Weekly Review 11

In the last review when we wanted to test

$$\begin{aligned} &\mathrm{H}_0 \colon \mu_1 = \mu_2 = \cdots = \mu_k \\ &\mathrm{H}_A \colon \mu_i \neq \mu_j \quad \text{for some } i \neq j \end{aligned}$$

under the assumption that $\sigma_1^2 = \cdots = \sigma_k^2 = \sigma^2$, we obtained two different estimates of the unknown σ^2. The first one is $ns_{\overline{x}}^2$ and the second one is $\overline{s^2}$.

If the first estimate $ns_{\overline{x}}^2$ is much smaller than the second one $\overline{s^2}$, this tells us that the sample means $\overline{x}_1, \ldots, \overline{x}_k$ have little variation (because $s_{\overline{x}}^2$ is a measure of variation in these sample means), and so it is likely that the population means are the same, i.e. it is likely that the null hypothesis is true. In contrast, if the first estimate is much larger than the second one, then the variation among the sample means is larger than expected under H_0, meaning that the variation comes not only from σ^2 but probably also from some variation in the true means (i.e. μ_i are not the same and hence their estimates $\overline{x}_i$ are estimating different numbers). Hence, we will reject the null hypothesis if the ratio of the first estimate to the second estimate is much larger than 1. The distribution of this ratio is $F_{k-1,\,k(n-1)}$, where the numerator degrees of freedom come from that of $s_{\overline{x}}^2$, i.e. the value of the numerator degrees of freedom is equal to $k-1$, and the denominator degrees of freedom come from $s_1^2 + \cdots + s_k^2$, i.e. the value of the denominator degrees of freedom is $k(n-1)$. Thus, we know how large is too large. (How large? Ans: when $F \geq F_{\alpha,\,k-1,\,k(n-1)}$.) Note that it is a one-sided test even though we have a two-sided alternative.

In order to allow us to generalise the above approach to k samples of different sample sizes, the procedure for doing these two estimation is summarised into the so-called *ANOVA table*:

Source of variation	degrees of freedom	(Sum of Squares) SS	(Mean Square) MS	F
Between	$k-1$	SSB	$MSB = \dfrac{SSB}{k-1}$	$\dfrac{MSB}{MSW}$
Within	$k(n-1)$	SSW	$MSW = \dfrac{SSW}{k(n-1)}$	
Total	$kn-1$	SST		

where

$$SST = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{..})^2,$$

$$SSB = \sum_{i=1}^{k} \sum_{j=1}^{n} (\overline{x}_{i.} - \overline{x}_{..})^2 = n \cdot \sum_{i=1}^{k} (\overline{x}_{i.} - \overline{x}_{..})^2,$$

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{i.})^2 = SST - SSB,$$

in which

$$\overline{x}_{i.} = \overline{x}_{i+} = \frac{T_{i.}}{n}, \qquad \overline{x}_{..} = \overline{x}_{++} = \frac{T_{..}}{kn},$$

and

$$T_{i.} = T_{i+} = \sum_{j} x_{ij}, \qquad T_{..} = T_{++} = \sum_{i} \sum_{j} x_{ij},$$

i.e., $\overline{x}_{i.}$ is the mean of the i^{th} sample, $\overline{x}_{..}$ is the mean of all data from k samples, $T_{i.}$ is the total (the sum) of the i^{th} sample, and $T_{..}$ is the total (the sum) of all data.

Note that I dislike that the textbook uses n to denote the total sample size. Here and in many books, n denotes the sample size of each sample.

Thus,

$$MSB = n \cdot \frac{\sum_{i=1}^{k} (\overline{x}_{i.} - \overline{x}_{..})^2}{k-1} = n s_{\overline{x}}^2$$

gives us an estimate of σ^2 if the null hypothesis is true. On the other hand,

$$MSW = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{i.})^2}{k(n-1)} = \sum_{i=1}^{k} \frac{1}{k} \left\{ \frac{\sum_{j=1}^{n} (x_{ij} - \overline{x}_{i.})^2}{n-1} \right\} = \sum_{i=1}^{k} \frac{s_i^2}{k} = \overline{s^2}$$

(where s_i^2 denotes the sample variance of the i^{th} sample) is the mean of the sample variances, which gives us another estimate of σ^2 that does not require the null hypothesis.

If the F-ratio is equal to or greater than $F_{\alpha, k-1, k(n-1)}$, then the variation among the sample means is too large and so we reject the null hypothesis at the α significance level.

Note that in many books, “Between” and “Within” are called “Treatments” and “Error”, respectively, because if each group receives a different treatment, then between-group variation is the same as between-treatment variation; the within-group variation comes from randomness in each group and hence is considered as variation from random errors. Consequently, SSB and SSW will then be denoted by $SS(Tr)$ and SSE, respectively.

Please go through the numerical calculation shown in Section 12.2 carefully.

The above formulae can also be expressed as

$$\begin{aligned}
SST &= \sum\sum x_{ij}^2 - \frac{T_{..}^2}{kn} = \sum\sum x_{ij}^2 - kn\overline{x}_{..}^2, \\
SSB &= \frac{1}{n}\sum_i T_{i.}^2 - \frac{T_{..}^2}{kn} = n\sum_i \overline{x}_{i.}^2 - kn\overline{x}_{..}^2, \\
SSW &= SST - SSB,
\end{aligned}$$

which will make the numerical calculation easier.

In the old time when ANOVA was done by pens and papers with pocket calculators, the ANOVA table would be very helpful and such a table becomes a standard procedure so that nowadays all computer software packages will report the full ANOVA table, even though our interest is usually in the test statistic value only.

Now it is straightforward to modify the ANOVA table and the corresponding formulae for unequal sample sizes cases by changing n to n_i and kn to $\sum_{i=1}^k n_i$:

Source of variation	degrees of freedom	SS	MS	F
Between	$k-1$	SSB	$MSB = \dfrac{SSB}{k-1}$	$\dfrac{MSB}{MSW}$
Within	$\sum_{i=1}^k (n_i - 1)$	SSW	$MSW = \dfrac{SSW}{\sum_{i=1}^k (n_i - 1)}$	
Total	$\sum_{i=1}^k n_i - 1$	SST		

where

$$\begin{aligned}
SST &= \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_{..})^2 = \sum\sum x_{ij}^2 - \frac{T_{..}^2}{\sum_i n_i}, \\
SSB &= \sum_{i=1}^k \sum_{j=1}^{n_i} (\overline{x}_{i.} - \overline{x}_{..})^2 = \sum_{i=1}^k n_i(\overline{x}_{i.} - \overline{x}_{..})^2 = \sum_i \frac{T_{i.}^2}{n_i} - \frac{T_{..}^2}{\sum_i n_i}, \\
SSW &= SST - SSB.
\end{aligned}$$

If $k = 2$, can we apply ANOVA instead of two-independent-sample t-test? The answer is yes if the alternative hypothesis of the t-test is two-sided. For marginal cases (i.e. the test statistics are close to the critical values), would it be possible that ANOVA and two-independent-sample t test give different conclusions? No! Mathematicians tell us that when

$k = 2$, the F-statistic in ANOVA is in fact equivalent to the square of the t-statistic of two-independent-sample t-test (under the equal variance assumption), and the critical value of F with numerator degrees of freedom $k - 1 = 1$ and denominator degrees of freedom $\sum_{i=1}^{k}(n_i - 1) = n_1 + n_2 - 2$ is the square of that of t with degrees of freedom $n_1 + n_2 - 2$ (I encourage you to check this equality in the F-table and t-table yourselves). Thus, the two testing procedures are equivalent when $k = 2$. The advantage of t-test is that the alternative hypothesis can be one-sided but the alternative hypothesis of ANOVA must be two-sided.

Note that the χ^2-table and the F-tables do not tabulate the critical values for all different degrees of freedom. When the exact critical values are not available from the tables given, if we are lucky, then we may apply some inequality arguments to draw conclusions without any approximation. In class, I showed you how to do this and I do not repeat it here. This concludes Chapter 12.

Perhaps you have already noticed that after finishing Chapters 9 and 10, we had proceeded with quicker and quicker pace because now you should have already acquired the necessary basic ideas of statistical inferences, and what we have learnt in Chapters 11 and 12 (as well as Chapter 15 and the confidence interval and hypothesis testing part in Chapter 13) are only technically more sophisticated procedures but not conceptually more difficult ideas. If you understand the philosophy behind those simple tests of hypotheses concerning one mean or two means, there is nothing really new in concept for you to understand; instead, there are only new mathematical formulae and computational procedures for you or waiting for you to get familiar with them. Of course, it does not mean that the final exam will be easy; for each question, you have to choose the appropriate method and this is the most challenging part; once the appropriate method is identified, then the implementation of the chosen method is entirely mechanical (but please make sure you can implement each method correctly; it would be a pity if you choose the appropriate method [the most difficult part] but fail to implement it correctly, as I could not give you marks if you fail to carry out the required statistical procedure).

The idea of linear regression, the theme of Chapter 13, is that we have a sample of paired data $\{(x_i, y_i), i = 1, \ldots, n\}$ (i.e. a sample of n points with x- and y-coordinates) and we want to find a straight line that describes the relationship between x and y. The model we have is

$$y = \alpha + \beta x + \varepsilon,$$

where ε (epsilon) represents random errors. That is to say, we have errors in the values of y but not the values of x, and so it makes sense to find a line

$$\hat{y} = a + bx$$

that minimises some function of the differences of the observed y and the predicted $\hat{y}$. (Finding a line means finding its intercept a and slope b.) A natural way is to minimise

$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$, i.e. to minimise $\sum_{i=1}^{n}(y_i-a-bx_i)^2$ by finding some optimal a and b. Using simple calculus argument (which will be demonstrated in MATH2205 Multivariate Calculus next semester) or using orthogonal projection argument (which will soon be demonstrated in MATH2207 Linear Algebra) [if you don't know calculus or linear algebra, it does not matter], we can obtain

$$b=\frac{SS_{xy}}{SS_{xx}} \qquad \text{and} \qquad a=\overline{y}-b\overline{x},$$

where

$$SS_{xy}=\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y}) \qquad \text{and} \qquad SS_{xx}=\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

The symbols SS_{xy} and SS_{xx} are easy to understand: SS stands for sum of squares (as in the ANOVA table), whilst the subscript x represents $(x_i-\overline{x})$ and y represents $(y_i-\overline{y})$ so that the double subscript xy in SS_{xy} represents the cross product terms $(x_i-\overline{x})(y_i-\overline{y})$ and the double script xx in SS_{xx} represents the squared terms $(x_i-\overline{x})^2$.

This method of estimation is called the *least squares estimation* (LSE for short). You should read the instruction of your own scientific calculator to see how to obtain a and b by using the `LR` mode (`LR` stands for `Linear Regression`) or the `Lin` option under the `Reg` mode. In the exam, it is possible that I would ask students to report these two values directly from their calculators without showing the intermediate calculation steps. Don't think it is a naïve question. For a given story, it may not be straightforward to determine which variable is x and which variable is y.

The estimators a and b give us point estimates only, and of course we would like to be able to construct confidence intervals for and perform tests of hypotheses concerning the two unknown parameters α and β.

Alright. What do we have to know in order to construct confidence intervals or carry out hypothesis testing? Yes, the distributions. And whose distributions? Yes, the distributions of the estimators. Why are the estimators random? Where does the randomness come from? From the random error ε! If $\varepsilon\sim \mathrm{N}(0,\sigma^2)$, then it can be shown

$$a\sim \mathrm{N}\left(\alpha,\sigma^2\left\{\frac{1}{n}+\frac{\overline{x}^2}{SS_{xx}}\right\}\right) \qquad \text{and} \qquad b\sim \mathrm{N}\left(\beta,\frac{\sigma^2}{SS_{xx}}\right). \tag{1}$$

These results will be proved rigorously next year, in MATH3805 Regression Analysis. The two ugly formulae for variances will be replaced by much more elegant matrix formulae. In fact, next year you will see how beautiful the mathematics for regression is, when we formulate and solve the problem in terms of matrices.

I explained that σ can be estimated by the sample standard deviation of *residuals*

$$e_i=y_i-\hat{y}_i,$$

which are estimates of the random error ε. However, the sample deviation of e_i does not involve the sample mean $\overline{e}$ because we know the true population mean of ε is zero, and so we need not estimate it (a known value) by $\overline{e}$; as a matter of fact, $\overline{e} = 0$ anyway. Nevertheless, two degrees of freedom are still lost because in the calculation of $\hat{y}_i$, we estimated the unknown parameters α and β by the estimators a and b. Thus, the sample standard deviation of e_i, denoted by s_e, is:

$$s_e = \sqrt{\frac{\sum e_i^2}{n-2}} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n-2}}.$$

Let me show you here (but not in class) why the last equality (solely for ease of computation) holds:

$$\begin{aligned}
\sum(y_i - \hat{y}_i)^2 &= \sum(y_i - \overline{y} + \overline{y} - \hat{y}_i)^2 = \sum(y_i - \overline{y})^2 + \sum(\overline{y} - \hat{y}_i)^2 + 2\sum(y_i - \overline{y})(\overline{y} - \hat{y}_i) \\
&= \sum(y_i - \overline{y})^2 + \sum(\overline{y} - a - bx_i)^2 + 2\sum(y_i - \overline{y})(\overline{y} - a - bx_i) \\
&= \sum(y_i - \overline{y})^2 + \sum\{\overline{y} - (\overline{y} - b\overline{x}) - bx_i\}^2 + 2\sum(y_i - \overline{y})\{\overline{y} - (\overline{y} - b\overline{x}) - bx_i\} \\
&= \sum(y_i - \overline{y})^2 + \sum(b\overline{x} - bx_i)^2 + 2\sum(y_i - \overline{y})(b\overline{x} - bx_i) \\
&= \sum(y_i - \overline{y})^2 + b^2\sum(x_i - \overline{x})^2 - 2b\sum(y_i - \overline{y})(x_i - \overline{x}) \\
&= SS_{yy} + b^2 SS_{xx} - 2bSS_{xy} = SS_{yy} + b\frac{SS_{xy}}{SS_{xx}}SS_{xx} - 2bSS_{xy} \\
&= SS_{yy} - bSS_{xy}.
\end{aligned}$$

Of course from the last two rows we can see that $bSS_{xy} = b^2 SS_{xx}$.

Consequently, $100(1-\alpha)\%$ confidence intervals for the intercept and the slope, respectively, are

$$a \pm t_{\alpha/2} \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}},$$

$$b \pm t_{\alpha/2} \cdot \frac{s_e}{\sqrt{SS_{xx}}},$$

where the number of degrees of freedom of $t_{\alpha/2}$ is, of course, equal to $n-2$. Note that α in the above confidence interval formulae are not the intercept of the regression line but is, say, 0.05 if we consider a 95% confidence interval. It looks a bit confusing but it is the usual practice to denote the significance level, the loss in confidence and the intercept of the model of the linear regression by α (some authors will use β_0 to denote the intercept and β_1 to denote the slope in regression). However, if you understand what we are doing, in fact no confusion is possible.

Sometimes we are interested not only in the confidence interval for the slope or the intercept individually but also in the confidence interval for the predicted value of y when e.g. $x = x_0$. There are two formulae, one is for the confidence interval for estimating **the mean** of y when $x = x_0$:

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_{xx}}},$$

and another one for the confidence interval for predicting **an individual** y when $x = x_0$:

$$(a + bx_0) \pm t_{\alpha/2} \cdot s_e \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{SS_{xx}}}.$$

The difference in the formulae is easy to understand. Intuitively it is clear that if we want to predict an individual y, there is more uncertainty than the estimation of the mean, because an individual y will include a random error ε, which will however be gone when we take the mean of y. Thus, a 95% confidence interval for an individual y has to be wider than a 95% confidence interval for the mean of y, even though they will be centred at the same mid-point (i.e. they have the same point estimate). More precisely, for predicting an individual y, there is an error term ε added to the point $a + bx_0$ and the variance of ε is σ^2, whilst for the mean of y this extra error term vanishes, because the mean of ε is zero. Thus, when the variance of the estimated mean of y is $\sigma^2\{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{SS_{xx}}\}$, the variance of an individual predicted y is $\sigma^2\{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{SS_{xx}}\}$, which results from adding σ^2 (from ε in every individual y) to the variance of the estimated mean of y.

Note that the term $(x_0 - \overline{x})^2$ in these two formulae suggest that confidence intervals are the narrowest when $x_0 = \overline{x}$, i.e. when x_0 is somewhere in the middle of the observed x-values, and the further away x_0 from $\overline{x}$, the wider the confidence intervals, because $(x_0 - \overline{x})^2$ are getting larger. This also alerts us that extrapolation (extending the regression line beyond the range of observed x-values) is not recommended because confidence intervals would be very wide; moreover, we in fact do not know whether the relationship between x and y is still well-described by our regression line when we go beyond our observations; thus, extrapolation is risky.

We also discussed how to carry out hypothesis testing for α and β. To test the null hypotheses that $\alpha = \alpha_0$ and $\beta = \beta_0$, respectively, from the normal distributions of a and b given in (1) above, we can use the test statistics:

$$t = \frac{a - \alpha_0}{s_e\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SS_{xx}}}} \qquad \text{and} \qquad t = \frac{b - \beta_0}{\frac{s_e}{\sqrt{SS_{xx}}}},$$

respectively. Each of them, under the null hypothesis, has a t-distribution with degrees of freedom $n - 2$.

After estimating the regression line, we should determine how well such a line actually describes the relationship between x and y (i.e. how strong the relationship is). That is the purpose of the *coefficient of determination*, denoted by R^2.

A simple algebraic calculation shows that

$$\sum(y_i - \overline{y})^2 = \sum(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = \cdots = \sum(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \overline{y})^2. \qquad (2)$$

The *total sum of squares* $\sum(y_i - \overline{y})^2$ tells you the total variation of y. The value y is not a constant because there are two sources of variation, namely the vertical variation comes from random error ε and the horizontal variation comes from variation in the values of x. The first term on the right-hand side of equation (2), called the *residual sum of squares*, $\sum(y_i - \hat{y}_i)^2$, is the sum of the squares of the vertical distances (the residuals) from each y_i to the regression line and hence is the vertical variation. It is also known as *error sum of squares*, denoted by *SSE*. It is also the sum that we want to minimise in the least squares estimation of the slope and the intercept of the regression line. Thus, the smaller the residual sum of squares, the smaller the vertical variation and hence the better the regression line describing the relationship (no vertical variation means the line can perfectly describe the relationship, as all points are lying on the line). Since the sum of the two terms on the right-hand side of equation (2) is the total variation, which is fixed for a given sample, the smaller the first term, the larger the second term. Thus, the larger the second term on the right-hand side of equation (2), called the *regression sum of squares* (*SSR* for short), $\sum(\hat{y}_i - \overline{y})^2$, the better the regression line describing the relationship. However, the variation in $\hat{y}_i$ (i.e. the horizontal variation of y_i) is entirely under my 'control' because $\hat{y}_i$ are all lying on the regression line and their variation comes from and only from the variation in x_i (the vertical variation is entirely out of my 'control' because it comes from random errors, which are not controllable). That is, this *SSR* is the variation that can be explained by the regression line.

If we divide the regression sum of squares by the total sum of squares, we will get a nonnegative ratio that is at most 1:

$$R^2 := \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2},$$

which is called the *coefficient of determination*.

A large coefficient of determination means a strong relationship between x and y. Next week we will give a more precise interpretation of its numerical value and then introduce a related notion called coefficient of correlation.

Next week we will finish Chapter 13 in 30 minutes or so and then move to Chapter 15 (we will skip Chapter 14). The pdf file of Chapter 15 can be downloaded from the homepage of the book (though it is free to download from the publisher, I do not have the copyright to re-distribute it, and so please download the file directly from the publisher; the link can be found in the course homepage). Chapter 15 will be our last chapter! Though the last, this chapter will still introduce new statistical arguments and new tests developed from the same idea that we have acquired in previous chapters. Believe me, we still have a lot to learn! So, please make sure that you are not lagging behind too much.

Cheers,
Heng Peng