Weekly Review 8

Chapter 9 is talking about one population. In Chapter 10 we are interested in not the case of only one population but the case of two populations, e.g. men and women. The most typical question is: whether their means are the same or not. Therefore, we consider the difference between two means:

$$\mathrm{H}_0\colon \mu_1 = \mu_2,$$

and we may have three different alternative hypotheses:

$$\mathrm{H}_A\colon \mu_1 \neq \mu_2, \qquad \mathrm{H}_A\colon \mu_1 > \mu_2, \qquad \text{or} \qquad \mathrm{H}_A\colon \mu_1 < \mu_2.$$

We first discuss the case that we have **two independent samples**, one from each population.

To test the above null hypothesis, we have to construct a test statistic and a very natural one is to consider the difference between two sample means $\overline{x}_1 - \overline{x}_2$. Corresponding to each of the three possible alternative hypotheses above, if the difference is very different from zero, much larger than zero, or much smaller than zero, respectively, then we should reject the null hypothesis and conclude that the difference is significantly different from zero, significantly greater than zero, or significantly smaller than zero, respectively. (As we mentioned above, in particular, "significantly different from zero" is often simply shortened to just one word: "significant".)

To see whether the difference is much larger than zero or much smaller than zero, we have to know its distribution and an important feature of the normal distribution is that if X and Y are two independent normally distributed random variables with means μ_X and μ_Y, respectively, and variances σ_X^2 and σ_Y^2, respectively, then

$$X \pm Y \sim \mathrm{N}(\mu_X \pm \mu_Y,\ \sigma_X^2 + \sigma_Y^2). \tag{1}$$

Note that the variance of the sum or the difference is the same and is equal to the sum of two individual variances. Let me emphasise again that the variance is the <u>sum</u> of the two variances, <u>not the difference</u>. Please keep it in mind. Moreover, we take the sum of the two <u>variances</u>, not the sum of the two standard deviations. We will **never** take the sum of standard deviations. The result given in (1) also explains why we use σ^2 instead of σ as the second parameter in the notation $\mathrm{N}(\mu, \sigma^2)$. From (1), we can see immediately that

$$\overline{X}_1 \pm \overline{X}_2 \sim \mathrm{N}\left(\mu_1 \pm \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right). \tag{2}$$

[Question: why do we have n_1 and n_2 in (2) but not in (1)?]

If the null hypothesis is true, from the distribution given in (2) we know that the test statistic

$$z = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$$

follows (by the central limit theorem) approximately the standard normal distribution for large n_1 and n_2; in particular, if the two populations themselves follow two independent normal distributions, z follows exactly the standard normal distribution.

If we rephrase the hypotheses as

$$\begin{aligned} &\mathrm{H}_0\colon \mu_1 - \mu_2 = 0 \\ &\mathrm{H}_A\colon \mu_1 - \mu_2 \neq 0 \;\; (\text{or} > 0, \text{ or} < 0) \end{aligned}$$

it is natural to generalise the hypotheses from asking whether the difference is zero to asking whether the difference is some number we have in mind. First, we introduce the Greek alphabet δ (delta) as a symbol for the difference and then add a subscript '$_0$' to it to get δ_0 to denote the hypothesised value specified in the null hypothesis (c.f. the symbols μ_0 and p_0). More precisely, we consider the following hypotheses:

$$\begin{aligned} &\mathrm{H}_0\colon \mu_1 - \mu_2 = \delta_0 \\ &\mathrm{H}_A\colon \mu_1 - \mu_2 \neq \delta_0 \;\; (\text{or} > \delta_0, \text{ or} < \delta_0) \end{aligned}$$

where δ_0 is an explicitly known number that we ourselves (or the question) hypothesised, and the test statistic is simply

$$z = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}},$$

which will have the standard normal distribution (which is an approximation, when n_1 and n_2 are large; or is an exact result [i.e. not an approximation but is rigorously correct], when the two populations are indeed normal) if the null hypothesis is true. [Questions: (i) why normal? (ii) why the denominator is $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$? Read the previous review if you don't know the answers.]

Now, consider the alternative hypothesis H_A: $\mu_1 - \mu_2 > \delta_0$. If the alternative is true, then z is likely to be a large value. In other words, a larger value of z provides stronger evidence to support the alternative and when z is too large, we reject the null hypothesis. When is it too large? When it is greater than or equal to the critical value z_α. By the same token, if the alternative hypothesis is $\mathrm{H}_A\colon \mu_1 - \mu_2 < \delta_0$, then the smaller the value of z, the more likely the alternative hypothesis is true and so we reject the null hypothesis if z is too small, i.e. if $z \le -z_\alpha$. Finally, if the alternative hypothesis is $\mathrm{H}_A: \mu_1 - \mu_2 \neq \delta_0$, then a large or a small value of z suggests the alternative is more likely. The same as in other two-sided tests discussed before, we do not use $\pm z_\alpha$ but use $\pm z_{\alpha/2}$ because the total area excluded by $\pm z_\alpha$ would become 2α. Thus, we use $\pm z_{\alpha/2}$ and we reject the null hypothesis whenever $|z| \ge z_{\alpha/2}$.

The above situation is quite ideal but unrealistic because we assume that we know σ_1^2 and σ_2^2. In reality, however, we seldom know them and so we have to estimate them. Nevertheless, before we estimate them, we have to know whether they are the same or not; if they are, then we simply have to estimate one unknown common variance. That is to say, we have three cases: (a) σ_1 and σ_2 are known, (b) σ_1 and σ_2 are unknown but $\sigma_1 = \sigma_2$, and (c) σ_1 and σ_2 are unknown and $\sigma_1 \neq \sigma_2$.

Case (a) has been discussed above.

For case (b), $\sigma_1 = \sigma_2$, we estimate the common variance by

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

I already explained the idea behind this formula. The pooled estimate s_p^2 can be considered either a weighted mean of the two sample variances (weighed by their degrees of freedom) or, better, the sum of the squared deviations from the sample mean divided by the degrees of freedom (where for data in sample 1 we consider their deviations from $\overline{x}_1$, while for data in sample 2, we consider deviations from $\overline{x}_2$, i.e. each deviation is the difference between the datum and its corresponding sample mean). The latter interpretation is better because then we would not mistakenly invent a weighted mean of the two sample standard deviations; it does not make any sense to take the mean, no matter weighted or unweighted, of standard deviations. Then, the test statistic becomes

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}},$$

which has the t-distribution with degrees of freedom $n_1 + n_2 - 2$. Therefore,

$\mathrm{H}_0: \mu_1 - \mu_2$	$\mathrm{H}_A: \mu_1 - \mu_2$	Reject H_0 at the α significance level if
$= \delta_0$	$\neq \delta_0$	$\lvert z\rvert \ge z_{\frac{\alpha}{2}}$, if we know σ, $\lvert t\rvert \ge t_{\frac{\alpha}{2}}$, if we do not know σ,
$= \delta_0$ or $\le \delta_0$	$> \delta_0$	$z \ge z_\alpha$, if we know σ, $t \ge t_\alpha$, if we do not know σ,
$= \delta_0$ or $\ge \delta_0$	$< \delta_0$	$z \le -z_\alpha$, if we know σ, $t \le -t_\alpha$, if we do not know σ,

where $t_{\frac{\alpha}{2}}$ and t_α have degrees of freedom $n_1 + n_2 - 2$.

Now, for case (c), σ_1 and σ_2 are unknown and $\sigma_1 \neq \sigma_2$. In this case we estimate the two variances by s_1^2 and s_2^2, and the test statistic becomes

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}.$$

For alternatives with $\neq$, $>$ and $<$, we again reject the null hypothesis whenever $|t| \geq t_{\alpha/2}$, $t \geq t_\alpha$ and $t \leq -t_\alpha$, respectively, where the degrees of freedom of t is

$$\nu = \frac{\left(\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}\right)^2}{\dfrac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \dfrac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}},$$

which is known as *Welch's approximation* (and ν is the Greek alphabet nu). For checking the critical values from the t-table, we need an integer for ν (in fact, mathematics allows us to have non-integer values for degrees of freedom, but it is beyond the scope of this course). We always **round it down** in order to get a *conservative* critical value (a critical value greater than the one you actually have to use so that it is more difficult to reject under this conservative critical value and so less likely to commit type I error). It would be helpful to keep the following inequality in mind:

$$t_{\alpha,\, df_1} \geq t_{\alpha,\, df_2} \qquad \text{whenever } df_1 \leq df_2,$$

where the second subscript in $t_{\alpha,\, df}$ denotes the value of the degrees of freedom. That is, e.g. if $\nu = 3.9$, we will round it down to 3 so that we can obtain the critical value $t_{\alpha,\,3}$ from the t-table and if

$$\text{test statistic} \geq t_{\alpha,\,3}$$

then because

$$t_{\alpha,\,3} \geq t_{\alpha,\,3.9}$$

it is actually true that

$$\text{test statistic} \geq t_{\alpha,\,3.9}$$

and using $t_{\alpha,\,3}$ as the critical value is *conservative* because then the probability of making a type I error will be at most α.

Unfortunately, there is no interpretation for this formula of ν and so I am afraid you have to remember it by heart and by brute force.

So, for each of the three cases, we have a different test statistic:

σ_1, σ_2 known	σ_1, σ_2 unknown but equal	σ_1, σ_2 unknown and unequal
$z = \dfrac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$	$t = \dfrac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{s_p \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$	$t = \dfrac{(\overline{x}_1 - \overline{x}_2) - \delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

and we can extend the above table to the following one to summarise the decision rules:

$\mathrm{H}_0 : \mu_1 - \mu_2$	$\mathrm{H}_A : \mu_1 - \mu_2$	Reject H_0 at the α significance level if	
$= \delta_0$	$\neq \delta_0$	$\lvert z\rvert \geq z_{\frac{\alpha}{2}}$,	σ_1, σ_2 known,
		$\lvert t\rvert \geq t_{\frac{\alpha}{2}, n_1+n_2-2}$,	σ_1, σ_2 unknown but equal,
		$\lvert t\rvert \geq t_{\frac{\alpha}{2}, \nu}$,	σ_1, σ_2 unknown and unequal,
$= \delta_0$ or $\leq \delta_0$	$> \delta_0$	$z \geq z_{\alpha}$,	σ_1, σ_2 known,
		$t \geq t_{\alpha, n_1+n_2-2}$,	σ_1, σ_2 unknown but equal,
		$t \geq t_{\alpha, \nu}$,	σ_1, σ_2 unknown and unequal,
$= \delta_0$ or $\geq \delta_0$	$< \delta_0$	$z \leq -z_{\alpha}$,	σ_1, σ_2 known,
		$t \leq -t_{\alpha, n_1+n_2-2}$,	σ_1, σ_2 unknown but equal,
		$t \leq -t_{\alpha, \nu}$,	σ_1, σ_2 unknown and unequal,

where the second subscript df in $t_{\frac{\alpha}{2}, df}$ or $t_{\alpha, df}$ denotes the value of its degrees of freedom.

The above problem concerns the means of two independent samples and is called the *two-independent-sample t-test* or simply the *two-sample t-test* if we do not know the variances, and is called the two-(independent-)sample z-test if we know the variances.

Since the basic idea is the same for all tests and I already explained the idea repeatedly in these two weeks, starting from here, the focus is no longer on the basic idea but on the particular feature of each test. Hence, the pace is getting quicker because a thorough understanding of the basic idea has been presumed. It would be very difficult to catch up with me if you do not understand the basic idea. Please, (I have said this many times:) read Chapters 9 and 10 carefully and repeatedly, at least three times, until you really understand all materials in these two chapters.

Cheers,
Heng Peng