Weekly Review 10

Last week we introduce the χ^2-distribution for the goodness-of-fit test, which is a one-population problem. Section 11.3 discusses the χ^2-test for the independence of two variables or the homogeneity of distributions in two or more populations. The data in either case can be tabulated as a so-called $R \times C$ *(contingency) table*.

Let's consider test of independence first. Under the null hypothesis that the two variables are independent, we know that $p_{ij} = p_{i\cdot}p_{\cdot j}$, where p_{ij} is the proportion of cases belong to the i^{th} row and j^{th} column, $p_{i\cdot}$ is the proportion of cases belong to the i^{th} row and $p_{\cdot j}$ is the proportion of cases belong to the j^{th} column. The independence null hypothesis means that the joint probability p_{ij} is equal to the product of two individual probabilities $p_{i\cdot}$ and $p_{\cdot j}$ and hence allows us to estimate the joint probability by estimates of two individual probabilities:

$$\hat{p}_{ij} = \hat{p}_{i\cdot} \times \hat{p}_{\cdot j} = \frac{i^{\text{th}} \text{ row total}}{\text{grand total}} \times \frac{j^{\text{th}} \text{ column total}}{\text{grand total}}.$$

To test the null hypothesis that the two variables are independent, we can calculate the expected count (under the null hypothesis, of course) for the i-j cell in the table by

$$E_{ij} = \text{grand total} \times \hat{p}_{ij} = \frac{i^{\text{th}} \text{ row total} \times j^{\text{th}} \text{ column total}}{\text{grand total}}.$$

After we get all the expected counts, we combine the information of the differences between the observed counts and the expected counts (under the null hypothesis) by the following test statistic:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum \frac{(O - E)^2}{E},$$

which has approximately a χ^2-distribution. This approximation works well if all E's are at least 5. I explained that the number of the degrees of freedom is $(R - 1)(C - 1)$.

For homogeneity, the null hypothesis states that either (i) for each fixed j, p_{ij}'s are the same for all i (where the i^{th} row represents the data from the i^{th} population), or (ii) for each fixed i, p_{ij}'s are the same for all j (then it means that the j^{th} column represents the data from the j^{th} population). These two statements are interchangeable because which variable is the row variable, and consequently which variable is the column variable, is arbitrarily chosen. In my lecture I used the formulation in case (i). Then under the homogeneity null hypothesis, for each fixed j (i.e. for the j^{th} column), $p_{1j} = p_{2j} = \cdots = p_{Rj}$ can be estimated by j^{th} column total divided by the grand total, and hence the expected count in the i-j position is the product of the estimated probability $\hat{p}_{\cdot j}$ and the i^{th} row total, i.e.

$$E_{ij} = \frac{j^{\text{th}} \text{ column total}}{\text{grand total}} \times i^{\text{th}} \text{ row total} = \frac{i^{\text{th}} \text{ row total} \times j^{\text{th}} \text{ column total}}{\text{grand total}},$$

which is the same formula as in the test of independence. The rest of the testing procedure is also the same and so I do not repeat here.

That is, for a given $R\times C$ table, no matter we test independence (when we have one sample in which everyone has two attributes) or homogeneity (when we have many populations and one sample from each population), except the null hypothesis and the alternative hypothesis, the testing procedures are exactly the same.

If we use our pocket calculator, the formula

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

is not too convenient because each calculated E has to be used twice (and E is not necessarily an integer). To make the calculation a bit less laborious, we may use the formula

$$\chi^2 = \sum \frac{O^2}{E} - n,$$

where $n = \sum O = \sum E$ is the grand total. In this expression each expected count will be used only once, leading to fewer calculation steps.

The χ^2-distribution has another important application, which comes from the mathematical fact that if the population has $\underline{\text{a normal distribution with variance } \sigma^2}$, then the distribution of the sample variance can be expressed in terms of:

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2\text{-distribution}, \tag{1}$$

where the degrees of freedom are $n-1$. A normal distribution population means that in a random sample $\{X_1, \dots, X_n\}$ of size n, each X_i follows $\mathrm{N}(\mu, \sigma^2)$. If this is true, then (as we mentioned before) $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n})$ $\underline{\text{exactly}}$ (i.e. this is not a large-sample approximation by the central limit theorem but is the true distribution of $\overline{X}$ even without taking the limit $n \to \infty$).

Suppose we want to estimate the population variance σ^2. The point estimate is naturally the sample variance s^2. In order to construct confidence intervals for σ^2, we need to know the sampling distribution of s^2. Now we know its distribution, and so it is straightforward to derive the formula of confidence interval. A $100(1-\alpha)\%$ confidence interval for σ^2 is simply

$$\chi^2_{1-\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{\alpha/2}$$

$$\Rightarrow \quad \frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}.$$

Note that because the normal distribution is symmetric, we have $z_{1-\alpha/2} = -z_{\alpha/2}$, but since s^2 is nonnegative, the χ^2-distribution is not symmetric and so $\chi^2_{1-\alpha/2} \neq -\chi^2_{\alpha/2}$. Hence, the values for $\chi^2_{1-\alpha/2}$ for various α are tabulated together with $\chi^2_{\alpha/2}$ in the χ^2-table.

Suppose we want to test

$$\mathrm{H}_0\colon \sigma^2 = \sigma_o^2,$$

against one of the following three different alternative hypotheses:

$$\mathrm{H}_A\colon \sigma^2 \neq \sigma_o^2, \qquad \mathrm{H}_A\colon \sigma^2 > \sigma_o^2, \qquad \text{or} \qquad \mathrm{H}_A\colon \sigma^2 < \sigma_o^2.$$

Again, what we need is the distribution given in (1). Thus, the natural test statistic is

$$\chi^2 = \frac{(n-1)s^2}{\sigma_o^2},$$

and if the null hypothesis is true, we know the distribution of the test statistic χ^2 and we know how large the observed χ^2 is too large and how small is too small. The decision rule can be tabulated as:

H_0	H_A	Reject H_0 at the α significance level if
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$\chi^2 \geq \chi^2_\alpha$
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \geq \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$\chi^2 \leq \chi^2_{1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 \geq \chi^2_{\alpha/2}$ or $\chi^2 \leq \chi^2_{1-\alpha/2}$

where $\chi^2_{1-\frac{\alpha}{2}}$, $\chi^2_{1-\alpha}$, χ^2_α and $\chi^2_{\alpha/2}$ all have the same degrees of freedom, namely, $n-1$.

Note that this decision rule table is not the same as that for z or t; we need two different critical values for χ^2 in a two-sided test. This has been explained above when we discussed confidence intervals for σ^2. Note also unlike the goodness-of-fit or the $R \times C$ table, the χ^2-test here is not one-sided but two-sided for a two-sided alternative hypothesis.

If we want to perform a two-independent-sample t-test, we have to know whether the two unknown variances are the same or not. Thus, when we have two independent samples from two populations, it is a very standard procedure to test

$$\mathrm{H}_0\colon \sigma_1^2 = \sigma_2^2,$$

against one of the following three different alternative hypotheses:

$$\mathrm{H}_A\colon \sigma_1^2 \neq \sigma_2^2, \qquad \mathrm{H}_A\colon \sigma_1^2 > \sigma_2^2, \qquad \text{or} \qquad \mathrm{H}_A\colon \sigma_1^2 < \sigma_2^2.$$

Assume the populations have normal distributions. If the null hypothesis is true, then

$$F = \frac{s_1^2}{s_2^2} \sim F_{n_1-1,\, n_2-1}\text{-distribution},$$

where the first subscript indicates that the numerator of F has $n_1 - 1$ degrees of freedom and the second subscript indicates that the denominator has $n_2 - 1$ degrees of freedom. Once we know the distribution, we can find the critical values from the corresponding table to determine whether an observed F is too large or is too small. The F-distribution is obviously not symmetric, as it is a ratio of two nonnegative values. Thus, we need two critical values for a two-sided test, and the decision rule for the F-distribution has the same form as that for the χ^2-distribution, i.e.

H_0	H_A	Reject H_0 at the α significance level if
$\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$F \geq F_{\alpha,\, n_1-1,\, n_2-1}$
$\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$F \leq F_{1-\alpha,\, n_1-1,\, n_2-1}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F \geq F_{\alpha/2,\, n_1-1,\, n_2-1}$ or $F \leq F_{1-\alpha/2,\, n_1-1,\, n_2-1}$

Interestingly, if we want to know whether s_1^2/s_2^2 is too small, it is equivalent to ask whether s_2^2/s_1^2 is too large, where

$$\frac{s_2^2}{s_1^2} \sim F_{n_2-1,\, n_1-1}\text{-distribution}.$$

Thus, we do not need a separate table for $F_{1-\alpha}$ because

$$F_{1-\alpha,\, df_1,\, df_2} = \frac{1}{F_{\alpha,\, df_2,\, df_1}},$$

(please pay particular attention to the change of the two values of the degrees of freedom) or equivalently, we have

$$\frac{s_1^2}{s_2^2} \leq F_{1-\alpha,\, n_1-1,\, n_2-1} \quad \Leftrightarrow \quad \frac{s_2^2}{s_1^2} \geq F_{\alpha,\, n_2-1,\, n_1-1}.$$

Thus, the decision rule table can be re-written as

H_0	H_A	Reject H_0 at the α significance level if
$\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$	$\frac{s_1^2}{s_2^2} \geq F_{\alpha, n_1-1, n_2-1}$
$\sigma_1^2 = \sigma_2^2$ or $\sigma_1^2 \geq \sigma_2^2$	$\sigma_1^2 < \sigma_2^2$	$\frac{s_2^2}{s_1^2} \geq F_{\alpha, n_2-1, n_1-1}$
$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$\frac{s_1^2}{s_2^2} \geq F_{\alpha/2, n_1-1, n_2-1}$ or $\frac{s_2^2}{s_1^2} \geq F_{\alpha/2, n_2-1, n_1-1}$

Be careful with the order of the two values of the degrees of freedom in the critical values. Also, as I mentioned, for the two-sided alternative, we only have to consider the ratio of the larger sample variance to the smaller sample variance, i.e. consider only the larger one between $\frac{s_1^2}{s_2^2}$ and $\frac{s_2^2}{s_1^2}$, because the smaller one is always not greater than 1 and so of course is not greater than the critical value $F_{\alpha/2}$, which is always greater than 1.

Note that both the χ^2-test and the F-test are not *robust*, i.e. they are sensitive to departures from the normality assumption. That is to say, inferences drawn from samples (no matter n is small or large) can be seriously misleading when the population distribution departs from normality. Thus, if the population does not have a normal distribution, we cannot use the χ^2-distribution or the F-distribution. (In contrast, the t-test is robust when the sample is reasonably large.)

Chapter 12 is devoted to a famous statistical analysis, called the *Analysis of Variance*, or *ANOVA* for short. The problem to tackle is: for k populations following normal distributions with population means μ_1, μ_2, ..., μ_k, we test at the α significance level the hypothesis that all the means are the same. That is to say, we test

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

The ANOVA is famous and widely used because in many contexts we will face the following situation: we apply k different treatments (e.g. different medicines) to k groups, assuming each group is a sample taken independently from the same population (e.g. patients suffering from the same illness) so that if there are any differences between groups, the differences are caused by the different treatments they received. We discussed that if we want to (actually, we have to) control the type I error probability for this yes-no question, just one question, it is more appropriate to test the null hypothesis by just one test, instead of testing the equality of each pair of population means by applying repeatedly the two-independent-sample t-tests. [How many times? Ans: $\binom{k}{2}$.] Thus, we need a new method.

Some weeks ago we already discussed how to test the above hypotheses if $k = 2$ by using the two-independent-sample t-test, in which we have to ask whether the two variances σ_1^2 and

σ_2^2 are the same or not; if they are the same, we simply pool the data to get one estimate of the common unknown variance, otherwise we have to estimate two unknown variances and have to use an approximation formula for the degrees of freedom.

For $k \geq 3$, we still have to know whether the variances are the same or not. Here we only consider the case that all variances are the same, and we still have to estimate this common value. The idea of the analysis is to estimate the common variance σ^2 by two methods, one assumes the null hypothesis is true while the other does not make such an assumption, and then we compare these two estimates to see if they differ significantly. That's why this approach is called analysis of variance.

Now, given k samples of data, for each sample we calculate its sample mean and sample variance. Then we can estimate the unknown common variance by the following argument.

First, assume the null hypothesis is true, the k sample means will then have the same mean μ (because of the null hypothesis) and the same variance σ^2 (assumption used in ANOVA), and hence the same (normal) distribution. We know from the Central Limit Theorem that for a sample of size n taken from the population $\mathrm{N}(\mu, \sigma^2)$, the sample mean will follow:

$$\overline{X} \sim \mathrm{N}\big(\mu, \frac{\sigma^2}{n}\big).$$

[Be careful: $\mathrm{N}(\mu, \sigma^2)$ is the distribution of the population, while $\mathrm{N}\big(\mu, \frac{\sigma^2}{n}\big)$ is the distribution of $\overline{X}$.] So, if we have many sample means $\{\overline{x}_1, \ldots, \overline{x}_k\}$ of independent samples of size n from the population $\mathrm{N}(\mu, \sigma^2)$, then the sample variance, denote by $s^2_{\overline{x}}$, of these sample means will be an estimate of σ^2/n (remember: though the statement is true no matter what the value of μ is, this is true only if all μ_k's are equal). That is to say, if the null hypothesis is true, then $s^2_{\overline{x}}$ is an estimate of σ^2/n; in other words, $ns^2_{\overline{x}}$ is an estimate of σ^2.

Second, we calculate the sample variance for each of the k samples and get k sample variances $\{s_1^2, \ldots, s_k^2\}$. Each of these k sample variances is an estimate of σ^2, no matter whether the population means are the same or not. Because the sample sizes of these k samples are the same, each s_i^2 is equally trustworthy. Thus, we take the mean of them, denoted by $\overline{s^2}$, and this mean will be another estimate of σ^2, which does not require the assumption that all μ_k's are equal.

We presented the numerical calculation of Example 12–2, in which we estimate the common variance by the two different methods explained above. Next week we will explain how to use these two different estimates of σ^2 to decide whether we reject H_0 or not and then generalise the above technique to the case that the sample sizes are different. Afterwards, we will move to the next chapter for *linear regression*, which is the starting point of all, or almost all, advanced statistical analyses.

Cheers,
Heng Peng