examine simple hypothesis testing of the mean of a normal distribution. If we have a sample $\mathbf{y} = (y_1, \ldots, y_n)$ of n data points from a normal population with unknown mean and known variance σ^2, then we can test the hypothesis

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

by first computing the ML estimate of μ

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

which is equivalent to the sample mean. Then we compute the z-score

$$z_{\hat{\mu}} = \frac{\hat{\mu} - \mu_0}{(\sigma/\sqrt{n})} \tag{1}$$

which can be interpreted as a standardised difference of the sample mean from the reference point μ_0 and is our test statistic for this problem. The p-value is then found as

$$\begin{aligned} p &= 2\,(1 - \mathbb{P}(Z < |z_{\hat{\mu}}|)) \\ &= 2\,\mathbb{P}(Z < -|z_{\hat{\mu}}|) \end{aligned} \tag{2}$$

which is the probability that a RV $Z \sim N(0,1)$ would be greater than $|z_{\hat{\mu}}|$ in either direction. While the above two formulas are the same, equation (2) is preferred as it is more numerically accurate if $z_{\hat{\mu}}$ is large. The p-value measures the strength of evidence against the null hypothesis.

1. For a given sample with sample mean $\hat{\mu}$, what happens to the z-score if the population variance increases?

   A: The z-score is the difference of the sample mean $\hat{\mu}$ from the hypothesised population mean μ_0 divided by the standard error, which is $(\sigma/\sqrt{n})$. If the population variance σ^2 increases, so does σ, and therefore the z-score will get smaller. This is because the z-score is a measure of strength of difference relative to the variability in the estimate, i.e., standardised by the variation we would expect to see in the estimate if we drew a new sample from our population.

2. For a given sample with sample mean, $\hat{\mu}$, what happens to the p-value if the population variance increases? How can you interpret this?

   A: If the population variance increases, the z-score will decrease as discussed above; as the p-value is smaller for larger z-scores (more extreme differences from the hypothesised population mean μ_0), a decrease in z-score will lead to an increase in p-value, and weaker evidence against the null.

3. Imagine we observed some data $\mathbf{y}$, and test $H_0 : \mu = \mu_0$ vs $H_A : \mu \neq \mu_0$ using the above procedure. We find that the p-value is 0.9.

   (a) What does a "p-value of 0.9" mean?

   A: A p-value of 0.9 means that, if the null was true, then 90% of possible samples that we could take from the population would lead to as extreme an observed difference from μ_0, or a more extreme difference, than the difference we have actually observed in our data. That is, 9 out of 10 possible samples we could draw from the population, if the null hypothesis was true, would lead to an absolute z-score as larger or larger than the one we observed from the sample we have.

   (b) Does this prove that the population mean $\mu = \mu_0$? If not, what does this p-value suggest?

   A: We can only collect evidence against the null, never in favour of the null – so this does not prove that the population mean $\mu = \mu_0$. A p-value of 0.9 means that almost any sample we could observe (90% of possible samples), if the null was true, would lead to a z-score as large or larger than the one we have observed; this says that the data we have observed is not at all at odds with the null distribution, and therefore offers no evidence against the null.

Of course, it is generally not possible to know what the population variance is, so this assumption is usually unrealistic. Instead, it is common to use the data itself to estimate the variance using the unbiased estimate of variance, which means that our test statistic becomes a t-score (see Lecture 5, Slides 34–35). We will now look at using the R function `t.test()` to compute t-test p-values. Use the command `?t.test` to get help on this function in RStudio.

4. Load the file `bpdata.csv` into R. This contains measurements on systolic blood pressure, weight pulse rates and stress measures for 20 males aged 47–56 years old.

5. A person is said to be "at risk" for high blood pressure if their systolic blood pressure is between $120 - 139$ $mmHg$, and is said to have high blood pressure if their blood pressure is greater than $139mmHg$. Knowing this, we could using `t.test()` to test to see if this group of males potentially comes from an "at risk" population by the two-sided test

$$\begin{aligned} H_0 &: \mu = 120 \\ &\text{vs} \\ H_A &: \mu \neq 120 \end{aligned}$$

   A: See `studio5.solns.R` for code.

6. As we are really interested to see if our group is an "at risk" or high blood pressure group, rather than just an "at risk" group, we can instead use the one-sided test

$$\begin{aligned} H_0 &: \mu \geq 120 \\ &\text{vs} \\ H_A &: \mu < 120 \end{aligned}$$

   You can do this by setting `alternative="less"` when calling `t.test()`.

   This alternative makes more sense as we can test $\mu \geq 120$ as the lower end of "at risk" people; if the evidence is strongly against this hypothesis when we are using the alternative that the

average blood pressure is *less* than $120mmHg$ then we have a stronger case to argue that the population our sample comes from has a healthy average blood pressure.

How do the p-values compare between the two-sided and one-sided tests?

A: See `studio5.solns.R` for code. The p-values in both cases are quite small; for the two-side test approximately 9×10^{-5}. This is strong evidence against the null – if the null hypotheses were true, only 1 in around $10,000$ samples from the population would lead to an observed standardised difference from $\mu_0 = 120$ as large, or larger than, the one we have observed just by chance. This is extremely unlikely.

The one-sided test is potentially more sensible here because we are testing the situation that our sample comes from a healthy population vs our sample comes from either an "at-risk" population or a high blood pressure population, i.e., we are just asking whether our population is healthy or not. The p-value halves when we perform the one-sided test, as now we are only treating large negative deviations from μ_0 as evidence against the null – large positive deviations (i.e., if the sample mean was greater than 120) would not be seen as evidence against the null, i.e., $H_0 : \mu \geq 120$).

7. Note that `t.test()` also produces confidence intervals. How do the confidence intervals vary when you use `alternative="two.sided"` (the default) as compared to `alternative="less"`?

   A: The two-sided confidence interval gives us a plausible range of values for the unknown population mean μ. When we select a one-sided alternative, such as `alternative="less"`, the confidence interval now gives us a plausible upper bound on the unknown population mean. In general, two-sided confidence intervals are more useful than one-sided confidence intervals as ranges of plausible values to report – however, one-sided tests can be more useful if we are testing specific hypotheses such as the one above, that is, testing whether the population our sample came from is healthy vs. they are either "at risk" population or a high blood pressure population.

8. You can control the coverage level $(1-\alpha)$ of the confidence interval by the parameter `conf.level`. Vary this value from 0.9, 0.95 and 0.99 for the two-sided test. How do the values of the confidence interval change?

   A: The confidence interval increases in size as $(1 - \alpha)$ grows; this is because to guarantee we cover the true population mean with higher confidence we need to cover a wider range of the possible parameter values.

3 Testing Differences of Means

A very common application of hypothesis testing is to test whether the means of two groups are the same, or whether they are different. This obviously has an enormous number of applications – testing if a drug has a positive effect on disease progression, testing whether a change in production techniques improvements quality, etc.

1. Let us begin with the following scenario: imagine we have run a small trial of a new drug to treat leukemia. The trial divided a group of leukemia patients into two "arms": (i) a treatment arm in which the patients were administered the drug, and (ii) a control arm in which the patients were administered a placebo. After the trial period of 12 months had elapsed, we found that the mortality in the group treated with the drug was half of that in the control group, and the p-value was 0.17. Think about the concepts of testing and decide, on the basis of this data, whether the following statements are true or false:

(a) The treatment is useless and has no effect.
A: False. The treatment appears to have a potentially strong effect on mortality.

(b) There is no point in continuing to develop the treatment.
A: False. Again, the treatment has shown a potentially strong effect on reducing mortality in our trial.

(c) As the reduction in mortality is so great we should look to immediately introduce the treatment.
A: False. The reduction may be great but the p-value is only 0.17; this says that even if there was no difference at the population level between the placebo and drug groups, we would expect to see a reduction in mortality this great, or greater, just by chance for 17% of possible samples, i.e., approximately a 1 in 6 chance.

(d) A larger trial should be conducted with a greater sample size.
A: True. The drug shows potential for a strong reduction and the p-value is borderline – a larger sample size is needed to make a more informed decision.

Now let us return to last week's question regarding the use of the Standard & Poor's economic Index as a surrogate for the US economy during the 2007-2009 financial crisis. We looked at the difference in average S&P Index value before and after the collapse of the Lehman Brothers investment bank at the end of September 2008, and calculated a confidence interval for the difference in mean index levels pre- and post-collapse. Let's call the two groups "pre" and "post" collapse. To summarise these results, we found that:

$$\hat{\mu}_{\text{pre}} = 1,381.703, \;\; \hat{\sigma}^2_{\text{pre}} = 9,383.026, \;\; n_{\text{pre}} = 58$$

and

$$\hat{\mu}_{\text{post}} = 886.916, \;\; \hat{\sigma}^2_{\text{post}} = 7,002.371, \;\; n_{\text{post}} = 50$$

The difference in mean S&P indices between the two groups was

$$\hat{\mu}_{\text{pre}} - \hat{\mu}_{\text{post}} = 494.78.$$

We calculated the approximate 95% confidence interval for the difference to be (460.735, 528.838). As both ends of the interval are quite far from zero, and the interval is reasonably narrow compared to the size of the difference, we believed it was strong evidence that there was a genuine difference in S&P indices pre- and post-collapse. We will now look at formally testing the hypothesis that the bank's collapse is associated with an adverse effect on the economy. We can do this by testing the hypothesis

$$\begin{aligned} H_0 \quad &: \quad \mu_{\text{pre}} = \mu_{\text{post}} \\ &\text{vs} \\ H_A \quad &: \quad \mu_{\text{pre}} \neq \mu_{\text{post}} \end{aligned}$$

under the assumption that the population variances in the two groups, σ^2_{pre} and σ^2_{post}, are unknown.

1. Load the data `S&P500.csv` into R. Write a script to calculate the difference and a p-value for the hypothesis outlined above, using the approximate difference in means approach *(see Lecture 5, Slide 42)*.

A: See `studio5.solns.R` for code. The z-score is

$$z = \frac{\text{diff}}{\text{se}_{\text{diff}}}$$

where

$$\text{diff} = \hat{\mu}_{\text{pre}} - \hat{\mu}_{\text{post}} = 1381.703 - 886.916 = 494.787$$

is the difference in means between the two samples, and

$$\text{se}_{\text{diff}} = \sqrt{\frac{\hat{\sigma}^2_{\text{pre}}}{n_{\text{pre}}} + \frac{\hat{\sigma}^2_{\text{post}}}{n_{\text{post}}}} = \sqrt{\frac{9383.026}{58} + \frac{7002.371}{50}} = 17.373$$

is the standard error for the difference.

The z-score is therefore $494.787/17.373 = 28.48$. We see that the observed difference (494.787) is large compared to the standard error for the difference (17.373), which means that the difference is large compared to the variability we would expected to see in the differences if we were able to repeat the experiment (somehow go back in time and cause the bank to collapse again!). Therefore we believe this difference will offer strong evidence against the null hypothesis that the population difference is zero (i.e. the S&P index before and after the collapse has the same mean); in fact

$$p = 2\,\mathbb{P}(Z < -28.48) = 2 * \texttt{pnorm}(-28.48) \approx 2 \times 10^{-178}.$$

This p-value is incredible small which suggests the observed difference is incredibly unlikely to have arisen just by chance under the null hypothesis that the Bank collapse and the change in economy are unassociated.

2. The `t.test()` function we examined in Section 2 can also be used to compare differences of means, and in the case that the variances are unknown will often be more accurate than the approximate method we used above. To specify that we are testing the means of two samples against each other we use the `y` argument to specify the second sample, and now the `mu` argument specifies the value to test the difference in means against. To test for equality of means, we use `mu = 0` (the default value). As previously mentioned, `t.test()` also returns the confidence intervals for either the mean (if one sample is used) or the difference in means (if two samples are used).

   (a) Calculate the p-value and 95% CI using `t.test()`, assuming the variances are unknown and different. To do this, use the Welch approximate degrees-of-freedom approach by using the `var.equal = F` option when calling `t.test()` (note: this is the default setting).

   (b) Calculate the p-value and 95% CI using `t.test()`, assuming variances are equivalent. To do this, you can use the `var.equal = T` option when calling `t.test()`.

   A: See `studio5.solns.R`. The p-values in both cases are incredibly small, just as in the case of the approximate method.

3. How do the three p-values compare and why?

   A: The p-values are all essentially the same as the difference is so enormous.

4. How do the three confidence intervals compare?

   A: The confidence intervals vary a little with the approximate CI of $(460.735, 528.838)$ being the narrowest (and therefore, a little "overconfident") because we did not take into account the variability in our estimates of the variances. Assuming the variances are the same leads to a little wider interval in this case, likely because the variances seem different: i.e., $\hat{\sigma}^2_{\text{pre}} \approx 7002$ vs $\hat{\sigma}^2_{\text{post}} \approx 9383$.

4 Hypothesis testing of binary data

Hypothesis testing from binary data occurs commonly in industry. This is because many reliability problems can be viewed in a success/failure framework. For example, we might be interested in the proportion of microprocessors being manufactured that are faulty. Recall our example of binary data from last week; we were playing "guess the coin" with our friend, who tossed a coin $n = 12$ times. We recorded the sequence of heads and tails as

$$\mathbf{x} = (0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1)$$

where a heads was coded as a "1", and a tails as a "0". Our friend claims she is using a fair coin (probability of heads $\theta = 1/2$), but we are not sure. Our best guess of the probability was

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{4}{12} = \frac{1}{3}.$$

Using the central limit theorem we derived an approximate 95% confidence interval for $\hat{\theta}$ (see Solutions for Studio 4):

$$\left(\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}},\ \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right). \tag{3}$$

Using the CLT we can also find an approximate p-value for the hypothesis test

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

by noting that $\hat{\theta} \xrightarrow{d} N(\theta_0, \theta_0(1-\theta_0)/n)$ (see Studio 4 solutions and Lecture 4), computing a z-score of the form

$$z_{\hat{\theta}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}} \tag{4}$$

and using

$$p = 2\,\mathbb{P}(Z < -|z_{\hat{\theta}}|) \tag{5}$$

where $Z \sim N(0, 1)$. For our observed data, the 95% confidence interval for $\hat{\theta}$ obtained using (3) was found to be $(0.066, 0.6)$.

1. Test our friend's claim that her coin is fair ($\theta_0 = 1/2$) using the above approximate procedure (4) and (5).

   A: See `studio5.solns.R` for code. Our estimate of $\hat{\theta}$ is 1/3; this gives us an approximate z-score of

   $$z_{\hat{\theta}} = \frac{1/3 - 1/2}{\sqrt{(1/2)(1-1/2)/12}} \approx -1.155,$$

   and a p-value of

   $$p = 2\,P(Z < -1.155) = 2 * \texttt{pnorm}(-\texttt{abs}(1.155)) \approx 0.248.$$

   A p-value of 0.248 which means that if we our coin was fair, and we tossed it $n = 12$ times then almost 25% of the time these 12 throws would result in 4 or less heads OR 8 or more heads (i.e., be biased towards heads or tails by the amount observed in our sample). This is not strong evidence against the null.

2. R provides a more exact test in the `binom.test()` function. Use this function to compute a p-value for the hypothesis that $\theta = \theta_0$, and a confidence interval for $\hat{\theta}$. How good are the values obtained by the approximate procedure?

   A: See `studio5.solns.R` for code. We use

   $$\texttt{binom.test}(\texttt{x} = 4, \texttt{n} = 12, \texttt{p} = 1/2) \approx 0.3877.$$

   The exact p-value is 0.3877, which is a bit bigger than our approximate procedure, but gives the same overall conclusion. If the sample size was larger we would expect the two p-values to be closer, as the normal approximation on which our approximate method is based would be better.

3. We observed $h = 4$ heads out of $n = 12$ throws; by changing the `x` parameter in `binom.test()` we can vary the number of observed heads in a dataset. For `n = 12`, how many/few heads would we need to see before we could start to suspect the coin is not fair?

   A: See `studio5.solns.R` for code. We see that

   $$\texttt{binom.test}(\texttt{x} = 3, \texttt{n} = 12, \texttt{p} = 1/2) \approx 0.146,$$

   which is still not strong evidence against the null. If we have

   $$\texttt{binom.test}(\texttt{x} = 2, \texttt{n} = 12, \texttt{p} = 1/2) \approx 0.0385$$

   which is starting to be strong evidence against the null, i.e., it is starting to be unlikely to see this much bias in $n = 12$ throws of a fair coin just by chance. So for 2 heads out of 12, or 10 heads out of 12, we would start to question the fairness of the coin.

While we are playing "guess the coin", we receive a phone call and are briefly distracted. When we return and play another 12 rounds, we observe the sequence:

$$\mathbf{y} = (0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1).$$

We become suspicious she may have swapped the coin while we were out of the room. We can formally verify our suspicions by testing the hypothesis

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

where θ_x is the population probability of success for the first coin, and θ_y is the population probability of success for the second coin. Using the central limit theorem we can derive a simple formula for an approximate p-value for the above test. Let $\hat{\theta}_x$ and $\hat{\theta}_y$ be the maximum likelihood estimates (equivalent to sample mean) of the success probabilities for sample one and two respectively, let m_x and m_y be the number of heads in the two samples and let n_x and n_y be the size of the two samples. Then we can calculate an approximate z-score for the difference of the probabilities (see Lecture 5, Slides 46–47)

$$z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{\hat{\theta}_x - \hat{\theta}_y}{\sqrt{\hat{\theta}_p(1 - \hat{\theta}_p)(1/n_x + 1/n_y)}} \tag{6}$$

where

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

is the pooled estimate of θ under the null hypothesis. Using this we can get an approximate two-sided p-value from

$$p = 2\,\mathbb{P}(Z < -|z_{\hat{\theta}_x - \hat{\theta}_y}|) \tag{7}$$

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

4. Test the hypothesis $H_0 : \theta_x = \theta_y$ vs $H_A : \theta_x \neq \theta_y$ using the approximate procedure given by equations (6) and (7).

   A: For our two samples, we have:

   $$m_x = 4,\ n_x = 12 \text{ and } m_y = 10, n_y = 12,$$

   so that our two estimates of probability of a heads for our two samples are

   $$\hat{\theta}_x = 1/3 \text{ and } \hat{\theta}_y = 5/6.$$

   Our pooled estimate of the success probability is

   $$\hat{\theta}_p = \frac{m_x + m_y}{n_x + n_y} = \frac{4+10}{12+12} = 14/24 \approx 0.5833.$$

   Our null hypothesis now does not specify a value of θ for the two coins, just that $\theta_x = \theta_y$. This pooled estimate would be a better estimate of the success probabilities of the two coins if the two coins were the same, i.e, under the null hypothesis. This leads to an approximate z-score of

   $$z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{1/3 - 5/6}{\sqrt{(14/24)(1-14/24)(1/12+1/12)}} \approx -2.484.$$

   This z-score is approximate because it is based on the normal approximation for the sampling distribution of the difference in probabilities – this will only be very accurate for large n_x and n_y. The p-value is then

   $$p = 2 * \texttt{pnorm}(-2.484) \approx 0.0130$$

   So, this approximate p-value suggests that if we repeated the experiment and the coin had not been changed between our two sequences of throws, then approximately 1 out of 77 throws of $n_x + n_y = 24$ coins would result in a difference as great as we observed, or greater, just by chance. This would lead us to believe that the coin that was used was likely different between the two sequences of coin tosses we observed.

5. There are more accurate tests for comparing binomial/Bernoulli probabilities (proportions). One of these is implemented in the R function `prop.test()`. This takes a parameter `x` which is a vector of counts of successes (the length of which is the number of different samples of Bernoulli trials to test), a vector `n` of number of trials in each of the samples, and a parameter `conf.level` which can be used to select the level of confidence interval the function additionally generates. Use this function to test whether the two coins are the same.

   A: We use

   $$\texttt{prop.test}(\texttt{x} = \texttt{c}(4,10), \texttt{n} = \texttt{c}(12,12)) \approx 0.0383$$

6. How do the two p-values compare? Do you think the data suggests that your friend swapped coins while you were out of the room?

   A: The p-value of the more exact test is 0.0384, or about 1 in 26 times we would expect to see a difference as great, or greater than the one we observed just by chance, if the coin was the same

for both sequences of throws. This is weaker evidence against the null than the approximate test (the approximate test is a overstating the evidence), but would still offers strong evidence against the null hypothesis that the coin was the same for the two sequences.

Important note: When interpreting p-values and making decisions based on the evidence against the null, it is important to note that whether you reject the null hypothesis or not might depend on the *consequences* of an incorrect decision. If incorrectly rejecting the null hypothesis (concluding for example, there is an effect or a difference at the population level) might lead to serious negative consequences (i.e., death or serious financial loss) then we should consider using a smaller threshold for rejection, i.e., we should demand stronger evidence before rejecting the null hypothesis.

For example, a drug might show a decrease in mortality but bring with it known, serious side effects. If we see a p-value in the order of 0.01 we might think that a 1 in a 100 chance of making a false discovery is small, but if the result is that large number of people will suffer serious side effects, with potential legal ramifications if we reject the null and say the drug reduces mortality, this may not be a small enough chance to warrant rejecting the null hypothesis. In this case we may decide only a chance in the order of 1 in $10,000$ (i.e., p-value of $\leq 10^{-4}$) or smaller is sufficient evidence.

This is why we are taught in this subject to evaluate the evidence rather than use a "rule-of-thumb" approach such as always rejecting the null hypothesis if $p < 0.05$. Statistical/data science procedures give you the objective evidence to help you make informed decisions, but cannot replace reasoning and judgement, which is often situation dependent.