Mathematical Statistics

Anna Janicka

Lecture XII, 19.05.2022

HYPOTHESIS TESTING IV:

PARAMETRIC TESTS: COMPARING TWO OR MORE POPULATIONS

Plan for today

- Parametric LR tests for one population cont.
- 2. Asymptotic properties of the LR test
- 3. Parametric LR tests for two populations
- 4. Comparing more than two populations
 - ANOVA

Notation

*x*_{something} **always** means a quantile of rank something

Model III: comparing the mean

Asymptotic model: $X_1, X_2, ..., X_n$ are an IID sample from a distribution with mean μ and variance (unknown), *n* – large.

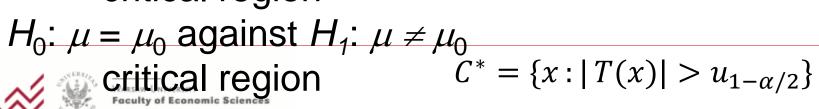
$$H_0$$
: $\mu = \mu_0$
Test statistic:
$$T = \frac{\bar{X} - \mu_0}{S} \sqrt{n}$$

has, for large *n*, an approximate distribution N(0,1)

$$H_0$$
: $\mu = \mu_0$ against H_1 : $\mu > \mu_0$ critical region $C^* = \{x : T(x) > u_{1-\alpha}\}$

H₀:
$$\mu = \mu_0$$
 against H₁: $\mu < \mu_0$
critical region
$$C^* = \{x : T(x) < u_\alpha = -u_{1-\alpha}\}$$

$$H_0$$
: $\mu = \mu_0$ against H_1 : $\mu \neq \mu_0$



Model IV: comparing the fraction

Asymptotic model: X_1 , X_2 , ..., X_n are an IID sample from a two-point distribution, n – large.

$$P_p(X = 1) = p = 1 - P_p(X = 0)$$

H₀:
$$p = p_0$$

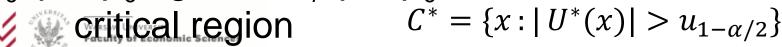
Test statistic: $U^* = \frac{\bar{X} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n} = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)}} \sqrt{n}$

has an approximate distribution N(0,1) for large n

$$H_0$$
: $p = p_0$ against H_1 : $p > p_0$ critical region $C^* = \{x : U^*(x) > u_{1-\alpha}\}$

$$H_0$$
: $p = p_0$ against H_1 : $p < p_0$ critical region $C^* = \{x : U^*(x) < u_\alpha = -u_{1-\alpha}\}$

$$H_0$$
: $p = p_0$ against H_1 : $p \neq p_0$



Model IV: example

We toss a coin 400 times. We get 180 heads. Is the coin symmetric?

$$H_0$$
: $p = \frac{1}{2}$ $U^* = \frac{(180/400 - 1/2)}{\sqrt{1/2(1 - 1/2)}} \sqrt{400} = -2$

for α = 0.05 and H_1 : $p \neq \frac{1}{2}$ we have $u_{0.975}$ = 1.96 \rightarrow we reject H_0 for α = 0.05 and H_1 : $p < \frac{1}{2}$ we have $u_{0.05}$ = $-u_{0.95}$ = -1.64 \rightarrow we reject H_0

for $\alpha = 0.01$ and H_1 : $p \neq \frac{1}{2}$ we have $u_{0.995} = 2.58$ \rightarrow we do not reject H_0

for $\alpha = 0.01$ and H_1 : $p < \frac{1}{2}$ we have $u_{0.01} = -u_{0.99} = -2.33$

 \rightarrow we do not reject H_0

p-value for $H_1: p \neq \frac{1}{2}: 0.044$

p-value for H_1 : $p < \frac{1}{2}$: 0.022

Likelihood ratio test for composite hypotheses – reminder

 $X \sim P_{\theta}$, $\{P_{\theta} : \theta \in \Theta\}$ – family of distributions We are testing H_0 : $\theta \in \Theta_0$ against H_1 : $\theta \in \Theta_1$ such that $\Theta_0 \cap \Theta_1 = \emptyset$, $\Theta_0 \cup \Theta_1 = \Theta$

Let

 H_0 : $X \sim f_0(\theta_0, \cdot)$ for some $\theta_0 \in \Theta_0$.

 H_1 : $X \sim f_1(\theta_1, \cdot)$ for some $\theta_1 \in \Theta_1$,

where f_0 and f_1 are densities (for $\theta \in \Theta_0$ and $\theta \in \Theta_1$, respectively)



Likelihood ratio test for composite hypotheses – reminder (cont.)

Test statistic:

$$\tilde{\lambda} = \frac{\sup_{\theta \in \Theta} f(\theta, X)}{\sup_{\theta_0 \in \Theta_0} f_0(\theta_0, X)}$$

or
$$\tilde{\lambda} = \frac{f(\hat{\theta}, X)}{f_0(\hat{\theta}_0, X)}$$

where $\hat{\theta}$, $\hat{\theta}_0$ are the ML estimators for the model without restrictions and for the null model.

We reject H_0 if $\tilde{\lambda} > \tilde{c}$ for a constant \tilde{c} .





Asymptotic properties of the LR test

We consider two nested models, we test

 H_0 : $h(\theta) = 0$ against H_1 : $h(\theta) \neq 0$

Under the assumption that

- ☐ h is a nice function
- □ Θ is a d-dimensional set
- $\square \ \Theta_0 = \{\theta : h(\theta) = 0\}$ is a d p dimensional set

Theorem: If H_0 is true, then for $n \rightarrow \infty$ the distribution of the statistic $2 \ln \tilde{\lambda}$ converges to a chi-square distribution with p degrees of freedom

Asymptotic properties of the LR test – example

Exponential model: $X_1, X_2, ..., X_n$ are an IID sample from $Exp(\theta)$.

We test H_0 : $\theta = 1$ against H_1 : $\theta \neq 1$

$$MLE(\theta) = \hat{\theta} = 1/\bar{X}$$

$$\tilde{\lambda} = \frac{\prod f_{\widehat{\theta}}(x_i)}{\prod f_1(x_i)} = \frac{\frac{1}{\bar{X}^n} \exp(-\frac{1}{\bar{X}} \Sigma x_i)}{\exp(-\Sigma x_i)} = \frac{1}{\bar{X}^n} \exp(n(\bar{X} - 1))$$

then:

$$\tilde{\lambda} > \tilde{c} \Leftrightarrow 2 \ln \tilde{\lambda} > 2 \ln \tilde{c}$$

from Theorem:
$$2\ln \tilde{\lambda} = 2n((\bar{X} - 1) - \ln \bar{X}) \xrightarrow{D} \chi^2(1)$$

for a sign. level $\alpha = 0.05$ we have $\chi_{0.95}^2(1) \approx 3.84 \approx 2 \ln \tilde{c}$ $\lambda > e^{3.84/2}$ so we reject H_0 in favor of H_1 if



Comparing two or more populations

We want to know if populations studied are "the same" in certain aspects:

- parametric tests: we check the equality of certain distribution parameters
- nonparametric tests: we check whether distributions are the same

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

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

$$H_0: \mu_X = \mu_Y \qquad U = \frac{\bar{X} - \bar{Y}}{\sqrt{\sigma_X^2/n_X + \frac{\sigma_Y^2}{n_Y}}} \sim N \ (0,1)$$
Test statistic:
$$\sqrt{\frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}} \sim N \ (0,1)$$

$$H_0: \mu_X = \mu_Y \text{ against } H_1: \mu_X > \mu_Y \qquad \text{assuming } H_0 \text{ is true}$$

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

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

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



Model I – comparison of means. Example

 X_1 , X_2 , ..., X_{10} are an IID sample from distr N(μ_X ,112), Y_1 , Y_2 , ..., Y_{10} are an IID sample from distr N(μ_Y ,132) Based on the sample:

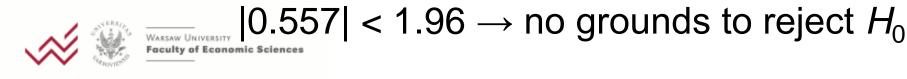
$$\bar{X} = 501, \bar{Y} = 498$$

Are the means equal, at significance level 0.05?

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

$$U = \frac{501 - 498}{\sqrt{\frac{13^2}{10} + \frac{11^2}{10}}} \approx 0.557$$

we have: $u_{0.975} \approx 1.96$.



Model II: comparison of means, variance unknown but assumed equal, significance level α

 $X_1, X_2, ..., X_{nX}$ are an IID sample from distr N(μ_X, σ^2), $Y_1, Y_2, ..., Y_{nY}$ are an IID sample from distr N(μ_Y, σ^2) with σ^2 **unknown**, samples are independent

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

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}} \sqrt{\frac{n_X n_Y}{n_X + n_Y}} (n_X + n_Y - 2) \sim t (n_X + n_Y - 2)$$
Assuming H_0 is

$$H_0$$
: $\mu_X = \mu_Y$ against H_1 : $\mu_X > \mu_Y$ true true 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 $C^* = \{x : |T(x)| > t_{1-\alpha/2}(n_x + n_y - 2)\}$



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

Model II: comparison of means, variance unknown but assumed equal, cont.

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

can be rewritten as

$$T = \frac{\bar{X} - \bar{Y}}{S_* \sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}} \sim t (n_X + n_Y - 2)$$

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 σ^2 based on the two samples jointly

Model II: comparison of variances, significance level α

 $X_1, X_2, ..., X_{nX}$ are an IID sample from distr N(μ_X, σ_X^2), $Y_1, Y_2, ..., Y_{nY}$ are an IID sample from distr N(μ_Y, σ_Y^2), σ_X^2, σ_Y^2 are **unknown**, samples are independent

$$H_0$$
: $\sigma_X = \sigma_Y$

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

Test statistic:

$$H_0$$
: $\sigma_X = \sigma_Y$ against H_1 : $\sigma_X > \sigma_Y$ assuming H_0 is true 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 $C^* = \{x : F(x) < F_{\alpha/2}(n_X - 1, n_Y - 1)$
 $\forall F(x) > F_{1-\alpha/2}(n_X - 1, n_Y - 1)\}$



Model II: comparison of means, variances unknown and no equality assumption

 $X_1, X_2, ..., X_{nX}$ are an IID sample from distr N(μ_X, σ_X^2), $Y_1, Y_2, ..., Y_{nY}$ are an IID sample from distr N(μ_Y, σ_Y^2), σ_X^2, σ_Y^2 are **unknown**, samples independent

$$H_0$$
: $\mu_{\rm X} = \mu_{\rm Y}$

The test statistic would be very simple, but:

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}} \sim ??$$

It isn't possible to design a test statistic such that the distribution does not depend on σ_X^2 and σ_Y^2 (values)...



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

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

 $X_1, X_2, ..., X_{nX}$ are an IID sample from distr. with mean μ_X $Y_1, Y_2, ..., Y_{nY}$ are an IID sample from distr. with mean μ_Y , both distr. have unknown variances, samples are independent, n_{χ} , n_{γ} – 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

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



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

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

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

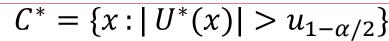
Model IV: comparison of fractions for large samples, significance level α

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:
$$U^* = \frac{\frac{X}{n_X} - \frac{Y}{n_Y}}{\sqrt{p_*(1 - p_*)\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \sim N(0,1)$$
where $p^* = \frac{X + Y}{n_X + n_Y}$ assuming H_0 is true, for large samples approximately

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





Model IV – example (equality of probabilities?)

157 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

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

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:

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



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

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

 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)

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

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

	А	В	С
sample mean	11	10	7
sample variance	3.5	2.8	3

 α =0.01

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

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

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

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

$$\rightarrow \text{reject } H_0 \text{ (equality of means)},$$

$$\text{consumption depends on city}$$

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



