Anna Janicka

Mathematical Statistics 2020/2021
Lecture 11

1. Hypothesis Testing – Examples of LR tests, cont.

Continuing our analysis of examples of LR tests, we will now turn to cases where we have more than one population sample to research.

1.2. **Two populations.** In the case where we have two populations (two samples), we might be interested in whether these two populations have the same characteristics (means, variances...). The types of models most commonly used in these cases are similar to the models used for the single population case where we compare with an external value; the test statistics are slightly different, however.

1.2.1. *Model I.* Lest us first assume that we have a random sample $X_1, \ldots, X_n$ from a normal distribution with parameters μ_X and σ_X^2, and a random sample $Y_1, \ldots, Y_n$ from a normal distribution with parameters μ_Y and σ_Y^2, where σ_X^2 and σ_Y^2 **are known**. Let us assume that we want to test the null hypothesis that $\mu_X = \mu_Y$, against different alternatives. We will use a test statistic slightly modified with respect to the single sample case:

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n_x + \sigma_Y^2/n_Y}},$$

which under the null hypothesis has a standard normal distribution, to construct critical regions in the following way:

- If the alternative is that $\mu_X > \mu_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : U(x) > u_{1-\alpha}\}$$
- If the alternative is that $\mu_X < \mu_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : U(x) < -u_{1-\alpha} = u_\alpha\}.$$
Please note, however, that this case is redundant, since we can just change the order of the samples and use the previous case instead. For this reason, in the models that follow, we will omit this type of alternative.
- If the alternative is that $\mu_X \neq \mu_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : |U(x)| > u_{1-\alpha/2}\},$$

where u_p signifies the quantile of rank p of the standard normal distribution.

Example:

(1) Suppose we have a random sample of 10 observations from a normal distribution with an unknown mean μ_1 and variance equal to 11^2, and 10 observations from a normal distribution with an unknown mean μ_2 and variance equal to 13^2. Let us assume that the average in the first sample amounts to 501, while the average in the second sample amounts to 498. Clearly, these two empirical averages differ. But does this mean that the means of the distributions differ, too? Or is this observed difference just due to pure chance?

Suppose that we wish to verify whether the means of the distributions are equal, at a significance level $\alpha = 0.05$. The value of the appropriate test statistic amounts to $U = \frac{501-498}{\sqrt{\frac{13^2}{10}+\frac{11^2}{10}}} \approx 0.557$. The critical value for a two-sided test at the $\alpha = 0.05$ significance level amounts to $u_{0.975} \approx 1.96$. The value of the test statistic does not fall into the critical region of $(-\infty, -1.96) \cup (1.96, \infty)$, so we do not have grounds to reject the null. We could also note that the p-value of the result amounts to $2 \cdot \Phi(-0.557) \approx 2 \cdot 0.289 = 0.578$. Since the p-value is larger than the adopted significance level, we do not have grounds to reject the null. In this case, we found

that the observed difference in empirical averages would be a very common outcome if the distributions were as specified by the null hypothesis. Therefore, we do not have any grounds to claim that the two theoretical means aren't in fact equal.

If we wished to verify the null hypothesis that the means are equal, against the alternative that the mean in the second sample is smaller at the same significance level of $\alpha = 0.05$, the result would stay the same: the value of the test statistic 0.557 does not fall into the critical region for the one-sided test, which is equal to $(1.64 = u_{0.95}; \infty)$. This time, the p-value of the result is equal to $1 - \Phi(0.557) \approx 0.289$, so it is smaller than in the case of the two-sided test, but again – this result is far from any extremity, so we do not have grounds to reject the null.

1.2.2. *Model II.* Lest us now assume that we have a random sample $X_1, \ldots, X_n$ from a normal distribution with parameters μ_X and σ^2, and a random sample $Y_1, \ldots, Y_n$ from a normal distribution with parameters μ_Y and σ^2, where σ^2 **is unknown, but assumed to be the same for both samples**. Let us further assume that we want to test the null hypothesis that $\mu_X = \mu_Y$, against different alternatives. We will use a test statistic:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}} \sqrt{\frac{n_X n_Y}{n_X + n_Y}(n_X + n_Y - 2)},$$

which under the null hypothesis has a t-Student distribution with $n_X + n_Y - 2$ degrees of freedom, to construct critical regions in the following way:

- If the alternative is that $\mu_X > \mu_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : T(x) > t_{1-\alpha}(n_X + n_Y - 2)\}$$
- If the alternative is that $\mu_X \neq \mu_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : |T(x)| > t_{1-\alpha/2}(n_X + n_Y - 2)\},$$

where $t_p(n_X + n_Y - 2)$ is the quantile of rank p of the t-Student distribution with $n_X + n_Y - 2$ degrees of freedom, and S_X^2 and S_Y^2 are unbiased estimators of the variance for the sample of Xs ad Ys, respectively.

The test statistic used in this case might be rearranged slightly, to become

$$T = \frac{\bar{X} - \bar{Y}}{S_* \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

where

$$S_*^2 = \frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}$$

is an estimator of the variance based on both samples jointly. This second form of the formula shows that the philosophy of the means testing procedure: take the difference of means and standardize by dividing by the standard deviation, is the same in all cases. If we assume that the two samples have the same variances, the S_*^2 estimator has a chi-squared distribution, just like in the single sample case, and it is easy to describe the distribution of the test statistic. However, if we could not assume that the variances were equal, and we allowed $\sigma_X^2 \neq \sigma_Y^2$, we would have a big problem. We would still be able to calculate the test statistic, but it would be impossible to find a general formula to describe the distribution of this test statistic without using the values of the unknown parameters σ_X^2 and σ_X^2 in some form (which obviously we cannot do, since they are assumed to be unknown).

Since in this model we must assume that the variances in the two populations are equal, we might wish to verify this assumption. Let us assume that we want to test $\sigma_X^2 = \sigma_Y^2$, against different types of alternatives. In this case, we can use a test statistic

$$F = \frac{S_X^2}{S_Y^2},$$

where F has the Fisher distribution (also called the Fisher-Snedecor or F distribution) with $n_X - 1$ and $n_Y - 1$ degrees of freedom, to construct critical regions in the following way:

- If the alternative is that $\sigma_X^2 > \sigma_Y^2$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : F > F_{1-\alpha}(n_X - 1, n_Y - 1)\}$$
- If the alternative is that $\sigma_X^2 \neq \sigma_Y^2$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : F > F_{1-\alpha/2}(n_X - 1, n_Y - 1) \vee F < F_{\alpha/2}(n_X - 1, n_Y - 1)\}$$

where $F_p(n, m)$ is the quantile of rank p of the F distribution with n and m degrees of freedom, and S_X^2 and S_Y^2 are unbiased estimators of the variance for the sample of Xs ad Ys, respectively.

1.2.3. *Model III.* Lest us now assume that we have a random sample $X_1, \ldots, X_n$ from a distribution with mean μ_X and variance σ_X^2, and a random sample $Y_1, \ldots, Y_n$ from a distribution with mean μ_Y and variance σ_Y^2, where σ_X^2 and σ_Y^2 **are not known and not assumed equal**. As we have stated above, in general it is not possible to test the equality of means in this case, even if we knew that the distributions were normal. However, if we want to test the null hypothesis that $\mu_X = \mu_Y$, against different alternatives and we have a **large sample size**, we might use a test statistic slightly modified with respect to the first model:

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/n_X + S_Y^2/n_Y}},$$

which under the null hypothesis has (for large sample sizes) approximately a standard normal distribution, to construct critical regions in the following way:

- If the alternative is that $\mu_X > \mu_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : U(x) > u_{1-\alpha}\}$$
- If the alternative is that $\mu_X \neq \mu_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : |U(x)| > u_{1-\alpha/2}\},$$

where u_p signifies the quantile of rank p of the standard normal distribution, and S_X^2 and S_Y^2 are unbiased estimators of the variance for the sample of Xs ad Ys, respectively.

1.2.4. *Model IV.* Again, as a special case of model III, we might consider two-point distributions and compare fractions. If we assume that the random variables $X_1, \ldots, X_n$ that we observe come from a distribution such that $P(X = 1) = p_X = 1 - P(X = 0)$, and $Y_1, \ldots, Y_n$ come from a distribution such that $P(Y = 1) = p_Y = 1 - P(Y = 0)$, and we are to test the null hypothesis that $p_X = p_Y$ against different types of alternatives, we might use the test statistic provided in Model III with a modified estimator of the variance:

$$U^* = \frac{\frac{X}{n_X} - \frac{Y}{n_Y}}{\sqrt{p_*(1 - p_*)(\frac{1}{n_X} + \frac{1}{n_Y})}},$$

where

$$p_* = \frac{X + Y}{n_X + n_Y}$$

is an estimator of the fraction based on both samples simultaneously. Under the null hypothesis, for large sample sizes the test statistic U^* has an approximate standard normal distribution, which allows us to construct the following critical regions

- If the alternative is that $p_X > p_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : U^*(x) > u_{1-\alpha}\}$$
- If the alternative is that $p_X \neq p_Y$, then the critical region of the test for significance level α is equal to
$$C^* = \{x : |U^*(x)| > u_{1-\alpha/2}\},$$

where u_p signifies the quantile of rank p of the standard normal distribution.

1.3. **Three or more populations – ANOVA.** What if we wanted to compare more than two populations simultaneously? The naive, simplest approach would be to compare all populations in pairs, and reject the null hypothesis if in any of the comparisons the decision was to reject the null hypothesis. In such a case, however, we do not control the significance level of the test. This is because the probability of incorrectly rejecting the null hypothesis is larger than the significance level adopted in the pairwise comparisons. To see this, let us assume that we have three populations, in which in reality the parameters under study are equal to each other. Let us also assume that we perform three pairwise tests for equality of parameters for each pair of populations, such that the significance level for each pairwise test is equal to α. In such a case, the chance that we will conclude that not all are the same is equal to the probability that in at least one case we will conclude that we should reject the null hypothesis. This probability may be calculated from the complementary event, i.e. as 1 minus the chance that in all three tests we will not find evidence against the null hypothesis. Assuming that the results of the pairwise tests are independent (which is a simplifying assumption!), we have that the probability of committing an error of the first type in the whole procedure amounts to $1-(1-\alpha)^3 = \alpha(1+\alpha+\alpha^2)$ and is not equal to α but larger than alpha. If the results of the pairwise tests are not independent (and we may expect them not to be independent), we do not know what this probability amounts to. Therefore, the procedure of a sequence of pairwise testing is not a good procedure.

Instead, if we want to check whether the means in more than two populations are equal, we may apply a procedure called the analysis of variance (ANOVA). Let us assume that we have samples from k populations, expressed as:

$$\begin{array}{cccc} X_{1,1}, & X_{1,2}, & \ldots & X_{1,n_1} \\ X_{2,1}, & X_{2,2}, & \ldots & X_{2,n_2} \\ & \ldots & & \\ X_{k,1}, & X_{k,2}, & \ldots & X_{k,n_k} \end{array}$$

where n_i is the number of observations in the i-th sample. Let us assume that all variables $X_{i,j}$ are independent, and that we have $X_{i,j} \sim N(\mu_i, \sigma^2)$ – the observations form the i-th sample have a normal distribution with mean μ_i, and all observations come from distributions having the same variance σ^2. All parameters ($\mu_1, ..., \mu_k$ and σ^2) are unknown. If $n = n_1 + \ldots + n_k$ is the overall sample size, we may test the null hypothesis that

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$$

(all means are equal to each other) against the alternative

$$H_1 : \neg H_0$$

(not all means are equal to each other) using a test statistic

$$F = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i,j} - \bar{X}_i)^2/(n-k)},$$

which, under the null hypothesis, has an F distribution with $k-1$ and $n-k$ degrees of freedom. This means that a test with a critical region

$$C^* = \{x : F(x) > F_{1-\alpha}(k-1, n-k)\},$$

where $F_{1-\alpha}(k-1, n-k)$ is the quantile of rank $1-\alpha$ of the F distribution with $k-1$ and $n-k$ degrees of freedom is the test that we were looking for, at a significance level α. For $n=2$, this test is equivalent to the two-population test (Model II) above.

The philosophy behind this test is as follows. We examine the variance in the whole sample (consisting of observations from all subsamples). We decompose the overall variance (or rather sum of squares) into two components: one coming from the variability within each subsample (within-group), and one reflecting the variability between samples (between-group). It may be shown that

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j}-\bar{X})^2}_{\text{sum of squares (SS)}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})^2}_{\text{sum of squares between (SSB)}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{X}_{i,j}-\bar{X}_i)^2}_{\text{sum of squares within (SSW)}}.$$

Therefore, the test statistic that we use has the form:

$$F = \frac{\text{SSB}/(k-1)}{\text{SSW}/(n-k)},$$

and we reject the null if the sum of squares between groups "dominates" over the sum of squares within groups (after scaling), meaning that the variability that we see in the data comes more from variability between groups than within groups. Note that if in our sample all subsample means are exactly equal to each other, then the numerator of the test statistic is zero, and the value of the test statistic is also zero. If the means start differing from each other, the numerator increases; if it becomes too large (with respect to the differences that we see within the particular samples), we reject the null.

Example

(1) Let us assume that we study the yearly chocolate consumption of milk chocolate bars of inhabitants of three cities: A, B and C. We base our reasoning on a sample of $n_A = 8$ observations from city A, $n_B = 10$ observations from city B and $n_C = 9$ observations from city C. Suppose that the average consumption levels amount to 11, 10 and 7 milk chocolate bars for cities A, B and C, respectively, while the sample variances amount to 3.5, 2.8 and 3, respectively. Assuming that the yearly consumption of milk chocolate bar follows a normal distribution with equal variances in the cities under study, we may verify whether the average consumption level depends on the city using an analysis of variance test (we will use a 1% significance level). We have:
• the average consumption level in the whole sample amounts to
$\bar{X} = \frac{1}{27}(11 \cdot 8 + 10 \cdot 10 + 7 \cdot 9) \approx 9.3$;
• the sum of squares between groups amounts to
$SSB = (11-9.3)^2 \cdot 8 + (10-9.3)^2 \cdot 10 + (7-9.3)^2 \cdot 9 = 75.63$;
• the sum of squares within groups amounts to
$SSW = 3.5 \cdot 7 + 2.8 \cdot 9 + 3 \cdot 8 = 73.7$.

Therefore, the value of the test statistic that we should use in this case amounts to

$$F = \frac{75.63/(3-1)}{73.7/(27-3)} \approx 12.31.$$

Given that the critical value of the ANOVA test is equal to the quantile of the F distribution with 2 and 24 degrees of freedom of rank 0.99: $F_{0.99}(2, 24) \approx 5.61$, we have that the value of the test statistic falls into the critical region $(5.61; \infty)$, and hence we reject the null hypothesis of the equality of means in the three groups considered. This means that we can't claim that in all cities the average yearly consumption of milk chocolate bars is the same. In at least one of the cities considered, the average is different.