476 Statistics, Spring 2022.
Lecturer: Sergey Nadtochiy.

Lecture 9. ANOVA test, analysis of categorical data. (Sections 13.1–13.3, 14.1–14.3)

1 ANOVA test

Consider k independent normal samples

$$\{Y_{1,1},\dots,Y_{1,n_i}\},\dots,\{Y_{k,1},\dots,Y_{k,n_k}\},$$

with the same variance σ^2 and with the means $\mu_1,\dots,\mu_k$.

Q 1. *How to test whether $\mu_1=\dots=\mu_k$ or not?*

More precisely, we test the hypotheses

$$H_0=\{\mu_1=\dots=\mu_k\},\quad H_a=\{\text{there exist at least two } i\neq j \text{ such that } \mu_i\neq\mu_j\}.$$

The ANOVA (ANalysis of VAriance) test of the above hypotheses is given by the test statistic

$$V=\frac{\frac{1}{k-1}\sum_{i=1}^k n_i(\bar{Y}_i-\bar{Y})^2}{\frac{1}{n-k}\sum_{i=1}^k\sum_{j=1}^{n_i}(Y_{i,j}-\bar{Y}_i)^2},\quad n=n_1+\dots+n_k,\quad \bar{Y}_i:=\frac{1}{n_i}\sum_{j=1}^{n_i}Y_{i,j},\quad \bar{Y}:=\frac{1}{n}\sum_{i=1}^k\sum_{j=1}^{n_i}Y_{i,j},$$

and a rejection region of the form

$$RR=[r,\infty),$$

where r is determined by the condition $\mathbb{P}^0(V\in RR)=\alpha$.

The higher is the value of V, the less likely it is that H_0 is correct. Indeed, under the null hyp., every $\bar{Y}_i$ should be close to $\bar{Y}$ (as both are close to the true mean), while the denominator is not expected to get smaller as it measures σ^2. Thus, large values of the test stat. V indicate deviation from H_0.

Exercise 1. *Show that the test statistic of the likelihood ratio test for $H_0=\{\mu_1=\dots=\mu_k\}$ vs. $H_a=\{$there exist at least two $i\neq j$ such that $\mu_i\neq\mu_j\}$ is given by*

$$U=\left(1+\frac{1}{n-1}V\right)^{-n/2},$$

where $n=n_1+\dots+n_k$. Thus, the ANOVA test is equivalent to the associated likelihood ratio test.

Q 2. *What is the distr. of V?*

We know that for every i

$$Z_i:=\sum_{j=1}^{n_i}(Y_{i,j}-\bar{Y}_i)^2\sim\sigma^2\chi^2(n_i-1).$$

Since $\{Z_i\}$ are indep. across $i = 1, \ldots, k$, we have

$$\sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{i,j} - \bar{Y}_i)^2 \sim \sigma^2 \chi^2(n-k)$$

This takes care of the denominator in the expression for V. Let us analyze the numerator.

Lemma 1. *Under H_0, we have*

$$\sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y})^2 \sim \sigma^2 \chi^2(k-1).$$

Moreover, the above random variable and $\{Z_i\}$ are independent.

Exercise 2. *Prove the first statement of the above lemma for the case where all $\{n_i\}$ are the same.*

Def 1. *Let $X \sim \chi^2(\nu_1)$ and $Z \sim \chi^2(\nu_2)$ be independent. Then,*

$$\frac{X/\nu_1}{Z/\nu_2}$$

has $F(\nu_1, \nu_2)$ distribution.

We conclude that, under H_0,

$$V \sim F(k-1, n-k),$$

and we need to choose $r = F_\alpha$, where the latter is the quantile of level $1-\alpha$ of the $F(k-1, n-k)$ distribution, to obtain an α-level test.

Rem 1. *When $k = 2$ and $n_1 = n_2$, we could also use t-test, applied to the reduced sample $\{Y_{1,j} - Y_{2,j}\}$, but it would not be as efficient (and the test would be different!): i.e. the power of the test would be lower (we do not make this statement rigorous).*

Ex 1. *Two methods, A and B, can be used to produce a certain type of plastic. A sample of several pieces of plastic produced by each method is taken, and the elasticity of each piece is recorded. The sample are described by: $n_1 = n_2 = 6$,*

$$Y_{1,j} : \ 6.1, 7.1, 7.8, 6.9, 7.6, 8.2,$$
$$Y_{2,j} : \ 9.1, 8.2, 8.6, 6.9, 7.5, 7.9.$$

Q 3. *Use ANOVA, level 0.05, to test the null hyp. that the true means are the same. Find the p-value.*

$$\bar{y}_1 \approx 7.283, \quad \bar{y}_2 \approx 8.03, \quad \bar{y} \approx 7.66,$$
$$v \approx \frac{6(7.283 - 7.66)^2 + 6(8.03 - 7.66)^2}{\frac{1}{12-2} \sum_{i=1}^{2} \sum_{j=1}^{6} (y_{ij} - \bar{y}_i)^2} \approx \frac{1.69}{0.59} \approx 2.88,$$
$$r = F_{0.05} \approx 4.96$$

Thus, we accept H_0.

The p-value is

$$1 - F_V(2.88) \approx 0.12,$$

where F_V is the cdf of V under H_0, which, in this case, coincides with the $F(1, 10)$ distribution.

Ex 2. *4 groups of students were subjected to different teaching techniques and then tested. The test scores are given by*

$$y_{1,j} : 65, 87, 73, 79, 81, 69, \quad n_1 = 6, \quad \bar{y}_1 = 75.67,$$
$$y_{2,j} : 75, 69, 83, 81, 72, 79, 90, , \quad n_2 = 7, \quad \bar{y}_2 = 78.43,$$
$$y_{3,j} : 59, 78, 67, 62, 83, 76, , \quad n_3 = 6, \quad \bar{y}_3 = 70.83,$$
$$y_{4j} : 94, 89, 80, 88, , \quad n_4 = 4, \quad \bar{y}_4 = 87.75$$

Q 4. *Construct the ANOVA test for the null hyp. that all teaching methods result in the same expected performance. Produce the p-value.*

$$\bar{y} = (6 \cdot 75.67 + 7 \cdot 78.43 + 6 \cdot 70.83 + 4 \cdot 87.75)/(6 + 7 + 6 + 4),$$
$$v = \frac{\frac{1}{4-1}\sum_{i=1}^{4} n_i(\bar{y}_i - \bar{y})^2}{\frac{1}{23-4}\sum_{i=1}^{4}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2} \approx \frac{237.5}{63} \approx 3.77.$$

The p-value is $1 - F_V(3.77) \approx 0.028$. Thus, we reject H_0 for any level above 2.8%.

2 Analysis of categorical data

Data resulting from an experiment whose set of possible outcomes is relatively small is called categorical. In other words, we consider a sample $\{Y_i\}$ of i.i.d. r.v.'s with Multinomial distribution, denoted $Y_i \sim Mult(p_1, \ldots, p_k)$: Y_i takes values in $\{1, \ldots, k\}$, with $\mathbb{P}(Y_i = j) = p_j$.

Statistical analysis of such data is more robust w.r.t. modeling assumptions, because there are no alternative choices of the distribution family.

There are plenty of practically relevant examples in which such data arises: voting, A-B testing, consumer preference over competing products, etc.

Q 5. *Given a set of candidate probabilities $(\bar{p}_1, \ldots, \bar{p}_k)$, how to test $H_0 = \{p_j = \bar{p}_j,\ j = 1, \ldots, k\}$ vs $H_a = \{p_j \neq \bar{p}_j$, for at least one $j = 1, \ldots, k\}$?*

For $j = 1, \ldots, k$, we denote by b_j the number of times that the value 'j' appears in the Multinomial sample $\{Y_i\}_{i=1}^n$.

Let us construct the likelihood ratio test.

$$\max_{p_j = \bar{p}_j,\, j=1,\ldots,k} L(y_1, \ldots, y_n; p_1, \ldots, p_k) = \bar{p}_1^{b_1} \cdots \bar{p}_k^{b_k},$$

$$\max_{p_1+\cdots+p_{k-1}\leq 1,\, p_j \geq 0} L(y_1, \ldots, y_n; p_1, \ldots, p_{k-1}, 1-p_1-\cdots-p_{k-1}) = \max_{p_1+\cdots+p_{k-1}\leq 1,\, p_j\geq 0} p_1^{b_1} \cdots p_{k-1}^{b_{k-1}}(1-p_1-\cdots-p_{k-1})^{b_k}$$

Taking log of the above and differentiating w.r.t. each p_j, we obtain

$$\frac{b_j}{p_j} - \frac{b_k}{1 - p_1 - \cdots - p_{k-1}} = 0, \quad j = 1, \ldots, k-1,$$

which means that

$$p_j = b_j \frac{1 - p_1 - \cdots - p_{k-1}}{b_k}$$

Summing up the above

$$p_1 + \cdots + p_{k-1} = \frac{n - b_k}{b_k}(1 - (p_1 + \cdots + p_{k-1})),$$

$$p_1 + \cdots + p_{k-1} = \frac{n - b_k}{n},$$

$$p_j = b_j/n,$$

$$\max_{p_1+\cdots+p_{k-1}\leq 1,\, p_j\geq 0} L(y_1, \ldots, y_n; p_1, \ldots, p_{k-1}, 1 - p_1 - \cdots - p_{k-1}) = (b_1/n)^{b_1} \cdots (b_k/n)^{b_k}.$$

Thus, the test statistic of the likelihood ratio test is

$$U = \frac{\bar{p}_1^{b_1} \cdots \bar{p}_k^{b_k}}{(b_1/n)^{b_1} \cdots (b_k/n)^{b_k}},$$

and

$$RR = [0, r].$$

The problem is that we don't know the distribution of U under H_0. Nevertheless, by Wilk's theorem, $-2\log U \to \chi^2(k-1)$, as $n \to \infty$. In principle, we now have a good (likelihood ratio) α-level test. However, there exists a simplification of the above test statistic U that is more intuitive and has similar asymptotic properties, and hence is more popular. It is derived below.

Considering $n \to \infty$, under H_0,

$$-2\log U := -2\sum_{j=1}^{k} b_j \log(\bar{p}_j n/b_j)) = -2\sum_{j=1}^{k} b_j \log(1 + (\bar{p}_j n - b_j)/b_j))$$

$$\approx -2\sum_{j=1}^{k}(\bar{p}_j n - b_j) + \sum_{j=1}^{k}\frac{(\bar{p}_j n - b_j)^2}{b_j} = \sum_{j=1}^{k}\frac{(\bar{p}_j n - b_j)^2}{b_j} \approx \sum_{j=1}^{k}\frac{(\bar{p}_j n - b_j)^2}{n\bar{p}_j}.$$

It turns out that the approximations used above are precise enough to preserve this conclusion.

Thm 1. *(Pearson) As $n \to \infty$, under H_0,*

$$\sum_{j=1}^{k}\frac{(n\bar{p}_j - b_j)^2}{n\bar{p}_j} \to \chi^2(k-1).$$

The above result allows us to choose a simplified test statistic

$$V = \sum_{j=1}^{k}\frac{(n\bar{p}_j - b_j)^2}{n\bar{p}_j},$$

and the associated rejection region

$$RR = [r, \infty),$$

with $r = \chi^2_\alpha$, and the latter is the quantile of level $1-\alpha$ of the $\chi^2(k-1)$ distribution. The resulting test has asymptotic confidence level α.

Rem 2. *The above test statistic is only asymptotically equal to the likelihood ratio test (hence it is expected to be efficient, asymptotically), but it has a natural meaning itself, hence, it is more popular than the actual likelihood ratio test statistic.*

Ex 3. *Each rat in a group of* 90 *is given a choice of one of three doors:* $n_1 = 23$ *choose 1st door,* $n_2 = 36$ *choose 2nd door,* $n_3 = 31$ *choose 3rd door.*

Q 6. *Test the hyp.* $H_0 = \{p_1 = p_2 = p_3 = 1/3\}$ *vs.* $H_a = \{(p_i, p_2, p_3) \neq (1/3, 1/3, 1/3)\}$*, at level* 0.05*. Find the p-value.*

We recall that $V \to \chi^2(2)$*, as* $n \to \infty$*, and compute:*

$$v = \sum_{i=1}^{3} \frac{(90 \cdot \frac{1}{3} - n_i)^2}{90 \cdot \frac{1}{3}} \approx 2.87,$$

$$\chi^2_{0.05} \approx 5.991.$$

Thus, we accept H_0*.*

The p-value is $1 - F_V(2.87) \approx 0.238$ *(the table at the end of the textbook does not give enough precision to compute this number, but there is a Python function for that).*