STAT 403 Tutorial · Week 2

One-Sample and Two-Sample Problems

1 Preliminaries

For this week's tutorial I'm going to give two numerical examples on one-sample and two-sample problems, including hypothesis testing and confidence intervals. We'll start with a review of some basic statistical theory that will be used.

1. Let $X \sim N(\mu, \sigma^2)$, then $Y := (X - \mu)/\sigma \sim N(0, 1)$;

2. (definitions of χ^2 and t distributions) Let $X, X_1, \dots, X_n \overset{iid}{\sim} N(0, 1)$, then $Q := X_1^2 + \cdots + X_n^2$ is said to have a chi-square distribution with n degrees of freedom, usually denoted by $Q \sim \chi_n^2$; $T := \frac{X}{\sqrt{Q/n}}$ is said to have a t distribution with n degrees of freedom, usually denoted by $T \sim t_n$;

3. Let $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, then $\hat{\mu} := \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ are unbiased estimators for μ and σ^2, respectively. In addition, $\hat{\mu} \sim N(\mu, \sigma^2/n)$, $\frac{n-1}{\sigma^2}S^2 \sim \chi_{n-1}^2$ and they are independent. Therefore, by Results 1 and 2, we have that

$$\frac{(\hat{\mu} - \mu)/\sqrt{\sigma^2/n}}{\sqrt{\frac{n-1}{\sigma^2}S^2/(n-1)}} = \frac{\hat{\mu} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

2 One-sample problem

Often in real world we are confronted with such a situation: we are given a bunch of values (a.k.a. sample) and asked to infer whether the "center" of the population where the sample comes from is equal to certain value. This is called the one-sample problem in statistics. To simplify the problem, we assumed that population follows a normal distribution.

Example 1 (one-sample z-test)**.** *Marks on an exam in a statistics course are assumed to be normally distributed with unknown mean μ but with variance σ^2 equal to 5. A sample of four students is selected, and their marks are 52, 63, 64, 84. Assess the hypothesis $H_0 : \mu = 60$ and compute a* 0.95 *confidence interval for unknown μ.*

To make it concise, we use X_1, X_2, X_3, X_4 to denote the marks of the four students and by statements above, we may assume that $X_1, \ldots, X_4 \overset{iid}{\sim} N(\mu, \sigma^2 = 5)$. Now we hope to say something about μ based on these 4 marks.

We starting by considering the hypothesis testing problem. Assessing the hypothesis $H_0 : \mu = 60$ is simply a more rigorous way of asking whether the population mean μ equals to 60 or not. Recall that a natural estimator of μ is the sample mean $\hat{\mu}$:

$$\hat{\mu} = \frac{1}{4}\left(X_1 + X_2 + X_3 + X_4\right),$$

which turns out to be $\frac{1}{4}(52 + 63 + 64 + 84) = 65.75$. Now if these marks really come from a normal distribution with mean mark $\mu = 60$, then the sample mean should also be close to 60. But is 65.75 close to 60? How close is close enough? This is where statistics plays its part. To answer these questions, we try to figure out the distribution of $\hat{\mu}$ if $\mu = 60$ is true.

Suppose that $X_1, \ldots, X_4 \sim N(\mu, \sigma^2)$, then by Result 3 in the previous section, we have $\hat{\mu} \sim N(\mu, \sigma^2/n)$ and using Result 1, we have

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

We can calculate the observed value of Z in our example $z = (65.75 - 60)/(\sqrt{5/4}) = 5.14$. We know Z is supposed to have a standard normal distribution, but 5.14 is a rather large value for it. In other words, we observed an event with very small probability, which may imply our hypothesis is not correct. Formally, we may calculate the p-value which is defined to be

$$p := \mathbb{P}(|Z| > |z|) = \mathbb{P}(Z > 5.14 \text{ or } Z < -5.14) \approx 2.747 \times 10^{-7}.$$

A common practice is to reject the null hypothesis if $p < 0.05$. Next we construct a confidence interval for μ, which can also be derived from the fact $Z \sim N(0,1)$. We know that, approximately, $\mathbb{P}(|Z| < 1.96) = 95\%$. By substituting the expression of Z into it, we have

$$\mathbb{P}(\hat{\mu} - 1.96 \cdot \frac{\sigma}{\sqrt{n}} < \mu < \hat{\mu} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = 95\%,$$

which means that we are 95% sure that mean score μ is contained in the interval $(\hat{\mu} - 1.96 \cdot \frac{\sigma}{\sqrt{n}}, \hat{\mu} + 1.96 \cdot \frac{\sigma}{\sqrt{n}}) = (65.75 - 1.96 \cdot \sqrt{5/4}, 65.75 - 1.96 \cdot \sqrt{5/4}) = (63.56, 67.94)$.

Example 2 (one-sample t-test)**.** *Drop the assumption that variance $\sigma^2 = 5$ in Example 1. Assess the hypothesis $H_0 : \mu = 60$ and compute a 0.95 confidence interval for unknown μ.*

In this case, the z-test method does not work—we can't compute z-score if σ^2 is unknown. However, since the sample variance S^2 is a good estimator of it, we can substitute σ for S in the expression of Z and obtain

$$T := \frac{\hat{\mu} - \mu}{S/\sqrt{n}},$$

which we already know is distributed as t_3 from Result 3. Here in our example,

$$s^2 = \frac{1}{4-1}[(52 - 65.75)^2 + (63 - 65.75)^2 + (64 - 65.75)^2 + (84 - 65.75)^2] = 177.58$$

and the observed value for T is $t = (65.75 - 60)/\sqrt{177.58/4} = 0.86$. This seems to be "reasonable" if $\mu = 60$ is true, because the p-value is

$$p = \mathbb{P}(|T| > |t|) = \mathbb{P}(|T| > 0.86) = 0.45.$$

Since 0.45 is larger than 0.05, we cannot reject the null hypothesis that $\mu = 60$; in other words, we did not find evidence strong enough to say $\mu = 60$ is false. Now let's construct a confidence interval. By looking up tables we know that for t_3 distribution, $\mathbb{P}(|T| < 3.18) = 95\%$. Thus with similar arguments in Example 1, the 95% confidence interval for μ is given by $(\hat{\mu} - 3.18 \cdot \frac{s}{\sqrt{n}}, \hat{\mu} + 3.18 \cdot \frac{s}{\sqrt{n}}) = (65.75 - 3.18 \cdot \sqrt{177.58/4}, 65.75 + 3.18 \cdot \sqrt{177.58/4}) = (44.56, 86.94)$.

3 Two-sample Problem

Experiments are often performed to compare two "treatments", such as two different fertilizers, machines, methods, processes or materials. We might be given two samples, each representing values obtained from one treatment, and asked to determine whether there is any real difference between them.

Example 3 (two-sample t-test)**.** *We get some marks of students from two classes. In class 1, we get marks of 4 students: 52, 63, 64, 84, which are assumed to be iid normally distributed with unknown mean μ_1 and unknown variance σ^2; in class 2, we get marks of 5 students: 51, 58, 66, 77, 80, which are assumed to be iid normally distributed with unknown mean μ_2 and unknown variance σ^2. Assess the hypothesis $H_0 : \mu_1 = \mu_2$ and determine a 95% confidence interval for $\mu_1 - \mu_2$.*

Again, for easy illustration, we denote the four marks in class 1 by X_1, X_2, X_3, X_4 and the five marks in class 2 by Y_1, Y_2, Y_3, Y_4, Y_5. We know that $X_1, \ldots, X_4 \overset{iid}{\sim} N(\mu_1, \sigma^2)$ and $Y_1, \ldots, Y_5 \overset{iid}{\sim} N(\mu_2, \sigma^2)$. We can calculate the sample mean first,

$$\hat{\mu}_1 = \frac{1}{4}(X_1 + X_2 + X_3 + X_4) = \frac{1}{4}(52 + 63 + 64 + 84) = 65.75$$

Similarly,

$$\hat{\mu}_2 = \frac{1}{5}(Y_1 + Y_2 + Y_3 + Y_4 + Y_5) = \frac{1}{5}(51 + 58 + 66 + 77 + 80) = 66.4.$$

If the hypothesis $\mu_1 = \mu_2$ holds, then $\hat{\mu}_2$ and $\hat{\mu}_2$ should be close, that is, $\hat{\mu}_1 - \hat{\mu}_2$ should be close to 0. Here $\hat{\mu}_1 - \hat{\mu}_2 = -0.65$. To determine whether this evidence is strong enough to say μ_1 and μ_2 are equal, we figure out the distribution of $\hat{\mu}_1 - \hat{\mu}_2$. It turns out

$$\frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{4} + \frac{1}{5}}} \sim N(0,1)$$

If $\mu_1 - \mu_2 = 0$, then with a probability of 95%, $\frac{\hat{\mu}_1 - \hat{\mu}_2}{\sigma\sqrt{\frac{1}{4}+\frac{1}{5}}} \sim N(0,1)$ should have a value lies in $(-1.95, 1.95)$. However, σ^2 is unknown, so we replace it by sample variance s^2. The tricky thing here is that this s^2 is calculated by combining the sample variances of marks from two classes. In class 1, the sample variance is

$$s_1^2 = \frac{1}{4-1}[(52-65.75)^2 + (63-65.75)^2 + (64-65.75)^2 + (84-65.75)^2] = 177.58$$

and in class 2, the sample variance is

$$s_2^2 = \frac{1}{5-1}[(51-66.4)^2 + (58-66.4)^2 + (66-66.4)^2 + (77-66.4)^2 + (80-66.4)^2] = 151.3.$$

Then s^2 is defined to be a weighted average of s_1^2 and s_2^2:

$$s^2 = \frac{1}{4+5-2}[(4-1)s_1^2 + (5-1)s_2^2] = 162.56$$

By some statistical theory, one may show that after substitution,

$$T' = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{4} + \frac{1}{5}}} \sim t_{4+5-2}.$$

Under the null hypothesis $\mu_1 = \mu_2$, the quantity $t' = \frac{\hat{\mu}_1 - \hat{\mu}_2}{s\sqrt{\frac{1}{4}+\frac{1}{5}}} = -0.08$. The p-value is $\mathbb{P}(|T'| > |t'|) = \mathbb{P}(|T'| > 0.08) = 0.94$ so cannot reject the null that $\mu_1 = \mu_2$. The confidence interval can be figured out by $\mathbb{P}(|T|' < 2.36) = 95\%$. Then using the expression of T', the 95% confidence interval for $\mu_1 - \mu_2$ is $(\hat{\mu}_1 - \hat{\mu}_2) \pm 2.36 \cdot s\sqrt{1/4 + 1/5} = (-20.84, 19.54)$.