Mathematical Statistics

Anna Janicka

Lecture XIII, 26.05.2022

ANOVA

NON-PARAMETRIC TESTS

Plan for Today

1. Comparing two populations – cont.
2. Analysis of variance tests (ANOVA)
3. Goodness-of-fit tests
   - Kolmogorov test
   - chi-square goodness-of-fit

Model I: comparison of means, variance known, significance level α – *reminder*

$X_1, X_2, \ldots, X_{nX}$ are an IID sample from distr $N(\mu_X, \sigma_X^2)$,
$Y_1, Y_2, \ldots, Y_{nY}$ are an IID sample from distr $N(\mu_Y, \sigma_Y^2)$,
σ_X^2, σ_Y^2 are **known**, samples are independent

H_0: $\mu_X = \mu_Y$

Test statistic:

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} \sim N(0,1)$$

assuming H_0 is true

H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X > \mu_Y$

critical region $C^* = \{x : U(x) > u_{1-\alpha}\}$

H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X \neq \mu_Y$

critical region $C^* = \{x : |U(x)| > u_{1-\alpha/2}\}$

Model II: variance unknown but assumed equal, significance level α – *reminder*

$X_1, X_2, \dots, X_{nX}$ are an IID sample from distr $N(\mu_X, \sigma^2)$,
$Y_1, Y_2, \dots, Y_{nY}$ are an IID sample from distr $N(\mu_Y, \sigma^2)$
with σ^2 **unknown**, samples are independent

H_0: $\mu_X = \mu_Y$ Test statistic: $T = \dfrac{\bar{X} - \bar{Y}}{S_* \sqrt{\dfrac{1}{n_X} + \dfrac{1}{n_Y}}} \sim t\,(n_X + n_Y - 2)$

Assuming H_0 is true

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

H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X > \mu_Y$

critical region $C^* = \{x : T\,(x) > t_{1-\alpha}(n_x + n_y - 2)\}$

H_0: $\mu_X = \mu_Y$ against H_1: $\mu_X \neq \mu_Y$

critical region $C^* = \{x : |\,T(x)| > t_{1-\alpha/2}(n_x + n_y - 2)\}$

Warsaw University
Faculty of Economic Sciences

$$S_X^2 = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \bar{X})^2, S_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2$$

Model II: comparison of variances, significance level α

$X_1, X_2, \ldots, X_{nX}$ are an IID sample from distr $N(\mu_X, \sigma_X^2)$,
$Y_1, Y_2, \ldots, Y_{nY}$ are an IID sample from distr $N(\mu_Y, \sigma_Y^2)$,
σ_X^2, σ_Y^2 are **unknown**, samples are independent

H_0: $\sigma_X = \sigma_Y$

Test statistic:

$$F = \frac{S_X^2}{S_Y^2} \sim F(n_X - 1, n_Y - 1)$$

assuming H_0 is true

H_0: $\sigma_X = \sigma_Y$ against H_1: $\sigma_X > \sigma_Y$

critical region $C^* = \{x : F(x) > F_{1-\alpha}(n_X - 1, n_Y - 1)\}$

H_0: $\sigma_X = \sigma_Y$ against H_1: $\sigma_X \neq \sigma_Y$

critical region $C^* = \{x : F(x) < F_{\alpha/2}(n_X - 1, n_Y - 1) \vee F(x) > F_{1-\alpha/2}(n_X - 1, n_Y - 1)\}$

$$S_X^2 = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \bar{X})^2, S_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2$$

Model III: comparison of means for large samples, significance level α

$X_1, X_2, \ldots, X_{nX}$ are an IID sample from distr. with mean μ_X, $Y_1, Y_2, \ldots, Y_{nY}$ are an IID sample from distr. with mean μ_Y, both distr. have unknown variances, samples are independent, n_X, n_Y – large.

H_0: $\mu_x = \mu_Y$ Test statistic:

$$U = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim N(0,1)$$

assuming H_0 is true, for large samples **approximately**

H_0: $\mu_x = \mu_Y$ against H_1: $\mu_x > \mu_Y$

critical region

$$C^* = \{x : U(x) > u_{1-\alpha}\}$$

H_0: $\mu_x = \mu_Y$ against H_1: $\mu_x \neq \mu_Y$

critical region

$$C^* = \{x : |U(x)| > u_{1-\alpha/2}\}$$

$$S_X^2 = \frac{1}{n_X - 1}\sum_{i=1}^{n_X}(X_i - \bar{X})^2, S_Y^2 = \frac{1}{n_Y - 1}\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2$$

Model III – example (equality of means?)

167 students take part in a probability calculus exam. Is attending lectures profitable? ($\alpha = 0.05$)

Among those, who participated 3 times (93 students):

mean = 3, variance = 0.70;

Among those, who participated less than 3 times (74 students): mean = 2.72, variance = 0.69.

Value of the test statistic

$$U = \frac{3 - 2.72}{\sqrt{{0.70}/{93} + {0.69}/{74}}} \approx 2.13$$

Model IV: comparison of fractions for large samples, significance level α

Two IID samples from two-point distributions. X – number of successes in n_X trials with prob of success p_X, Y – number of successes in n_Y trials with prob of success p_Y. p_X and p_Y unknown, n_X and n_Y large.

H_0: $p_X = p_Y$

Test statistic:

$$U^* = \frac{\frac{X}{n_X} - \frac{Y}{n_Y}}{\sqrt{p_*(1-p_*)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim N\,(0,1)$$

assuming H_0 is true, for large samples **approximately**

where $p^* = \frac{X+Y}{n_x + n_y}$

H_0: $p_X = p_Y$ against H_1: $p_X > p_Y$

critical region

$$C^* = \{x : U^*(x) > u_{1-\alpha}\}$$

H_0: $p_X = p_Y$ against H_1: $p_X \neq p_Y$

critical region

$$C^* = \{x : |U^*(x)| > u_{1-\alpha/2}\}$$

Model IV – example (equality of probabilities?)

167 students take part in a probability calculus exam. Is attending lectures profitable? ($\alpha = 0.05$)

Among those, who participated 3 times (93 students):

64 passed (68.8%);

Among those, who participated less than 3 times (74 students): 36 passed (48.6%).

Value of the test statistic

$$U = \frac{0.688 - 0.486}{\sqrt{{}^{100}/_{167} \cdot {}^{67}/_{167} \cdot \left({}^{1}/_{93} + {}^{1}/_{74}\right)}} \approx 2{,}55$$

Tests for more than two populations

A naive approach:

pairwise tests for all pairs

But:

in this case, the type I error is higher than the significance level assumed for each simple test...

More populations

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)
- $X_{i,j} \sim N(m_i, \boldsymbol{\sigma}^{\mathbf{2}})$
- 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

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:

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

assuming H_0 is true

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

$$\underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j}-\bar{X})^2}_{\text{Sum of Squares (SS)}} = \underbrace{\sum_{i=1}^{k} n_i(\bar{X}_i-\bar{X})^2}_{\text{Sum of Squares Between (SSB)}} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{i,j}-\bar{X}_i)^2}_{\text{Sum of Squares Within (SSW)}}$$

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

α=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: $X_1, X_2, \ldots, X_n$ 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

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

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

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

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

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 $(1, \ldots, k)$.

H_0: the distribution probabilities are equal to

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

value labels

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

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.**

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 $(1, \ldots, k)$.

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 $\chi^2(k\text{-}d\text{-}1)$ 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.