Engineering Mathematics and Statistics (B39AX)
Fall 2023

Tutorial 6

Problem A. The Michelson-Morley experiment consisted in trying to measure the speed of light in the vacuum. The following values were obtained (in hundreds of millions m/s):

$$3.0\,,\ 3.2\,,\ 3.2\,,\ 3.2\,,\ 3.0\,,\ 3.4\,,\ 3.3\,,\ 3.3\,,\ 3.1\,.$$

Compute a 90% confidence interval for the speed of light (population mean).

Let X denote the RV "speed of light", and let X_i denote "measurement i of the speed of light", for $i = 1, \dots, 9$. We want to compute a confidence interval for $\mu := \mathbb{E}[X]$. Assuming the measurements are independent and the measurement error has zero mean, the X_i's are i.i.d. copies of X and, for all $i = 1, \dots, 9$, $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \mathrm{Var}(X) =: \sigma^2$. An estimator for μ is the sample mean $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, of which we observed

$$\overline{x}_n = \frac{3.0 + 3.2 + 3.2 + 3.2 + 3.0 + 3.4 + 3.3 + 3.3 + 3.1}{9} = 3.19\,.$$

The observed (unbiased) sample variance was

$$\begin{aligned}\widehat{\sigma}_n^2 = \frac{1}{9-1}\Big[&(3.0-3.19)^2 + (3.2-3.19)^2 + (3.2-3.19)^2 + (3.2-3.19)^2 + (3.0-3.19)^2 \\ &+ (3.4-3.19)^2 + (3.3-3.19)^2 + (3.3-3.19)^2 + (3.1-3.19)^2\Big] \\ = 0.0186\,.\end{aligned}$$

As $n = 9$ is small ($n < 50$), the sample variance does not provide a good estimate of the variance, and we have to use the t-Student distribution. We have

$$T_n = \frac{\overline{X}_n - \mu}{\widehat{S}_n/\sqrt{n}} \sim t\text{-Student}(8)\,.$$

And since

$$0.9 = \mathbb{P}\Big(-t \le T_n \le t\Big) = \mathbb{P}\left(-t \le \frac{\overline{X}_n - \mu}{\widehat{S}_n/\sqrt{n}} \le t\right) = \mathbb{P}\left(\overline{X}_n - t\frac{\widehat{S}_n}{\sqrt{n}} \le \mu \le \overline{X}_n + t\frac{\widehat{S}_n}{\sqrt{n}}\right),$$

the confidence interval will be

$$\left[\overline{X}_n - t\frac{\widehat{S}_n}{\sqrt{n}}\ ,\ \overline{X}_n + t\frac{\widehat{S}_n}{\sqrt{n}}\right].$$

To determine t, we use the t-Student table, looking up the value for which

$$0.9 = \mathbb{P}\Big(-t \le T_n \le t\Big) = 1 - 2\,\mathbb{P}\Big(T_n \ge t\Big) \qquad \Longleftrightarrow \qquad \mathbb{P}(T_n \ge t) = 0.05\,,$$

i.e., $t = 1.86$. Replacing the values, we obtain the 90% confidence interval

$$\left[3.19 - 1.86\frac{\sqrt{0.0186}}{\sqrt{9}}\ ,\ 3.19 + 1.86\frac{\sqrt{0.0186}}{\sqrt{9}}\right] = \Big[3.105\,,\,3.275\Big]\ \times 10^8 \text{ m/s}.$$

Problem B. A Heriot-Watt statistician takes a random sample of 100 Heriot-Watt students traveling to work in 2019. The average value of the travel times this year (2019) was 47.2 minutes. Based on data from previous years, the statistician knows that travel times are well modeled by a normal distribution with standard deviation $\sigma = 15$ minutes, and are independent across students. Last year the mean value of the travel times was 45 minutes. Can the statistician say with 10% significance level that the mean value of travel times has increased?

Let X_i be the random variable representing the travel time of student i. We have $X_i \sim \mathcal{N}(\mu, \sigma^2)$, with $\sigma = 15$ min, for all $i = 1, \dots, n = 100$. The hypotheses are

$$\begin{aligned} \text{Null hypothesis:} \quad & H_0 : \mu \leq 45 \\ \text{Alternative hypothesis:} \quad & H_1 : \mu > 45. \end{aligned}$$

Note that this is a one-sided hypothesis test. A statistic (estimator) of μ is the sample mean,

$$\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n} \sim \mathcal{N}(\mu, \sigma^2/n)\,,$$

which is normal distributed because sums of normals are normal, and has variance σ^2/n because the X_i's are independent. The rejection region of H_0 is $\overline{X}_n > \overline{\mu}$, where the threshold $\overline{\mu}$ is determined such that false rejections (type I error) happen with probability smaller than 10%. To compute $\overline{\mu}$, we use the fact that

$$Z := \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)\,.$$

We then have

$$\Phi(c) = \mathbb{P}(Z \leq c) = \mathbb{P}\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \leq c\right) = \mathbb{P}\Big(\overline{X}_n \leq \underbrace{\mu + c\,\frac{\sigma}{\sqrt{n}}}_{\overline{\mu}}\Big) \geq 0.9\,.$$

From the normal tables, we compute $c = 1.29$. Therefore, $\overline{\mu} = \mu + c\,\frac{\sigma}{\sqrt{n}} = 45 + 1.29\,\frac{15}{\sqrt{100}} \simeq 46.94$. Since $47.2 = \overline{x}_n > \overline{\mu} = 46.94$, the statistic falls in the rejection region, and we reject the null hypothesis H_0 with 10% significance level. In other words, we conclude that the average travel time has increased.

Problem C. The average price of statistics textbooks last year was £24.96. This year a random sample of 40 such textbooks yielded a sample average of £26.10. Can we say that the mean value of statistics textbooks has increased, with a 10% significance level? Assume σ is £8.33.

Let X_i be the random variable representing the price of textbook i. As $n = 40 > 30$, we can use the central limit theorem and assume

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)\,,$$

where $\overline{X}_n = \sum_{i=1}^n X_i/n$, and $\sigma = 8.33$. The hypotheses are

$$\text{Null hypothesis:} \quad H_0 : \mu \leq \mu_0 =: 24.96$$

$$\text{Alternative hypothesis:} \quad H_1 : \mu > \mu_0.$$

The rejection region has the format $\overline{X}_n > \overline{\mu}$, where $\overline{\mu}$ is determined from the normal table using the significance level 10%:

$$\Phi(c) = \mathbb{P}\left(\frac{\overline{X}_n - \mu_0}{\sigma/\sqrt{n}} \le c\right) = \mathbb{P}\Big(\overline{X}_n \le \underbrace{\mu_0 + c\frac{\sigma}{\sqrt{n}}}_{\overline{\mu}}\Big) \ge 0.9$$

From the normal table, $c = 1.29$. This yields $\overline{\mu} = 24.96 + 1.29 \cdot 8.33/\sqrt{40} \simeq 26.66$. Since $26.10 = \overline{x}_n < \overline{\mu} = 26.66$, the statistic does not fall into the rejection region, and therefore we do not reject H_0 with significance 10%. In other words, there is no reason to believe that the average price increased.

Problem D. We want to determine whether or not students who work 20 or less hours/week get better grades than students who work more than 20 hours/week, at a significance level of 5%. Assume that the GPA of each student is normal distributed and independent from each other. We observed the data in Table 1.

Table 1: Data for problem D.

# students	Work hours	GPA average	GPA sample std
120	≤ 20	2.98	0.44
120	> 20	2.01	0.38

Define two types of random variables:

$$X_i = \text{"GPA of student } i \text{ in the group that studies} \le 20 \text{ hours"}, \quad i = 1, \dots, n_X = 120\,,$$
$$Y_i = \text{"GPA of student } i \text{ in the group that studies} > 20 \text{ hours"}, \quad i = 1, \dots, n_Y = 120\,.$$

They satisfy

$$X_i \sim \mathcal{N}(\mu_X, \sigma_X^2) \quad \text{and} \quad Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)\,.$$

The hypotheses are

$$\text{Null hypothesis:} \quad H_0 : \mu_X \le \mu_Y$$
$$\text{Alternative hypothesis:} \quad H_1 : \mu_X > \mu_Y.$$

Consider

$$W := \overline{X}_{n_X} - \overline{Y}_{n_Y} := \frac{\sum_{i=1}^{n_X} X_i}{n_X} - \frac{\sum_{i=1}^{n_Y} Y_i}{n_Y}\,.$$

We reject H_0 if $W > \xi$, where ξ is determined such that $\mathbb{P}(W > \xi) \le 0.05$. To do that, we need to find the distribution of W. As the difference (or sum) between normals is normal, W has normal distribution. Its mean and variance are

$$\mu_W := \mathbb{E}[W] = \mathbb{E}[\overline{X}_{n_X} - \overline{Y}_{n_Y}] = \mathbb{E}[\overline{X}_{n_X}] - \mathbb{E}[\overline{Y}_{n_Y}] = \mu_X - \mu_Y = 0 \qquad \text{(under } H_0\text{)}$$

$$\sigma_W^2 := \mathrm{Var}(W) = \mathrm{Var}(\overline{X}_{n_X} - \overline{Y}_{n_y}) = \mathrm{Var}(\overline{X}_{n_X}) + \mathrm{Var}(\overline{Y}_{n_Y}) = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\,.$$

However, we neither have access to σ_X^2 nor to σ_Y^2; we only know the sample standard deviations $s_X = 0.44$ and $s_Y = 0.38$. As $n_X = n_Y = 120 \gg 50$ is large, the sample standard deviations provide good approximations to σ_X and σ_Y. We then have

$$\Phi(c) = \mathbb{P}\Big(\frac{W - \mu_W}{\sigma_W} \le c\Big) = \mathbb{P}\Big(W \le \underbrace{\mu_W + c \cdot \sigma_W}_{\xi}\Big) \ge 0.95$$

According to the normal table, $c = 1.64$, and using $\mu_W = 0$ and $\sigma_W = \sqrt{0.44^2/120 + 0.38^2/120} = 0.0531$, we obtain $\xi = \mu_W + c \cdot \sigma_W = 0 + 1.64 \cdot 0.0531 = 0.087$. Since $w = 2.98 - 2.01 = 0.97 > \xi = 0.087$, we reject H_0 with significance level of 5%.