CE2100/CZ2100
Probability and Statistics for Computing

Tutorial 4 - Inference from Small Samples

Problem 1

A small component in an electronic device has two small holes where another tiny part is fitted. In the manufacturing process, the average distance between the two holes must be tightly controlled at 0.02 mm, else many units would be defective and wasted. Many times throughout the day quality control engineers take a small sample of the components from the production line, measure the distance between the two holes, and make adjustments if needed. Suppose at one time four units are taken and the distances are measured as

$$0.021 \quad 0.019 \quad 0.023 \quad 0.020$$

Determine, at the 1% level of significance, if there is sufficient evidence in the sample to conclude that an adjustment is needed. Assume the distances of interest are normally distributed.

Solution:

The null hypothesis is $H_0 : \mu = 0.02$, and the alternative hypothesis is $H_a : \mu \neq 0.02$. Consider the test statistic

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

which follows student-t distribution with degree of freedom $4 - 1 = 3$.

From the data we can compute that $\bar{x} = 0.02075$ and $s = 0.00171$, which gives $T = 0.877$. The rejection region is $(-\infty, -t_{3,0.005}) \cup (t_{3,0.005}, \infty) = (-\infty, -5.841) \cup (5.841, \infty)$. Since our test statistic does not fall in the rejection region, we do not have evidence to reject the null hypothesis.

Problem 2

A state agency requires a minimum of 5 parts per million (ppm) of dissolved oxygen in order for the oxygen content to be sufficient to support aquatic life. Six water specimens taken from a river at a specific location during the low-water season (July) gave readings of 4.9, 5.1, 4.9, 5.0, 5.0 and 4.7 ppm of dissolved oxygen. Do the data provide sufficient evidence to indicate that the dissolved oxygen content is less than 5 ppm? Test using $\alpha = 0.05$.

Solution:

The null hypothesis is $H_0 : \mu = 5$ and the alternative hypothesis is $H_a : \mu < 5$. Calculate that $\bar{x} = 4.933$ and $s = 0.137$, and the test statistic is

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = -1.195.$$

Therefore, we do not have sufficient evidence to reject the null hypothesis since $t = -1.195 > -t_{5,0.05} = -2.015$.

Problem 3

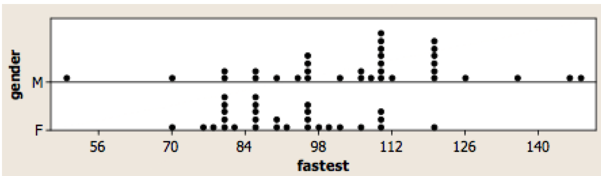
A psychologist was interested in exploring whether or not male and female college students have different driving behaviours. There were a number of ways that she could quantify driving behaviours. She opted to focus on the fastest speed ever driven by an individual. Therefore, the particular statistical question she framed was as follows:

Is the mean fastest speed driven by male college students different from the mean fastest speed driven by female college students?

She conducted a survey of a random $n = 34$ male college students and a random $m = 29$ female college students. Here is a descriptive summary of the results of her survey:

Males (X)	Females (Y)
$n = 34$	$m = 29$
$\bar{x} = 105.5$	$\bar{y} = 90.9$
$s_x = 20.1$	$s_y = 12.2$

and here is a graphical summary of the data in the form of a dotplot:



Is there sufficient evidence at the $\alpha = 0.05$ level to conclude that the mean fastest speed driven by male college students differs from the mean fastest speed driven by female college students?

Solution:

The null hypothesis is $H_0 : \mu_x - \mu_y = 0$, and the alternative hypothesis is $H_a : \mu_x - \mu_y \neq 0$. First, we can calculate the pooled sample variance by

$$S_p = \sqrt{\frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}} = \sqrt{\frac{33(20.1^2) + 28(12.2^2)}{61}} = 16.9$$

The test statistic is

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{S_p\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{(105.5 - 90.9) - 0}{16.9\sqrt{\frac{1}{34} + \frac{1}{29}}} = 3.42$$

Thus, we reject the null hypothesis since $t = 3.42 > t_{61,0.025} = 1.9996$

Problem 4

In a packing plant, a machine packs cartons with jars. It is supposed that a new machine will pack faster on average than the machine currently used. To test that hypothesis, the times it takes each machine to pack ten cartons are recorded. The results, in seconds, are shown in the tables.

New machine	Old machine
42.1	42.7
41	43.6
41.3	43.8
41.8	43.3
42.4	42.5
42.8	43.5
43.2	43.1
42.3	41.7
41.8	44
42.7	44.1

Do the data provide sufficient evidence to conclude that, on average, the new machine packs faster?

Solution:

The null hypothesis is that there is no difference in the two population means, *i.e.*, $H_0\colon \mu_1 - \mu_2 = 0$. The alternative is that the new machine is faster, *i.e.*, $H_a\colon \mu_1 - \mu_2 < 0$.

We first calculate the pooled standard deviation by

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(10-1)(0.683)^2 + (10-1)(0.750)^2}{10+10-2}} = 0.7173$$

The test statistic is

$$t = \frac{\bar{x}_1 - \bar{x}_2 - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = -3.398$$

The alternative is left-tailed so the critical value is the value a such that $Pr(T < a) = 0.05$, with degrees of freedom $10 + 10 - 2 = 18$. The critical value is -1.7341. The rejection region is $t < -1.7341$. Thus, we reject the null hypothesis since our test statistic is in the rejection region.

Problem 5

In a study on the effect of an oral rinse on plaque buildup on teeth, 14 people whose teeth were thoroughly cleaned and polished were randomly assigned to two groups of seven subjects each. Both groups were assigned to use oral rinses (no brushing) for a 2-week period. Group 1 used a rinse that contained an antiplaque agent. Group 2, the control group, received a similar rinse except that the rinse contained no antiplaque agent. A measure of plaque buildup was recorded at 14 days with means and standard deviations for the two groups shown in the table

	Control Group	Antiplaque Group
Sample Size	7	7
Mean	1.26	0.78
Standard Deviation	0.32	0.32

Do the data provide sufficient evidence to indicate that the oral antiplaque is effective? Test using $\alpha = 0.05$.

Solution:

The null hypothesis is $H_0\colon \mu_1 - \mu_2 = 0$, and the alternative is $H_a\colon \mu_1 - \mu_2 > 0$. We can calculate the pooled estimator of σ^2 by

$$s^2 = \frac{(n_1 - 1)\, s_1^2 + (n_2 - 1)\, s_2^2}{n_1 + n_2 - 2} = \frac{6(.32)^2 + 6(.32)^2}{7 + 7 - 2} = .1024$$

and the test statistic is

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{1.26 - .78}{\sqrt{.1024 \left(\frac{1}{7} + \frac{1}{7}\right)}} = 2.806$$

Thus, we can reject the null hypothesis since $t = 2.806 > t_{12,0.05} = 1.782$.

Additional Questions (Not discuss in tutorial)

Problem 6

We perform a t-test for the null hypothesis $H_0 : \mu = 10$ at significance level $\alpha = 0.05$ by means of a dataset consisting of $n = 16$ elements with sample mean 11 and sample variance 4. Use p-value to solve the following questions.

(a) Should we reject the null hypothesis in favour of $H_a : \mu \neq 10$?

(b) What if we test against $H_a : \mu > 10$?

Solution:

(a) This is a two-sided alternative. The t-statistic is

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1}{2/4} = 2$$

Since we have $n = 16$, our t statistic has 15 degrees of freedom. There are two critical values for the two-tailed test $H_0 : \mu = 10$ versus $H_a : \mu \neq 3$. We want to find the value $t_{15,\alpha/2}$ such that the probability to the right of it is $\alpha/2$, and the value $-t_{15,\alpha/2}$ such that the probability to the left of it is also $\alpha/2$. It can be shown using a t-table that the critical value $t_{15,0.025}$ is 2.13145. That is, we would reject the null hypothesis in favour of the alternative hypothesis if the test statistic T is less than -2.13145 or greater than 2.13145.

Since $-2.13145 < 2 < 2.13145$, we do not have sufficient evidence to reject the null hypothesis.

(b) This is a one-sided alternative. The t-statistic is the same

$$\frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{1}{2/4} = 2$$

We want to find the value $t_{15,\alpha}$ such that the probability to the right of it is α. According to the t-table, $t_{15,0.5}$ is 1.70305. We would reject the null hypothesis in favour of the alternative hypothesis if the test statistic is greater than 1.70305. Since $2 > 1.70305$, we reject the null hypothesis.

Problem 7

Suppose μ is the average height of a college male. You measure the heights (in inches) of twenty college men, getting data $x_1, \ldots, x_{20}$, with sample mean $\bar{x} = 69.55$ in. and sample variance $s^2 = 14.26$ in^2. Suppose that the x_i are drawn from a normal distribution with unknown mean μ and unknown variance σ^2.

(a) Using $\bar{x}$ and s^2, construct a 90% t-confidence interval for μ.

(b) Now suppose you are told that the height of a college male is normally distributed with a standard deviation 3.77 in. Construct a 90% z-confidence interval for μ.

(c) In (b), how many people in total would you need to measure to bring the width of the 90% z-confidence interval down to 1 inch?

(d) Consider again the case of unknown variance in (a). Based on this sample variance of 14.26 in^2, how many people in total should you expect to need to measure to bring the width of the 90% t-confidence interval down to 1 inch? Is it guaranteed that this number will be sufficient? Explain your reasoning.

Solution:

(a) Since we have $n = 20$ and $\alpha = 0.1$, and

$$t_{\alpha/2} = t_{19,0.05} = 1.729$$

Thus, the confidence interval is given by

$$\left[\bar{x} - t_{19,0.05} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{19,0.05} \cdot \frac{s}{\sqrt{n}}\right] = [68.08993, 71.01007]$$

(b) The 90% confidence interval can be computed as

$$\left[\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right] = [68.15839, 70.94161]$$

(c) We want $2 \cdot z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 1$, so

$$n = \left(2\sigma z_{\alpha/2}\right)^2 = 153.8$$

Thus, at least 154 people are needed.

(d) We need to find n such that $2 \cdot t_{n-1,\alpha/2} \cdot s/\sqrt{n} = 1$. However, it worth noting that the value of $t_{n-1,\alpha/2}$ depends on n. We need to check the width regarding the different choices of n. A `for` loop can be used. We find $n = 157$ is the first value of n where the width 90% interval is less than 1. This is not guaranteed. In an actual experiment, the value of s^2 won't necessarily equal 14.26. If it happens to be smaller then the 90% t confidence interval will have width less than 1.

Problem 8

Consider a machine that is known to fill soda cans with amounts that follow a normal distribution with (unknown) mean μ and standard deviation $\sigma = 3$ mL. We measure the volume of soda in a sample of bottles and obtain the following data (in mL):

$$352, 351, 361, 353, 352, 358, 360, 358, 359$$

(a) Construct a precise 95% confidence interval for the mean μ

(b) Now construct a 98% confidence interval for the mean μ

(c) Suppose now that σ is not known. Redo parts (a) and (b), and compare your answers to those above.

Solution:

(a) The sample mean is $\bar{x} = 356$. Since $z_{0.025} = 1.96$, $\sigma = 3$ and $n = 9$, the 95% confidence interval is

$$\left[\bar{x} - z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.025} \cdot \frac{\sigma}{\sqrt{n}}\right] = [354.04, 357.96]$$

(b) Similarly, we use $z_{0.01} = 2.33$, the confidence interval is

$$\left[\bar{x} - z_{0.01} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.01} \cdot \frac{\sigma}{\sqrt{n}}\right] = [353.67, 358.33]$$

(c) The sample variance is

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 15.5$$

Since $n = 9$ the number of degrees of freedom for the t-statistic is 8. Thus, the 95% confidence interval is

$$\left[\bar{x} - t_{8,0.025} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{8,0.025} \cdot \frac{s}{\sqrt{n}}\right] = [352.97, 359.03]$$

and 90% confidence interval is

$$\left[\bar{x} - t_{8,0.01} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{8,0.01} \cdot \frac{s}{\sqrt{n}}\right] = [352.20, 359.80]$$

These intervals are larger than the corresponding intervals from parts (a) and (b). The new uncertainty regarding the value of σ means we need larger intervals to achieve the same level of confidence. This is reflected in the fact that the t distribution has fatter tails than the normal distribution.

Problem 9

The following data comes from a real study in which 1408 women were admitted to a maternity hospital for (i) medical reasons or through (ii) unbooked emergency admission. The duration of pregnancy is measured in complete weeks from the beginning of the last menstrual period. We can summarize the data as follows:
Medical: 775 observations with $\bar{x}_M = 39.08$ and $s_M^2 = 7.77$.
Emergency: 633 observations with $\bar{x}_E = 39.60$ and $s_E^2 = 4.95$
Set up and run a two-sample t-test using p-value to investigate whether the mean duration differs for the two groups. What assumptions did you make?

Solution:

The null hypothesis is $H_0\colon \mu_M - \mu_E = 0$, and the alternative is $H_a\colon \mu_M - \mu_E \neq 0$. Note that the pooled variance for this data is

$$s_p^2 = \frac{(774 \times 7.77 + 632 \times 4.95)}{775 + 633 - 2} \left(\frac{1}{775} + \frac{1}{633} \right) = 0.0187$$

The t statistic for the null distribution is

$$t = \frac{\bar{x}_M - \bar{y}_E}{s_p} = -3.8064$$

We have 1406 degrees of freedom. Note that the p-value is

$$p = P(|T| > |t|) = 2 \times P(T > 3.8064) = 0.00015$$

Since the p-value is significantly smaller than commonly used significance level ($\alpha = 0.01$ or $\alpha = 0.05$), we reject the null hypothesis.