FIT2086 Studio 5
Hypothesis Testing

Daniel F. Schmidt

August 19, 2017

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

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

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?

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?

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 the the p-value is 0.9.

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

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

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 $139 mmHg$. 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

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

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

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

   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 $120 mmHg$ 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?

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"`?

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?

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.

(b) There is no point in continuing to develop the treatment.

(c) As the reduction in mortality is so great we should look to immediately introduce the treatment.

(d) A larger trial should be conducted with a greater sample size.

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 had an adverse effect on the economy. We can do this by testing the hypothesis

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

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)*.

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()`.

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

4. How do the three confidence intervals compare?

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

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

by noting that $\hat{\theta} \xrightarrow{d} N(\theta, \theta(1-\theta)/n)$ (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).

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?

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?

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

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

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).

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.

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?

5 Predictors of Diabetes

In this last question we will use hypothesis testing to try and determine which measured variables are good predictors of diabetes in a study of Pima ethnic indians (native Americans) from the United States. To do this, we note that diabetes status is a binary variable, and that each of the variables that has been collected on the participants can therefore be divided into two groups. One of these groups will be the values of that variable for people with diabetes, and one will be the values of that variable for people without diabetes. We call this stratifying by diabetes status.

We can then use a t-test to see if there is a difference in the average value of the variable between people with and without diabetes – if there is, then the larger the (standardised) difference between the two groups, the better the predictor of diabetes the variable will be. The the p-value can be used as a measure of strength of (standardised) association between the variable and diabetes status. Essentially we are looking for variables which are markedly differently in people with and without diabetes.

Begin by loading the `pima.csv` dataset into R; this file contains the following predictors variables:

- Number of Pregnancies (`PREG`)
- Plasma Glucose Concentration (`PLAS`)
- Diastolic Blood Pressure (`BP`)
- Triceps Skin Fold Thickness (`SKIN`)
- 2-hour Serum Insulin Levels (`INS`)
- Body Mass Index (`BMI`)
- Diabetes Pedigree Function (`PED`)
- Age (`AGE`)

as well as the outcome variable Diabetes (`DIAB`).

1. Examine the dataset in R and determine the basic characteristics of the dataset and each variable.

2. Write a script that loops over each of the predictors. For each predictor, it should stratify the predictor by diabetes status into two data vectors, and then use `t.test()` to compute a p-value for the hypothesis that the mean in the two groups is the same, assuming the variances in the two groups are the same.

3. If you rank the predictors from smallest to largest p-value, which varibles are the best potential predictors of diabetes? Which are possibly not useful?

4. Repeat the same operations but this time do not assume the variances in the two groups are the same. Do the results change?

5. Take the most strongly associated predictors and produce side-by-side box-plots for the two groups. You can do this using the `boxplot()` command and the special formula operator "∼". For example, the side-by-side boxplots for Blood Pressure (`BP`) stratified by diabetes can be produced using `boxplot(BP ∼ Diabetes, data=pima)` (replacing `pima` by whatever your dataframe is called).

   What do the boxplots reveal?

6. Finally, once we have identified good predictors, how could we go about using them to try and predict whether someone has diabetes or not?

These operations allow us to identify potential predictors of diabetes, but it is not immediately obvious how we can use the variables to actually make predictions. This is the topic of the next two weeks when we study regression models, for both continuous and binary outcome data.