Ma 3b Practical – Recitation 7

February 27, 2025

Definition 11.0.1 (Significance level). This is the probability of rejecting the hypothesis H_0 if it is true. For a given confidence interval, the significance level $\alpha \in [0, 1]$ is also the parameter which controls the confidence level $1 - \alpha$.

Definition 11.0.2 (p-value). For a given hypothesis test, the p-value is the probability that an event occurs which is as extreme or more extreme than what is observed under the assumption that H_0 is true.

General idea of hypothesis tests :

- We reject H_0 if the p-value of the test is below the level of significance. We don't reject H_0 otherwise. In other words, we want to reject the default hypothesis H_0 if the probability of our observation (or a more extreme observation) under H_0 is too small compared to the threshold α that we set;

- WE NEVER ACCEPT H_1, because no matter how small the p-value is (i.e. no matter how improbable our observation (or a more extreme observation) is under H_0), it is never a proof that H_1 is true. The same idea is at the heart of scientific research and the justice system. We accumulate evidence to support a theory, but we never prove a theory. The philosophical idea (going back to Karl Popper) is that a scientific theory must be falsifiable, inductive reasoning is not proof. Likewise, an accused is innocent until there is enough evidence to dismiss his innocence with a level of confidence deemed appropriate.

Exercise 1. (hypothesis test μ with known σ^2) We know that the standard deviation in a certain study population is 20 . Determine the p-value for a hypothesis test that the population mean is 50 (or not), if the observed sample mean for 64 observations is 52.5

Exercise 2. (hypothesis test: test μ with unknown σ^2)

Miguel is interested in studying average years of schooling in various countries around the world. His initial research focused on Costa Rica. He hypothesized that the mean years of school for people 18 years old or above is higher than 8.69 years.

In order to test his hypothesis, he drew a random sample from the 2011 census of $\mathbf{299,071}$ people. He found out that the mean number of years of schooling for his sample population is 8.70, with a standard deviation of 4.52. Based on these results, with an alpha of 0.05, can Miguel reject the null hypothesis and conclude that the mean number of years of schooling in the population is higher than 8.69? (Hint: Give the null hypothesis and alternative hypothesis with respect to the exercise, use the data to calculate t-test. Search for the table https://www.mathsisfun.com/data/standard-normal-distribution-table.html to get the p-value.)

Exercise 3. (confidence interval with unknown variance σ^2)

There are 80 customers visiting a restaurant, among them, 60 reply they are satisfied with the service they received

Calculate a 95% confidence interval for the proportion of satisfied customers.

Exercise 4. (confidence interval with known σ^2)

A sample of 20 cigarettes is tested to determine the nicotine content per cigarette. The sample mean value is 1.2mg. Find a 99% confidence interval (two-sided) for the expected value of nicotine if we know that the standard deviation is $\sigma = 0.2$mg.

Exercise 5. (confidence interval with unknown σ^2)

Assume that in the above exercise, the (theoretical) variance is not known in advance and assume that in our sample of 20 cigarettes, we observe a sample variance of 0.04 . Find a 99% confidence interval (two-sided) for the expected value of nicotine. (Hint : Use the Student distribution.)

Exercise 6. (test $\mu_A = \mu_B$, with σ_A, σ_B known)

A sample of 10 fish from Lake A shows, using some technique, that their concentration of PCB toxin, in parts per million (ppm), is :

$$11.5, 10.8, 11.6, 9.4, 12.4, 11.4, 12.2, 11, 10.6, 10.8.$$

With a different technique, the same type of measurement is carried out on 8 fish from lake B, and the following PCB concentrations are obtained :

$$11.8, 12.6, 12.2, 12.5, 11.7, 12.1, 10.4, 12.6.$$

If we know that the variance is $\sigma_A^2 = 0.09$ for the measurements made with the technique of lake A and we know that the variance is $\sigma_B^2 = 0.16$ for the measurements made with the technique of lake B, can we reject the proposition " the two lakes have the same level of PCB contamination " at the level of significance $\alpha = 0.05$?

Remark: p value isn't the probability that null hypothesis is true, but the possibility of observing the even worse result assuming the null is true.

Solution. Exercise 1 We have $\sigma = 20$ and $n = 64$. The null hypothesis is $H_0 : \mu = \mu_0 = 50$. If we denote the observed sample mean by $\bar{x}_n$, the p-value is

$$P\left(\left|\bar{X} - \mu_0\right| \geqslant |\bar{x}_n - \mu_0|\right) = P\left(\left|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right| \geqslant \left|\frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}\right|\right) = 2P\left(Z \geqslant \left|\frac{\bar{x}_n - 50}{20/8}\right|\right) = 0.3173$$

Solution. Exercise 2

We first give the null hypothesis: "The mean number of years of schooling in the population is equal or lower than 8.69 years." as well as the alternative hypothesis: "The mean number of years of schooling in the population is higher than 8.69 years." Then we calculate the t-test

$$t = \frac{\bar{X} - \mu_X}{s_X/\sqrt{N}}.$$

Here $\bar{X} = 8.70$, $\mu_X = 8.69$, $s_X = 4.52$, and $N = 299071$ are given by the sample data. Using the given data, we get $t = 1.845$. Finally, by searching the z-table of standard normal distribution, we know the p-value is $1 - \Phi(1.845) = 0.0325$. Thus, the data rejects the null hypothesis.

Solution. Exercise 3

We follow the way to give confidence interval without variance on slides. Calculate $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{80}{79}p(1-p)$, and thus standard deviation $s_n = \sqrt{S^2} = 0.4357$, here $p = 60/80 = 0.75$ is the observed value of probability. Then we just use the student distribution with n-1=79 degrees of freedom. The confidence interval is given by

$$\left[\bar{x} \pm t_{0.025,79}\frac{s_n}{\sqrt{n}}\right] = \left[0.75 \pm 2.284 * \frac{0.4357}{\sqrt{80}}\right] = [0.6388; 0.8612].$$

Solution. Exercise 4 We have $\bar{x}_n = 1.2, \sigma = 0.2, n = 20$ and $\alpha = 0.01$. The confidence interval is thus

$$\left[\bar{x}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[1.2 \pm 2.575\frac{0.2}{\sqrt{20}}\right] = [1.0848; 1.3152].$$

Solution. Exercise 5 We have $\bar{x}_n = 1.2, s_n = \sqrt{0.04} = 0.2, n = 20$ and $\alpha = 0.01$. This time, since the variance that we used isn't the exact variance (it is an estimation), we use the Student distribution with $n - 1 = 19$ degrees of freedom. The confidence interval is thus

$$\left[\bar{x}_n \pm t_{\alpha/2,19}\frac{s_n}{\sqrt{n}}\right] = \left[1.2 \pm 2.86\frac{0.2}{\sqrt{20}}\right] = [1.0721; 1.3279]$$

Solution. Exercise 6 We have $\sigma_A = \sqrt{0.09} = 0.3, \sigma_B = \sqrt{0.16} = 0.4, n_A = 10$ and $n_B = 8$. Moreover, given the two samples, we can calculate the averages $\bar{x}_A = 11.17$ and $\bar{x}_B = 11.9875$. The null hypothesis is $H_0 : \mu_A = \mu_B$. The test statistic is

$$z = \frac{\bar{x}_A - \bar{x}_B}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} = \frac{11.17 - 11.9875}{\sqrt{\frac{0.09}{10} + \frac{0.16}{8}}} = -4.8.$$

Since

$$|z| > z_{\alpha/2} = 1.96$$

we reject H_0 at the significance level $\alpha = 5\%$.