Mathematical Statistics

Anna Janicka

Lecture XIII, 31.05.2021

Anova – cont.

Non-Parametric Tests

Plan for Today

1. Analysis of variance tests (ANOVA)
2. Goodness-of-fit tests
   - Kolmogorov test
   - Kolmogorov-Smirnov (two samples)
   - Kolmogorov-Lilliefors
   - chi-square goodness-of-fit
3. Tests of independence
   - chi-square test

ANOVA assumptions – reminder

Assume we have *k* samples:

$$X_{1,1}, X_{1,2}, \ldots, X_{1,n_1},$$
$$X_{2,1}, X_{2,2}, \ldots, X_{2,n_2},$$
$$\ldots$$
$$X_{k,1}, X_{k,2}, \ldots, X_{k,n_k}$$, and

- all $X_{i,j}$ are independent (i=1,...,k, j=1,.., n_i)
- $X_{i,j} \sim N(m_i, \sigma^2)$
- we do not know m_1, m_2, ..., m_k, nor σ^2

let $n=n_1+n_2+\ldots+n_k$

Test of the Analysis of Variance (ANOVA) for significance level α – reminder

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

H_1: $\neg H_0$ (i.e. not all μ_i are equal)

A LR test; we get a test statistic:

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

with critical region

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

$$\bar{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} X_{i,j}\,, \bar{X} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} X_{i,j} = \frac{1}{n} \sum_{i=1}^{k} n_i \bar{X}_i$$

for k=2 the ANOVA is equivalent to the two-sample t-test.

ANOVA – interpretation

we have

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j}-\bar{X})^2=\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j}-\bar{X}_i)^2$$

Sum of Squares (SS) — Sum of Squares Between (SSB) — Sum of Squares Within (SSW)

$$\frac{1}{k-1}\sum_{i=1}^{k}n_i(\bar{X}_i-\bar{X})^2$$ – between group variance estimator

$$\frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j}-\bar{X}_i)^2$$ – within group variance estimator

ANOVA test – table

source of variability	sum of squares	degrees of freedom	value of the test statistic F
between groups	SSB	k-1	–
within groups	SSW	n-k	–
total	SS	n-1	*F*

ANOVA test – example

Yearly chocolate consumption in three cities: *A*, *B*, *C* based on random samples of $n_A = 8$, $n_B = 10$, $n_C = 9$ consumers. Does consumption depend on the city?

	A	B	C
sample mean	11	10	7
sample variance	3.5	2.8	3

$\alpha=0.01$

$$\bar{X} = \frac{1}{27}(11 \cdot 8 + 10 \cdot 10 + 7 \cdot 9) = 9.3$$

$$SSB = (11 - 9.3)^2 \cdot 8 + (10 - 9.3)^2 \cdot 10 + (7 - 9.3)^2 \cdot 9 = 75.63$$

$$SSW = 3.5 \cdot 7 + 2.8 \cdot 9 + 3 \cdot 8 = 73.7$$

$$F = \frac{75.63/2}{73.7/24} \approx 12.31 \quad \text{and} \quad F_{0.99}(2,24) \approx 5.61$$

$\rightarrow$ reject H_0 (equality of means), consumption depends on city

ANOVA test – table – example

source of variability	sum of squares	degrees of freedom	value of the test statistic F
between groups	75.63	2	–
within groups	73.7	24	–
total	149.33	26	12.31

Non-parametric tests

- we check whether a random variable fits a given distribution (goodness-of-fit tests).
- we check whether random variables have the same distribution
- we check whether variables/characteristics are independent (test of independence)

Kolmogorov goodness-of-fit test

Model: X_1, X_2, ..., X_n are an IID sample from distribution with CDF F.

H_0: $F = F_0$ (F_0 specified)

H_1: $\neg H_0$ (i.e. the CDF is different)

If F_0 is continuous, we use the statistic

$$D_n = \sup_{t \in R} |F_n(t) - F_0(t)| = \max\{ D_n^+, D_n^- \}$$

where

$$D_n^+ = \max_{i=1,\ldots,n} \left| \frac{i}{n} - F_0(x_{i:n}) \right|, \; D_n^- = \max_{i=1,\ldots,n} \left| F_0(x_{i:n}) - \frac{i-1}{n} \right|$$

and $F_n(t)$ – n-th empirical CDF

Kolmogorov goodness-of-fit test – cont.

The test: we reject H_0 when:

$$D_n > c(\alpha, n)$$

for a critical value $c(\alpha, n)$.

Theorem. If H_0 is true, the distribution of D_n does not depend on F_0.

Problem: This distribution needs tables, for each different *n.*

Theorem. In the limit

$$P(\sqrt{n}D_n \leq d) \xrightarrow[n\to\infty]{} K(d) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2d^2}$$

the approximation may be used for $n \geq 100$

Kolmogorov goodness-of-fit test – cont. (2)

Tables of the asymptotic distribution $K(d)$

$1-\alpha$	0.8	0.9	0.95	0.99
quantile of $K(d)$	1.07	1.22	1.36	1.63
$c(n, \alpha)$ for $n \geq 100$	$1.07/\sqrt{n}$	$1.22/\sqrt{n}$	$1.36/\sqrt{n}$	$1.63/\sqrt{n}$

Kolmogorov goodness-of-fit test – example

Does the sample

0.4085	0.5267	0.3751	0.8329	0.0846
0.8306	0.6264	0.3086	0.3662	0.7952

come from a uniform distribution U(0,1)?

Source: W. Niemiro

Kolmogorov goodness-of-fit test – example cont.

$X_{i:10}$	(i-1)/10	i/10	i/10 - F($X_{i:10}$)	F($X_{i:10}$)-(i-1)/10
0.0846	0	0.1	0.0154	0.0846
0.3086	0.1	0.2	-0.1086	**0.2086**
0.3662	0.2	0.3	-0.0662	0.1662
0.3751	0.3	0.4	0.0249	0.0751
0.4085	0.4	0.5	0.0915	0.0085
0.5267	0.5	0.6	0.0733	0.0267
0.6264	0.6	0.7	0.0736	0.0264
0.7952	0.7	0.8	0.0048	0.0952
0.8306	0.8	0.9	0.0694	0.0306
0.8329	0.9	1	**0.1671**	-0.0671

$D_n = 0.2086$ $\qquad c(10; 0.9) = 0.369$

→ no grounds to reject the null hypothesis that the distribution is uniform

Kolmogorov-Smirnov test of equality of distributions

Model: $X_1, X_2, \ldots, X_n$ are an IID sample from a distribution with CDF F, $Y_1, Y_2, \ldots, Y_m$ are an IID sample from a distribution with CDF G.

H_0: $F = G$

H_1: $\neg H_0$ (i.e. the CDF functions/distributions differ)

If F (and G) is continuous, we test with

$$D_{n,m} = \sup_{t \in R} |F_n(t) - G_m(t)|$$

where $F_n(t)$ – n-th empirical CDF for the first sample, and $G_m(t)$ – m-th empirical CDF for the second sample

Kolmogorov-Smirnov test of equality of distributions – cont.

The test: we reject H_0 if:

$$D_{n,m} > c(\alpha, n, m)$$

for a critical value $c(\alpha, n, m)$.

Theorem. If H_0 is true, the distribution of $D_{n,m}$ does not depend on F (or G).

Theorem. In the limit

$$P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \le d\right) \xrightarrow[n\to\infty, m\to\infty]{} K(d) = \sum_{k=-\infty}^{+\infty} (-1)^k e^{-2k^2 d^2}$$

the approximation is OK for $n,m \ge 100$

Kolmogorov-Lilliefors goodness-of-fit test

Model: $X_1, X_2, \ldots, X_n$ are an IID sample from a distribution with CDF F.

H_0: F is a CDF of a normal distribution

(**with unknown parameters**)

H_1: $\neg H_0$ (i.e. the distribution is not normal)

We test with

$$D_n = \max\{D_n^+, D_n^-\}$$

where

$$D_n^+ = \max_{i=1,\ldots,n}\left|\frac{i}{n} - z_i\right|,\ D_n^- = \max_{i=1,\ldots,n}\left|z_i - \frac{i-1}{n}\right|$$

and

$$z_i = \Phi\left(\frac{X_{i:n} - \bar{X}}{S}\right)$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Kolmogorov-Lilliefors goodness-of-fit test – cont.

The test: we reject H_0 if:

$$D_n > D_n(\alpha)$$

for a critical value $D_n(\alpha)$.

Theorem. If H_0 is true, the distribution of D_n does not depend on the parameters of the normal distribution.

Problem: we need tables and do not know the analytical form of this distribution...

Used for small samples (n ≤30), when it performs better than the chi-square test

Kolmogorov-Lilliefors goodness-of-fit test – critical values

Sample Size N	Level of Significance for $D = \text{Max}\left\|F^*(X) - S_N(X)\right\|$				
	.20	.15	.10	.05	.01
4	.300	.319	.352	.381	.417
5	.285	.299	.315	.337	.405
6	.265	.277	.294	.319	.364
7	.247	.258	.276	.300	.348
8	.233	.244	.261	.285	.331
9	.223	.233	.249	.271	.311
10	.215	.224	.239	.258	.294
11	.206	.217	.230	.249	.284
12	.199	.212	.223	.242	.275
13	.190	.202	.214	.234	.268
14	.183	.194	.207	.227	.261
15	.177	.187	.201	.220	.257
16	.173	.182	.195	.213	.250
17	.169	.177	.189	.206	.245
18	.166	.173	.184	.200	.239
19	.163	.169	.179	.195	.235
20	.160	.166	.174	.190	.231
25	.149	.153	.165	.180	.203
30	.131	.136	.144	.161	.187
Over 30	$\frac{.736}{\sqrt{N}}$	$\frac{.768}{\sqrt{N}}$	$\frac{.805}{\sqrt{N}}$	$\frac{.886}{\sqrt{N}}$	$\frac{1.031}{\sqrt{N}}$

Source: H. Lilliefors

Chi-square goodness-of-fit test

Model: $X_1, X_2, \ldots, X_n$ are an IID sample from a discrete distribution with k values (1, ..., k).

H_0: the distribution probabilities are equal to

i	1	2	3	...	k
$P(X=i)$	p_1	p_2	p_3	...	p_k

H_1: $\neg H_0$ (i.e. the distribution is different)

value labels

If the results of the experiment are

i	1	2	3	...	k
N_i	N_1	N_2	N_3	...	N_k

where N_i denotes the number of outcomes equal to i:

$$N_i = \sum_{j=1}^{n} 1_{X_j = i}$$

Chi-square goodness-of-fit test – cont.

General form of the test:

$$\chi^2 = \sum \frac{(\text{observed value} - \text{expected value})^2}{\text{expected value}}$$

here:

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

Theorem. If H_0 is true, the distribution of the χ^2 statistic converges to a chi-square distr with k-1 degrees of freedom $\chi^2(k-1)$ for $n \to \infty$

Procedure: we reject H_0 if $\chi^2 > c$, where $c = \chi^2_{1-\alpha}(k-1)$ is a quantile of rank 1-α from a chi-square distr with k-1 degrees of freedom

Chi-square goodness-of-fit test – example

Is a die symmetric? For a significance level α=0.05

n=150 tosses. Results:

i	1	2	3	4	5	6
N_i	15	27	36	17	26	29

H_0: $(N_1, N_2, N_3, N_4, N_5, N_6)$

$\sim$Mult(150, 1/6, 1/6, 1/6, 1/6, 1/6, 1/6)

H_1: $\neg H_0$

$$\chi^2 = \frac{(15-25)^2}{25} + \frac{(27-25)^2}{25} + \frac{(36-25)^2}{25} + \frac{(17-25)^2}{25} + \frac{(26-25)^2}{25} + \frac{(29-25)^2}{25} = 12.24$$

$\chi^2_{1-0.05}(5) \approx 11.7$ $\rightarrow$ **we reject H_0.**

Warsaw University
Faculty of Economic Sciences

Source: W. Niemiro

Chi-square goodness-of-fit test – distribution with an unknown parameter.

Model: $X_1, X_2, \ldots, X_n$ are an IID sample from a discrete distribution with k values (1, ..., k).

H_0: distribution probabilities are equal to

i	1	2	3	...	k
$P(X=i)$	$p_1(\theta)$	$p_2(\theta)$	$p_3(\theta)$	...	$p_k(\theta)$

where θ is an unknown parameter of dimension d.

H_1: $\neg H_0$ (i.e. the distribution is different)

Chi-square goodness-of-fit test – distribution with an unknown parameter, cont.

Test statistics are constructed like in the previous case, with the expected values calculated using ML estimators of the parameter θ. Only the number of degrees of freedom changes:

Theorem. If H_0 is true, the distribution of the χ^2 statistic converges to a chi-square distribution with k-d-1 degrees of freedom $\chi^2(k-d-1)$ for $n \to \infty$

Chi-square goodness-of-fit test – version for continuous distributions

Kolmogorov tests are better, but the chi-square test may also be used

Model: $X_1, X_2, \ldots, X_n$ are an IID sample from a continuous distribution.

H_0: The distribution is given by F

H_1: $\neg H_0$ (i.e. the distribution is different)

It suffices to divide the range of values of the random variable into classes and count the observations. The expected values are known (result from F).Then: the chi-square test.

Chi-square goodness-of-fit test – practical notes

- The test should be used for large samples only.
- The expected counts can't be too small (<5). If they are smaller, observations should be grouped.
- The classes in the „continuous" version may be chosen arbitrarily, but it is best if the theoretical probabilities are balanced.

Chi-square test of independence

Model: (X_1, Y_1), ..., (X_n, Y_n) are an IID sample from a two-dimensional distribution with $r \cdot s$ values (denoted by the set $\{1, ..., r\} \times \{1, ..., s\}$).

Let the theoretical distribution be

$$p_{ij} = P(X = i, Y = j) \quad i = 1, \ldots, r \;\; j = 1, \ldots, s$$

Denote $$p_{i\bullet} = \sum_{j=1}^{s} p_{ij}, \qquad p_{\bullet j} = \sum_{i=1}^{r} p_{ij}$$

We want to verify independence of X and Y:

H_0: $p_{ij} = p_{i\bullet} \cdot p_{\bullet j} \;\; i = 1, \ldots, s, \qquad j = 1, \ldots, r$

H_1: $\neg H_0$

Chi-squared test of independence – cont.

The empirical distribution may be summarized by a table (so-called contingency table, or crosstab)

$i \setminus j$	1	2	...	s	$N_{i\bullet}$
1	N_{11}	N_{12}		N_{1s}	$N_{1\bullet}$
2	N_{21}	N_{22}		N_{2s}	$N_{2\bullet}$
...					
r	N_{r1}	N_{r2}		N_{rs}	$N_{r\bullet}$
$N_{\bullet j}$	$N_{\bullet 1}$	$N_{\bullet 2}$		$N_{\bullet s}$	n

Chi-squared test of independence – cont. (2)

This is a special case of a goodness-of-fit test with (r-1) + (s-1) parameters to be estimated:

The test statistic:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(N_{ij} - N_{i\bullet} N_{\bullet j}/n)^2}{N_{i\bullet} N_{\bullet j}/n}$$

has a chi-square distribution with (r-1)(s-1) degrees of freedom (if H_0 is true)

Chi-squared test of independence – example

We verify independence of political and musical preferences, at signif. level α =0.05

	Support X	Do not support X	Total
Listen to jazz	25	10	35
Listen to rock	20	20	40
Listen to hip-hop	15	10	25
Total	60	40	100

$$\chi^2 = \frac{(25-60*35/100)^2}{60*35/100} + \frac{(20-60*40/100)^2}{60*40/100} + \frac{(15-60*25/100)^2}{60*25/100} + \frac{(10-40*35/100)^2}{40*35/100} + \frac{(20-40*40/100)^2}{40*40/100} + \frac{(10-40*25/100)^2}{40*25/100}$$

$$\approx 3.57$$

$$\chi^2_{1-0.05}((2-1)(3-1)) = \chi^2_{0.95}(2) \approx 5.99$$

→ **no grounds to reject *H*$_0$.**

Source: W. Niemiro