Weekly Review 13

As we can see from the signed-rank test, the rank sum test and the Kruskal–Wallis test, ranks are important when we talk about nonparametric methods. Let us now consider a nonparametric correlation, namely the *Spearman rank-correlation*, which is in fact the correlation between the ranks of two variables. Suppose we have a given set of n paired data. We rank the x's among themselves from low to high and then, independently, rank the y's among themselves also from low to high. If there is no tied observation, I explained to you briefly that the *Pearson* correlation coefficient r between ranks of x and ranks of y is equal to

$$\rho_{\mathrm{S}} = \frac{\sum(r_{x_i} - \overline{r}_x)(r_{y_i} - \overline{r}_y)}{\sqrt{\sum(r_{x_i} - \overline{r}_x)^2 \sum(r_{y_i} - \overline{r}_y)^2}}$$

$$= \frac{\sum_{i=1}^n r_{x_i} r_{y_i} - \frac{(\sum_{k=1}^n k)^2}{n}}{\sum_{k=1}^n k^2 - \frac{(\sum_{k=1}^n k)^2}{n}}, \tag{1}$$

because when the ranks r_{x_i} are all distinct, $\sum r_{x_i} = 1 + \cdots + n$ and $\sum r_{x_i}^2 = 1^2 + \cdots + n^2$. Using the (high school level) identities that

$$\sum_{k=1}^n k = \frac{n(n+1)}{2}, \quad \text{and} \quad \sum_{k=1}^n k^2 = \frac{n(n+1)(2n+1)}{6},$$

and the fact that

$$\sum_{i=1}^n (r_{x_i} - r_{y_i})^2 = 2\sum_{k=1}^n k^2 - 2\sum_{i=1}^n r_{x_i} r_{y_i},$$

after a few more algebraic steps, we can re-express the Spearman correlation between the ranks as

$$\rho_{\mathrm{S}} = \frac{\sum k^2 - \frac{(\sum k)^2}{n} - \frac{\sum d^2}{2}}{\sum k^2 - \frac{(\sum k)^2}{n}} = 1 - \frac{6\sum d^2}{n(n^2-1)}, \tag{2}$$

where each $d := d_i$ is the difference between the ranks of a pair. In this course, we will still use formula (2) even if there are tied observations[1].

To test the null hypothesis that the population Spearman rank-correlation is zero

$$\mathrm{H}_0 \colon \rho_{\mathrm{S}} = 0$$

[1]There are many other authors who suggested that in the presence of tied observations, expression (2) for r_{S} is not correct and the original expression (1) should be used. However, the difference is often small, and so here we simply stick to expression (2) even if there are ties.

we just compare the sample rank-correlation r_S with the critical values in Table XI.

When the sample size is large, we can use the same t-statistic as that for testing zero Pearson correlation in Chapter 13:

$$t = r_S \sqrt{\frac{n-2}{1-r_S^2}},$$

which under the null hypothesis of zero rank-correlation, follows the t-distribution with degrees of freedom $n-2$ approximately.

All the methods we studied in this course are based on the assumption that our samples are random, and one feature of a random sample is that the order of the observations does not have any special pattern. Consider a sequence of binary outcomes, such as "Heads" and "Tails", 0s and 1s, "above the median" and "below the median", "successes" and "failures" and so on. A way to test whether the arrangement is random or not is to consider the *total number of runs*, which is denoted by R. A run is a succession of identical symbols that is followed and preceded by different symbols or no symbols at all. If R is too small or too large, it indicates that there is a clustering or cyclical pattern, respectively. We reject the null hypothesis that the arrangement is random if R is too small or too large. Again, how small is too small and how large is too large? If n_1, the number of symbols of one kind, and n_2, the number of symbols of another kind, are small, then we use Tables XII on page 696 to find the critical values for R at the 0.05 significance level. If n_1 and n_2 are large, then the distribution of R can be approximated by the normal distribution with mean μ_R and variance σ_R^2, where the formulae for μ_R and σ_R^2 can be found on page 671, so that we can reject the null hypothesis if $z = (R - \mu_R)/\sigma_R$ is too large or too small. This is the so-called *runs test*, which is the last topic of the textbook.

In this course, we have covered all chapters, except Chapter 14, of the textbook. The first three chapters tell us how to summarise data by using various measures of locations and measures of variation, as well as graphical and tabular presentations. All these things have been taught at primary school and high school level but in this course we re-considered them from a different perspective. We explained the idea behind formulae, graphs and tables so that we now know why they are useful and what they are telling us. The first three chapters form the basis of what we call *Descriptive Statistics*.

However, if we want to do *Statistical Inference*, i.e. to generalise our findings from a sample to the population from which we got our sample, we have to acquire some basic knowledge and skills in the context of *Probability Theory* first. Thus, the focus of Chapter 4–6 shifts from Statistics to Probability. Chapter 4 explains what probability is and introduces some elementary calculation rules. An important notion is the conditional probability and one of the most famous theorems in conditional probability is the *Bayes Theorem*. Chapters 5 and 6 introduce the notion of discrete and continuous random variables and their distributions. Of particular interest are

- the binomial distribution,

- the hypergeometric distribution,
- the geometric distribution,
- the negative binomial distribution,
- the Poisson distribution,
- the uniform distribution,
- the normal distribution.

The normal distribution is perhaps the most important distribution in Statistics.

Chapter 7 builds a bridge from Probability to Statistics, and the name of the bridge is the magic word: **distribution**, or more precisely, *sampling distribution*. Since a sample is random, the sample mean and any other statistics are random and so they have their distributions, which are all called sampling distributions. The beautiful central limit theorem says that if the sample size is large, then the sampling distribution of the sample mean is approximately a normal distribution.

The above materials form our starter and the main course of the meal started from Chapter 8. After learning various distributions, we are now able to construct formulae for confidence intervals for μ, σ and p. In addition to confidence interval, Chapter 9 to Chapter 12 discuss a lot of different hypothesis testing procedures. The general procedure for hypothesis testing is as follows.

(i) We write down the null hypothesis and the alternative hypothesis,

(ii) fix α,

(iii) calculate the value of the corresponding test statistic and/or the p-value, and then

(iv) reject the null hypothesis

  (a) if the test statistic is in the critical region:
    - for a right-sided test, greater than or equal to the one-tailed right-sided (or upper) critical value;
    - for a left-sided test, less than or equal to the one-tailed left-sided (or lower) critical value;
    - for a two-sided test, either
      - greater than or equal to the two-tailed right-sided critical value or
      - less than or equal to the two-tailed left-sided critical value,
  (b) or if the p-value (the calculation of which depends on the alternative hypothesis) is less than or equal to α.

For each of the three parameters μ, σ and p, we have learnt

- one-sample tests, and
- two-independent sample tests.

In particular, for μ, we have also learnt

- the paired-sample t-test, and
- one-way ANOVA for k independent samples.

For p, we have discussed also the χ^2-test for

- goodness-of-fit,
- independence (of two attributes) in an $R \times C$ table,
- homogeneity (of distributions in different populations) in an $R \times C$ table.

The above tests require the assumption that the populations from which the samples are drawn follow normal distributions, and these tests are parametric because we do have parameters. Chapter 15 introduces *nonparametric tests* for hypotheses concerning:

- one median (one sample or paired-sample): sign test and signed-rank test,
- two medians (two independent samples): rank-sum test, and
- k medians (k independent samples): Kruskal–Wallis H-test.

In Chapter 13, we studied the *simple linear regression*

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

which is a model for paired data. Under this model, we have two parameters α and β (the same as μ and σ, these parameters are denoted also by Greek alphabets). We learnt how to estimate the slope β and the intercept α by the least squares estimation (LSE). Under the normal distribution assumption for the random ε, the distributions of these two estimators are also normal, and hence we are able to construct confidence intervals and test statistics. We also discussed how to measure the strength of the linear relationship between x and y by the *coefficient of determination* and the *(Pearson) coefficient of correlation*.

In the regression analysis, the variable x is not considered as random. However, the idea of correlation can also apply to the case that both x and y are normally distributed variables

and a t-test for testing zero Pearson correlation was introduced. A nonparametric version of the correlation, the *Spearman rank-correlation*, was introduced in Chapter 15.

Finally, we considered how to test whether or not the arrangement of binary data (e.g. `H` and `T`, or `above` and `below`) is random by the *runs test*.

That's all I want you to learn in this course. I hope you have enjoyed this course and wish you a successful final exam.

Cheers,
Heng Peng

— T H E E N D —

P.S. You can find the following formulae in the final exam paper.

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i} {x_{ij}}^2 - \frac{{T_{..}}^2}{\sum_i n_i}, \qquad SSB = \sum_{i=1}^{k}\frac{{T_{i.}}^2}{n_i} - \frac{{T_{..}}^2}{\sum_i n_i}, \qquad SSW = \sum_{i=1}^{k}\sum_{j=1}^{n_i} {x_{ij}}^2 - \sum_{i=1}^{k}\frac{{T_{i.}}^2}{n_i}$$

$$SS_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}, \qquad SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}, \qquad SS_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n},$$

$$b = \frac{SS_{xy}}{SS_{xx}}, \qquad s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}}, \qquad t = \frac{b-\beta}{s_e/\sqrt{SS_{xx}}},$$

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}, \qquad t = r\sqrt{\frac{n-2}{1-r^2}}, \qquad H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(n+1),$$

$$r_S = 1 - \frac{6(\sum d^2)}{n(n^2-1)}, \qquad z = r_S\sqrt{n-1}, \qquad \chi^2 = \sum\frac{(O-E)^2}{E} = \sum\frac{O^2}{E} - \sum O.$$

$$\mu_{T^+} = \frac{n(n+1)}{4}, \qquad \mu_T = \frac{n_1(n_1+n_2+1)}{2}, \qquad \mu_R = \frac{2n_1n_2}{n_1+n_2} + 1,$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}, \qquad \sigma_T = \sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}, \qquad \sigma_R = \sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}},$$

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