Weekly Review 6

In the three hours this week, we finished Chapter 8, which discusses in details the so-called confidence intervals. Such an interval is used to estimate an unknown parameter, not by one single value but by giving you a range, indicating the possible size of error. The 60-minute mid-term test on 8 March will cover everything from Chapter 1 up to the end of Chapter 8. Then I started Chapter 9, which is devoted to one of the central questions in this course: hypothesis testing. **What I explained this week (and the week after next) is very important** because once you understand the philosophy behind the hypothesis testing for the population mean, you will understand the rationale behind the procedures of hypothesis testing for all other parameters. If you do not understand it, you will be lost very soon. And **you have to memorise and understand all technical terms introduced, or to be introduced**, because they will be repeatedly used in the future without repeating the explanations. If you do not know them, we do not have a common language (more precisely, you do not understand my language) and are not able to have intellectual discussion anymore. **Please read this (and the next) review carefully and repeatedly, at least three times, until you really understand <u>all</u> materials presented here.**

But before we talked about hypothesis testing, let us finish our discussion on estimation (which is also an important part of statistical inference).

Last week we were discussing the estimation of the population mean μ by the sample mean $\overline{X}$. As we mentioned, we have to know the sampling distribution of $\overline{X}$ and the central limit theorem tells us that if n is large ($n \geq 30$), then $\overline{X}$ has, approximately, a normal distribution. In fact, if the population itself has a normal distribution, then $\overline{X}$ has exactly, no matter whether n is large or small, a normal distribution.

Based on the fact that $\overline{X}$ follows a normal distribution, I showed you with full mathematical details that

$$\Pr\left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95,$$

and in view of this result, we say that a 95% *confidence interval* for μ is

$$\left(\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}\ ,\ \ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right),$$

or simply

$$\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}.$$

Thus, if, say, $\overline{x} = 151$, $\sigma = 10$ and $n = 16$, a 95% confidence interval for μ is 151 ± 4.9. We say 151 is a *point estimate* of μ, whilst 151 ± 4.9 is an *interval estimate*.

[Writing e.g. 151 ± 4.9 is an acceptable way to present your answer for a confidence interval; you need not report $(146.1, 155.9)$. However, every year there were, there are and there will be many students still writing e.g. $(146.1, 155.9)$ instead of 151 ± 4.9 in the final exam. Thus, when you spend time on reading my reviews now, you will save your time in the exam!]

Note that we should not say $\Pr(151 - 4.9 < \mu < 151 + 4.9) = 0.95$. It is wrong because no matter how many times I repeat the statement "$151 - 4.9 < \mu < 151 + 4.9$", the true μ is either in the interval $(146.1, 155.9)$ all the time or not in the interval all the time. Nothing is random here! The probability 0.95 applies to the method used (i.e. the way we derived the formula, in which we have a random variable $\overline{X}$) and not directly to a particular data set at hand. One sample gives us one confidence interval, and if we repeatedly take samples, then different samples give us many different intervals (and so we say **a** 95% confidence interval, not **the** 95% confidence interval, because there are many 95% confidence intervals; even for a given sample, there are other ways to construct confidence intervals but we would not consider these other ways in this course). Thus, if we repeat and repeat, among these (infinitely many) intervals 95% of them contain the true μ while 5% of them do not; this is the correct interpretation of the meaning of 95% confidence. In general, we make probability statements about *potential* error of an estimate and make confidence statements once the data have been obtained.

Now, go back to the formula $\overline{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$. The value 1.96 comes from the requirement of 95% confidence. If we want to have a 99% (90%, respectively) confidence interval, we simply change 1.96 to 2.576 (1.645, respectively). More generally, if we want to have a $100(1-\alpha)\%$ confidence interval, we replace 1.96 by $z_{\alpha/2}$, i.e.

$$\overline{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

where $z_{\alpha/2}$ is the value that the area to its right under the standard normal distribution curve is equal to $\alpha/2$, i.e. $\Pr(Z \geq z_{\alpha/2}) = \alpha/2$, and we know

$$z_{0.05} = 1.645, \qquad z_{0.025} = 1.96, \qquad z_{0.005} = 2.576.$$

Of course, the smaller the value $\alpha/2$, the larger (i.e. more to the right-hand side) the value $z_{\alpha/2}$, and so a higher level of confidence can be achieved only by a longer interval.

In most situations we do not know the population σ. In this case we simple replace σ by its (natural) estimate, i.e. the sample standard deviation s. However, the interval

$$\overline{x} \pm 1.96\frac{s}{\sqrt{n}}$$

will not give you 95% confidence because now you know less, and naturally you lose some confidence. Nevertheless, we are not interested in calculating how much confidence we have

for this interval, because there is no reason to stick to 1.96. Instead, we are interested in how to construct a 95% confidence interval when σ is unknown and estimated by s.

In order to achieve 95% confidence, we have to make the interval wider by changing 1.96 to a larger value. Since if s is a very precise estimate of σ, we may get 95% confidence by replacing 1.96 by, say, 2.1. However, if s is rather rough, we may have to replace 1.96 by a much larger value, say, 4.3 in order to achieve 95% confidence. Hence, it is clear that the replacement of 1.96 is a number depending on the precision of s and the precision of s depends on n. We denote by $t_{\alpha/2}$ the number that replaces $z_{\alpha/2}$, where $t_{\alpha/2}$ depends on n in some way (to be explained below). Thus, a $100(1-\alpha)\%$ confidence interval for μ, when σ is unknown, is:

$$\overline{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}.$$

These values of $t_{\alpha/2}$ are tabulated in Table V on the last two sheets of paper in the textbook. This table is known as the t-table, and when α is fixed, exactly which value should be taken from this table depends on the *degrees of freedom*, or *df* for short, which in this particular context $df = n-1$. The degrees of freedom of t comes from the degrees of freedom we have in the formula of $s^2 = \frac{\sum(x_i-\overline{x})^2}{n-1}$. One degree of freedom is lost because the sum of all $(x_i - \overline{x})$ must be zero and so only $n-1$ such terms are free. The denominator of s^2 is actually its degrees of freedom; it sounds very natural that we divide a sum by the total number of free terms that we summed up (i.e. divide the sum by its degrees of freedom).

Sometimes, we write $t_{\alpha/2}(n-1)$ to emphasise that the value depends on α **and** $n-1$, and please remember that it is **not** the product of $t_{\alpha/2}$ and $n-1$. Sometimes it is also written as $t_{\alpha/2,\,n-1}$. But in this course we just use the simplest form $t_{\alpha/2}$ to avoid a lengthy notation with two subscripts, unless the value of the degrees of freedom is explicitly required for some reasons.

These values of $t_{\alpha/2}$ are tabulated in Table V on the last two sheets of paper in the textbook and I have shown you how to check the values. In particular, if the value of the degrees of freedom is infinity, then you can interpret that s is the same as σ and so it is understandable that $t_{\alpha/2,\infty} = z_{\alpha/2}$ (and that is how we got the factor 1.645 and 2.576 for the formulae of the 90% and 99% confidence intervals in the case that σ is known). The same as the standard normal table, I may intentionally change the entries of the t-table in the mid-term and in the final exam so that you must use the tables provided to get the corresponding values in the mid-term/final exam. Your super-calculators do not help!

If the population has a normal distribution, we need only two parameters, namely, the mean μ and the variance σ^2, because for a normal distribution, if we know the mean and the variance, we know everything about this distribution.

However, in some applications, we definitely do not have a normally distributed population. For example, when we have *dichotomous data* (also known as *binary data*), i.e. the population consists of 0's (failures) and 1's (successes) only. For such a population, only one

parameter will be of interest, namely, the proportion p of successes. The most natural point estimate of p is the sample proportion $\hat{p}$. Note that the notation system here is different from that of the normal distribution case. In the latter we use Greek alphabets to denote parameters in the population and Roman alphabets to denote statistics calculated from a sample. In an elementary Statistics course such as this one, probably it is not a good idea to use π to denote the population proportion, and hence here we use the Roman p to denote the parameter and put a hat on top of it to denote the corresponding estimator. Putting a hat on top of a parameter (no matter it is denoted by a Greek or a Roman alphabet) is a widely accepted way to denote its estimator in Statistics, e.g. $\hat{\mu}$ denotes an estimator of μ and $\hat{\sigma}^2$ an estimator of σ^2.

Now, an important observation has to been made: the sample proportion of successes, i.e. the total number of successes in the sample divided by the sample size, happens to be the sample mean! Also, the population proportion p is the population mean μ. It is because each datum is either 0 or 1; the sum of 0's and 1's in a sample/population is just equal to the count of 1's (i.e. the number of successes) and hence the mean is the same as the proportion of 1's (successes). Therefore, if we have a large sample, then by the central limit theorem $\hat{p}$ follows approximately a normal distribution. However, what is σ or s in a population of dichotomous data?

Thinking more carefully, we could see that actually our sample is $\{X_1, X_2, \ldots, X_n\}$, where each X_i is either 0 or 1 with $\Pr(X_i = 1) = p$ and so the expectation of each X_i is $\mu = p \times 1 + (1-p) \times 0 = p$. However, we need not derive the variance from the definition of variance. We just have to note that $X_i \sim \mathrm{B}(1, p)$, and consequently the population variance of X_i is the variance of a binomial distribution with one trial, i.e. $\sigma^2 = p(1-p)$. Hence, approximately, by the central limit theorem

$$\hat{p} \sim \mathrm{N}\left(p\,,\ \frac{p(1-p)}{n}\right),$$

and so a $100(1-\alpha)\%$ confidence interval for p is

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

(which can be derived in exactly the same way as we derive the formula $\overline{x} \pm 1.96\sigma/\sqrt{n}$). It is important to note that for the standard error $\sqrt{\frac{p(1-p)}{n}}$, because we do not know p, we have to estimate p by $\hat{p}$, but we still use $z_{\alpha/2}$, not $t_{\alpha/2}$; we use $t_{\alpha/2}$ only when we use s to replace, or estimate, σ. Remember: t **is the alphabet after** s**; no** s**, no** t. Note that the normal distribution for $\hat{p}$ is a large-sample approximation anyway and changing $z_{\alpha/2}$ to $t_{\alpha/2}$ neither is mathematically correct nor can reduce the approximation error.

These formulae for confidence intervals can help us determine the sample size. The question of how large a sample to take arises early in the planning of any survey or experiment. This is an important question that should not be treated lightly. To take a larger sample

than is needed to achieve the desired results is wasteful of resources, whereas very small samples often lead to results that are of no practical value.

Suppose that we want to estimate p so that the margin of error is at most 3% with 95% confidence. That is to say, we have to find an n such that

$$n \geq p(1-p)\Big(\frac{1.96}{0.03}\Big)^2.$$

Of course we do not know p but (by simple calculus if you know, or by substituting different values into the expression) it is to easy see that the maximum of $p(1-p)$ is $1/4$. So if we choose

$$n \geq \frac{1}{4}\Big(\frac{1.96}{0.03}\Big)^2, \tag{1}$$

then because

$$\frac{1}{4}\Big(\frac{1.96}{0.03}\Big)^2 \geq p(1-p)\Big(\frac{1.96}{0.03}\Big)^2$$

we can achieve the following consequently:

$$n \geq p(1-p)\Big(\frac{1.96}{0.03}\Big)^2,$$

The inequality given in (1) means

$$n \geq 1067\frac{1}{9}.$$

Note that if you require that the margin of error must be not greater the prescribed 3%, you should not correct the answer to the nearest integer but take the smallest integer that is greater than the value you obtained. In the above example, the minimum sample size is 1068. Of course, after you fix the sample size and then take your sample, the sample proportion $\hat{p}$ from your sample is not necessarily 0.5 and hence your margin of error is not 3% but must be some value not greater than 3%.

That is the end of Chapter 8 and the mid-term test, to be held on Thursday 20 October 2016 will cover everything we have discussed up to here.

~~~~~~~~~~~~~~~~~ Mid-term test up to here ~~~~~~~~~~~~~~~~~

Good Luck!
~~~~~~~~~~~~~~~~~

Now, let us start the basic idea of hypothesis testing, which can be illustrated by the following example. If I want to know whether $\mu = 50$ or not, I am testing the *null hypothesis* that $\mu = 50$, denoted by "$\mathrm{H}_0 \colon \mu = 50$".By saying "*testing* the null hypothesis", we mean carrying out the statistical procedure below to determine whether the evidence provided by the given sample is strong enough to conclude that null hypothesis is not correct. Suppose we take a random sample and obtain, say $\overline{x} = 53$. Our job now is to determine whether an estimation error of size $\overline{x} - 50 = 3$ is too large or not. If I think 3 is a large estimation error, then I would *reject* the null hypothesis that $\mu = 50$, based on the believe that the actual estimation error (which is equal to $\overline{x} - \mu$, where μ is unknown) is not 3 but should be a smaller and hence a more reasonable value; if I think 3 is not a large estimation error, then I could only say that I do not have any strong evidence to reject the null hypothesis that $\mu = 50$.

How large the estimation error is too large? In Statistics, we usually measure a value in terms of standard deviation (so that the physical unit of measurement does not matter), and hence we have to know how many standard deviations the estimation error is, i.e. we should calculate not $\overline{x} - 50$ but $z = (\overline{x} - 50)/(\sigma/\sqrt{n})$. A large z means the estimate $\overline{x}$ deviates from the hypothesised value 50 by many standard deviations, which should not be often seen. Most of the time we should get not too large and not too small values of z. However, we still have not yet solved the problem: How large is too large? When we say a value is too large, implicitly we have two assumptions: (i) this value is an outcome of a random variable, and (ii) the probability of getting such a large value is very small. How small a probability is a small probability? We have to decide the cutoff value ourselves. We may decide that a too-large value is a value that will be obtained with probability less than, say, 0.05. (That is to say, we define "most of the time = 95% of the time", and thus 95% of the time we should get something smaller than the cutoff value.) Then, it is easy to determine whether an estimation error of 3 is a large error or not.

Alright, before we continue, we have to fix another hypothesis, viz. the *alternative hypothesis* (also called *research hypothesis*), which is the statement that we will conclude when we reject the null hypothesis. In this example it seems that naturally the alternative hypothesis should be that $\mu \neq 50$; this would be simply written as: "$\mathrm{H}_A \colon \mu \neq 50$". (However, we will see below that it is not the only alternative hypothesis that we may have and we will also see why it is important to write down the alternative hypothesis.) In real applications, the alternative hypothesis is usually, if not always, the statement we actually want to conclude. Otherwise, why should we waste our time and effort on getting evidence for rejecting the null hypothesis?

The idea of hypothesis testing is to check whether we have strong evidence against H_0 (for a trial in court, H_0: innocence) and supporting H_A (in court, H_A: guilty). In this example of testing $\mathrm{H}_0 \colon \mu = 50$, if an estimation error of 3 is considered large, then an estimation error of -3 is equally large, and an estimation error of -4 is an even larger error. Hence, either a very positive or a very negative estimation error offers strong evidence suggesting the rejection of H_0.

We know, from the central limit theorem, that $\overline{X}$ has a normal distribution with mean μ and variance σ^2/n. Thus, if the null hypothesis is true, then $Z = (\overline{X} - 50)/(\sigma/\sqrt{n})$ should have the standard normal distribution. For the standard normal distribution, the probability that $Z \geq 1.96$ or $Z \leq -1.96$ (which will often be neatly expressed as $|Z| \geq 1.96$, where $|Z|$ denotes the absolute value of Z) is 0.05. Hence, if it happens that $Z \geq 1.96$ (or $Z \leq -1.96$, respectively), we can say that the estimation error is too large (or too negative, respectively) and so we reject the null hypothesis at the 0.05 *significance level*. The value 1.96 is called the *critical value* and the statistic Z is called a *test statistic*. (Recall that any number calculated from a sample is called a statistic, and the procedure above is called a statistical test.)

Note that if we get a sample and have data, then we get a numeric value for the sample mean and we denote this numeric value by the lower case $\overline{x}$ and consequently the numeric value of standardised quantity $(\overline{x} - 50)/(\sigma/\sqrt{n})$ by the lower case z. Thus, when we are really talking about data of a sample, we use lower case symbols.

Clearly, the value 50 does not play a special role and can be replaced by any other number. Thus, in the general setting, we usually use μ_0 (which is a value explicitly specified in the question), i.e.

$$\begin{aligned} \mathrm{H}_0 &: \mu = \mu_0 \\ \mathrm{H}_A &: \mu \neq \mu_0 \end{aligned}$$

and then the test statistic is simply

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}.$$

The critical value does not depend on μ_0 and so remains 1.96 if we stick to the 0.05 significance level. More generally, if we want to test at the α significance level, we reject the null hypothesis at the α significance level whenever $|z| \geq z_{\alpha/2}$. (Forget what $z_{\alpha/2}$ is? If so, then check the discussion on confidence intervals.) If we do not know the population standard deviation σ, we replace it by s and so the test statistic becomes $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}}$. We reject the null hypothesis at the α significance level whenever $|t| \geq t_{\alpha/2}$ where the degrees of freedom is $n - 1$.

What is the interpretation of the significance level? The probability 0.05 (or in general, α) guarantees that if the null hypothesis is indeed correct, we will reject the null hypothesis **by such a decision rule** with probability 0.05 (or α) only. (We reject the null because $|z| \geq z_{\alpha/2}$. If this is a wrong decision, it means the z is really coming from the standard normal and we are just unlucky enough to observe the unlikely event $\{|z| \geq z_{\alpha/2}\}$, which has a probability of α when the null hypothesis is in fact correct.) Such an error is called the *type I* error. If we carry out such a statistical procedure, we say we *test the hypothesis at the 0.05 (or α) significance level*.

If the null hypothesis is in fact incorrect but we do not reject it, then we commit the so-called *type II* error, but the probability of committing the type II error, denoted by β, is

not under our control and is a function depending on the true μ. The probability $1 - \beta$ is called the *power*, which is of course a function of the true μ. The power is telling us how likely we can correctly reject the null hypothesis, given that the null hypothesis is wrong. Clearly, when we apply the same procedure repeatedly to different samples, the higher the α, the more times we would reject the null hypothesis by this procedure; the more times we reject, the likelier we could correctly reject a wrong null hypothesis and hence the higher the power. That is to say, α is the price we are willing to pay to buy power. If you pay less (smaller α), you get something of poorer quality (less powerful test); if you pay more (higher α), you get better quality (more powerful test). Thus, to choose α, we have to balance the price (type I error probability) and the quality (power). Lowest price ($\alpha = 0$, never reject) will give you poorest quality (zero power, never be able to reject correctly), while highest quality (power = 1, always reject) costs all you have (type I error probability = 1; the correct null hypothesis will always be mistakenly rejected).

Enjoy your Assignment 3!

Cheers,
Heng Peng