Lecture 5. Statistical tests (the criteria of consent)

Let $\vec{X} = (X_1, \dots, X_n)$ – sample $\in \mathcal{F}$ – distribution family.

The basic hypothesis $H_1 = \{\mathcal{F} = \mathcal{F}_1\}$ with $H_2 = \{\mathcal{F} \neq \mathcal{F}_1\}$ alternative.

Definition 1. (cc) Let Borel function $\rho(\vec{X})$ exists and has properties:

$$\begin{cases} (C1)\text{: if } H_1 - \text{true (if } X_i \in \mathcal{F}_1 \text{ then } \rho(\vec{X}) \Rightarrow G, G - \text{known continuous distribution}) \\ (C2)\text{: if } H_1 - \text{incorrect (wrong), it means: if } X_i \text{ have } \mathcal{F}_2 \neq \mathcal{F}_1, \text{ then} \\ |\rho(\vec{X})| \xrightarrow{P} \infty, \text{ if } n \to \infty, \text{ for } \forall\, \mathcal{F}_2 \neq \mathcal{F}_1 \end{cases} \quad (1)$$

Definition 2. *Criterion of consent*

For random variable η, $\eta \in G$ it is necessary to determine constant C, satisfying the equation $\varepsilon = P(|\eta| \geq C)$ and construct criterion:

$$\delta(\vec{X}) = \begin{cases} H_1, & if \quad |\rho(\vec{X})| < C \\ H_2, & if \quad |\rho(\vec{X})| \geq C \end{cases} \quad (2)$$

Explanation: if H_1 – true, then behavior of ρ is the same as η, ε – small, so it is extremely difficult to get H_2 .

Theorem

The criterion of consent δ (1)-(2) has the asymptotical size ε and the property of consistency

Proof Property (C1) provides criterion size:

$\alpha_1(\delta) = P_{H_1}(|\rho(\vec{X})| \geq C) \to P(|\eta| \geq C) = \varepsilon$.

(C2) – consistency: according to definition $\eta_n \xrightarrow{P} \infty$ means that for any C>0

$P(\eta_n < C) \to 0$ as $n \to \infty$, so for $\forall\, \mathcal{F}_2$ (may be set of distributions) probability of the second kind errors tends to zero: $\alpha_2(\delta, \mathcal{F}_2) = P_{\mathcal{F}_2}(|\rho(\vec{X})| < C) \to 0$, *proven.*

Remark

If instead the weak convergence in (C1) we have $\rho(\vec{X}) \in G$, then criterion (2) has *exact* size ε.

Statistical criterion

I. Kolmogorov criterion (Kc)

$\vec{X} = (X_1...,X_n)$ have distribution function $\mathcal{F}$, $H_1=\{\mathcal{F}=\mathcal{F}_1\}$, $H_2=\{\mathcal{F}\neq\mathcal{F}_1\}$. If $\mathcal{F}_1$ has continuous distribution function F_1, then (Kc) use! Let's check the validity of the criterion, namely properties C1-C2 for given $\rho(\vec{X})$:

Proof

Let consider $\rho(\vec{X}) = \sqrt{n}\sup_y|F_n^*(y) - F_1(y)|$.

C1: If H_1 – true , then X_i for $\forall i$ has distribution $\mathcal{F}_1$. According to Kolmogorov theorem $\rho(\vec{X}) \overset{as.}{\Rightarrow} \eta$, η has distribution with Kolmogorov distribution function K(y).

C2: If H_1 – wrong, X_i for $\forall i$ has distribution $\mathcal{F}_2$. ($\mathcal{F}_2 \neq \mathcal{F}_1$). By Glivenko-Cantelli theorem, the empirical distribution $F_n^*(y) \overset{P}{\to} F_2(y)$ for $\forall y,\ n \to \infty$. If $\mathcal{F}_2 \neq \mathcal{F}_1$ then $\exists y_0$: $|F_2(y_0)-F_1(y_0)|>0$, so $\sup_y |F_n^*(y) - F_1(y)| \geq |F_n(y_0)-F_1(y_0)| \overset{P}{\to} |F_2(y_0) - F_1(y_0)| > 0$ that is

for $\forall\varepsilon$ find C: $\varepsilon = P\{\eta \geq c\}, \eta \in K(y)$,

this allows us to the Kolmogorov criterion:

$$\delta(\vec{X})=\begin{cases} H_1, & \text{if} \quad \rho(\vec{X}) < C \\ H_2, & \text{if} \quad \rho(\vec{X}) \geq C \end{cases}$$

II. Pearson Criterion (PC) *based on grouped data.*

For given sample $\vec{X}=(X_1,\ldots, X_n)$ belongs to distribution $\mathcal{F}$ the simple hypothesis is $H_1=\{\mathcal{F}=\mathcal{F}_1\}$, $H_2=\{\mathcal{F}\neq\mathcal{F}_2\}$.

$A_1, \ldots, A_k$ – grouping intervals *:* $A_i \cap A_j = \varnothing$, $i\neq j$, $\nu_j = \{$ number $X_i \in A_j \} =$

$= \sum_{i=1}^{n} I(X_i \in A_j)$, $p_j=P_{H_1}(X_i \in A_j)>0$ (p_j - theoretical probability of H_1 - true, $X_i \in F_1$),

$$\sum_{j=1}^{k} p_j = 1, \quad \rho(\vec{X}) = \sum_{j=1}^{k} \frac{(\nu_j - np_j)^2}{np_j}.$$

Remark. *On correction PC*

$$\begin{cases} H'_1 = \{\mathcal{F}: P(X_1 \in A_j) = p_j, \forall\, j = 1,..,k\} \\ H'_2 = \{\exists j: p_j^{H_1} \neq p_j^{H_2}, \quad H'_1 - \text{wrong}\}, \end{cases}$$ so need large number of intervals!

Theorem Pearson:

If H_1 or H_2 – true, then for fixed k, and $n \to \infty$

$$\rho(\vec{X}) = \sum_{j=1}^{k} \frac{(v_j - np_j)^2}{np_j} => H_{k-1}$$

Here H_{k-1} is χ^2_{k-1} distribution, so consent criterion is: it is necessary to determine constant C: $\varepsilon = P(\eta \geq C), \eta \in H_{k-1}$ and

$$\delta(\vec{X}) = \begin{cases} H'_1, & if\ \rho(\vec{X}) < C \\ H'_2, & if\ \rho(\vec{X}) \geq C \end{cases}$$

III Criterion χ^2 *(parametric case)*

Let $\vec{X} = \{X_1, X_2, .., X_n\} \in$ uknown distribution

$$H_1 = \{\mathcal{F} \in \{\mathcal{F}_\theta, \boldsymbol{\theta} \in\ \Theta \subseteq R^d\}\}$$

$\boldsymbol{\theta}$ – unknown vector of parameters, d – size of $\boldsymbol{\theta}$ dimension.

There are $A_1, .., A_k$ – intervals, $k > d + 1$, v_j – number of $X_i \in A_j$

$p_j = P_{H_1}(X_1 \in A_j) = P_j(\boldsymbol{\theta})$ and $\rho(\vec{X}, \boldsymbol{\theta}) = \sum_{j=1}^{k} \frac{(v_j - np_j(\boldsymbol{\theta}))^2}{np_j(\boldsymbol{\theta})}$ and $\boldsymbol{\theta}^*$

$\left(\boldsymbol{\theta}^* = \boldsymbol{\theta}^*(\vec{X}):\ \rho(\vec{X}, \boldsymbol{\theta}^*)\right)$ reaches its minimum value

Theorem Fisher.

If hypothesis H_1, d – dimension of vector $\boldsymbol{\theta}$, then for fixed k and $n \to \infty$

$\rho(\vec{X}, \boldsymbol{\theta}^*) = \sum_{j=1}^{k} \frac{\left(v_j - np_j(\boldsymbol{\theta}^*)\right)^2}{np_j(\boldsymbol{\theta}^*)} \Rightarrow H_{k-1-d}$, where H_{k-1-d} family of χ^2_{k-1-d} distribution, note that $\boldsymbol{X^2_{k-1-d}}$ – ***parametric*** **criterion**:

$\eta \in H_{k-1-d}$, for given ε it is necessary to determine constant C: $\varepsilon = P(\eta \geq C)$ and

$$\delta(\vec{X}) = \begin{cases} H_1, & \text{if}\ \ \rho(\vec{X}, \boldsymbol{\theta}^*) < C \\ H_2, & \text{if}\ \rho(\vec{X}, \boldsymbol{\theta}^*) \geq C \end{cases}$$

Remark. To find $\boldsymbol{\theta}^*$numerically – is not simple idea, so we can use instead $\boldsymbol{\theta}^*$ likelihood estimation $\widehat{\boldsymbol{\theta}}$.

IV Criterion Fisher $(\boldsymbol{\sigma_1^2 = \sigma_2^2})$

If we know that variances of two samples are the same, for estimation of its average we use Student criterion. *The first step* – to estimate variance.

Let $\vec{X} = \{X_1, X_2, .., X_n\}$ are independent and $X_j \in N(a_1, \boldsymbol{\sigma_1^2})$; $\vec{Y} = \{Y_1, Y_2, .., Y_m\}$, $Y_j \in N(a_2, \boldsymbol{\sigma_2^2})$. Fisher criterion: $H_1 = \{\sigma_1^2 = \sigma_2^2\}$. It is known that $S_0^2(\vec{X})$, $S_0^2(\vec{Y})$ is an unbiased estimate of variances and we introduce a measure

$$\rho(\vec{X}, \vec{Y}) = \frac{S_0^2(\vec{X})}{S_0^2(\vec{Y})},$$

to get criteria statistics – *the second step*.

Theorem

If hypothesis H_1 is true, then $\rho(\vec{X}, \vec{Y})$ has Fisher distribution $F_{n-1,m-1}$ with $n-1$ and $m-1$ degrees of freedom.

Proof

$\chi_{n-1}^2 = \frac{(n-1)S_0^2(\vec{X})}{\sigma_1^2} \in H_{n-1}$ and $\chi_{m-1}^2 = \frac{(m-1)S_0^2(\vec{Y})}{\sigma_2^2} \in H_{m-1}$.

If $\sigma_1 = \sigma_2$, then $\frac{\sigma_1}{\sigma_2} = 1$ and

$$\rho(\vec{X}, \vec{Y}) = \frac{S_0^2(\vec{X})}{\sigma_1^2} \frac{\sigma_2^2}{S_0^2(\vec{Y})} = \frac{\frac{\chi_{n-1}^2}{n-1}}{\frac{\chi_{m-1}^2}{m-1}} \in F_{n-1,m-1}$$

Take quantile $f_{1-\varepsilon} \epsilon F_{n-1,m-1}$ and obtain Fisher criterion $\delta(\vec{x}, \vec{y})$

$$\delta(\vec{X}, \vec{Y}) = \begin{cases} H_1, & if\ \rho(\vec{X}, \vec{Y}) < f_{1-\varepsilon} \\ H_2, & if\ \rho(\vec{X}, \vec{Y}) \geq f_{1-\varepsilon} \end{cases}$$

V Student criterion (SC) *t-test*

Let $\vec{X} = \{X_1, X_2, .., X_n\} \in N(a_1, \sigma^2)$ and $\vec{Y} = \{Y_1, Y_2, .., Y_m\} \in N(a_2, \sigma^2)$ are independent with the same variance; $H_1 = \{a_1 = a_2\}$

Theorem *(without proof)*

Introduced random variable t_{n+m-2}. In according to the following expression

$$t_{n+m-2} = \sqrt{\frac{n\,m}{n+m}}\,\frac{(\bar{X}-a_1)-(\bar{Y}-a_2)}{\sqrt{\frac{(n-1)S_0^2(\vec{X})+(m-1)S_0^2(\vec{Y})}{n+m-2}}},$$

random variable has a Student distribution with n+m-2 random of freedom,

because $D(\bar{X}-a_1)-(\bar{Y}-a_2)=\frac{n+m}{nm}\sigma^2$; and according to Lemma Fisher:

$(n-1)S_0^2(\vec{X})/\sigma^2 \in H_{n-1}$ and $(m-1)S_0^2(\vec{Y})/\sigma^2 \in H_{m-1}$

so their sum $\frac{(n-1)S_0^2(\vec{X})}{\sigma^2}+\frac{(m-1)S_0^2(\vec{Y})}{\sigma^2} \in H_{n+m-2}$

Introduce criteria statistics

$$\rho(\vec{X},\vec{Y}) = \sqrt{\frac{n\,m}{n+m}}\,\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{(n-1)S_0^2(\vec{X})+(m-1)S_0^2(\vec{Y})}{n+m-2}}}$$

here $\bar{X},\bar{Y}$ – average (mean).

If H_1 – true, then $\rho = t_{n+m-2} \in T_{n+m-2}$, and quantile of T_{n+m-2} – symmetrical distribution is $C = \tau_{1-\varepsilon/2}$ and Student criterion $\delta(\vec{X},\vec{Y})$ is:

$$\delta(\vec{X},\vec{Y}) = \begin{cases} H_1, & \text{if } |\rho(\vec{X},\vec{Y})| < C \\ H_2, & \text{if } |\rho(\vec{X},\vec{Y})| \geq C \end{cases}$$

VI One-dimensional analysis of variance (ODAV)

ODAV solves the problem of *checking the equality of the means of several independent normally distributed samples with the same variances*

(for two samples we suggest using a t-test).

For given k-samples:

$$X^{(1)} = \left(X_1^{(1)}, X_2^{(1)}, .., X_{n_1}^{(1)}\right), .., X^{(k)} = \left(X_1^{(k)}, X_2^{(k)}, .., X_{n_k}^{(k)}\right)$$

$X_i^{(j)} \in N(a_j, \sigma^2)$. The main hypothesis: $H_1 = \{a_1 = a_2 = \cdots = a_k\}$. For each sample $\bar{X}^{(j)} = \frac{1}{n_j}\sum X_i^{(j)}$, $\left(S^{(j)}\right)^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}(X_i^{(j)} - \bar{X}^{(j)})^2$.

If $n = n_1 + \cdots + n_k$ then overall sample mean

$\bar{X} = \frac{1}{n}\sum_{i,j} X_i^{(j)} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{X}^{(j)}$; and overall sample variance

$S^2 = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_i^{(j)} - \bar{X})^2$.

The test for the H_1 is based on comparison of intra-group and inter-group variances, let's to determine:

Between groups variance is equal

$$(SSB)\ S_{\boldsymbol{B}}^2 = \frac{1}{n}\sum_{j=1}^{k} n_j\,(\bar{X}_i^{(j)} - \bar{X})^2;$$

Inter group (**E**ach group) variance:

$$(SSE)\ S_{\boldsymbol{E}}^2 = \frac{1}{n}\sum_{j=1}^{k} nS^{(j)}{}_j = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(X_i^{(j)} - \bar{X}^{(j)}\right)^2.$$

In accordance with Fisher's Lemma for $\forall j = 1,..,k$ value $n_j \frac{(S^{(j)})^2}{\sigma^2} \in H_{n_j-1}$ and doesn't depend on $\bar{X}^{(j)}$. Based on sample independence and stability on summation of χ^2, we can obtain:

$n\frac{S_E^2}{\sigma^2} = \sum_{j=1}^{k} \frac{n_j(S^{(j)})^2}{\sigma^2} \in H_{n-k}$, here $n - k = n_1 - 1 + \cdots + n_k - 1$ and S_E^2 doesn't depend on $\bar{X}^{(1)}, \ldots, \bar{X}^{(k)}$ and as a result, it does not depend on $\bar{X}$ and on $S_{\boldsymbol{B}}^2$ (because $S_{\boldsymbol{B}}^2$ function of $\bar{X}^{(1)}, \ldots, \bar{X}^{(k)}$ and $\bar{X}$).

These properties are true always and regardless of the true values of equality or inequality (also H_1).

Let H_1 – true. Consider all k samples as single volume of n size. Variances S^2, S_B^2, S_E^2 satisfy dispersion relation (easy to check):

$$\frac{nS^2}{\sigma^2} = \frac{nS_B^2}{\sigma^2} + \frac{nS_E^2}{\sigma^2}$$

and $\frac{nS^2}{\sigma^2} \in H_{n-1}$, $\frac{nS_B^2}{\sigma^2} \in H_{k-1}$, $\frac{nS_E^2}{\sigma^2} \in H_{n-k}$; [$\frac{nS_B^2}{\sigma^2} \in H_{k-1} - \ let's\ show!$]

Lemma *(without proof).*

Let random variable ξ and η are independent and $\xi \in H_m$, $\xi + \eta \in H_s$, then $\eta \in H_{s-m}$

H_1 – true, according to Lemma we obtain two Chi-square distributions of independent random variables:

$$\chi^2 = \frac{nS_B^2}{\sigma^2} \in H_{k-1} \text{ and } \psi^2 = \frac{nS_E^2}{\sigma^2} \in H_{n-k}.$$

Let's build Fisher statistics $F_{k-1,n-k}$:

$$\rho = \frac{\chi^2}{k-1}\frac{n-k}{\psi^2} = \frac{n-k}{k-1}\frac{S_B^2}{S_E^2} \in F_{k-1,n-k}.$$

For given $\varepsilon > 0$ find quantile C level $1-\varepsilon$ of Fisher distribution $F_{k-1,n-k}$ and formulate the follow criterion of exact size.

$$\delta = \begin{cases} H_1, & \text{if } \rho < C \\ H_2, & \text{if } \rho \geq C \end{cases}$$

VII Criterion of independence χ^2

For sample $(X_1, Y_1), \dots, (X_n, Y_n)$ of two jointly observed random variables X,Y in *n* independent experiments

$$H_1 = \{X, Y - independent\}$$

Introduce *k* – grouping intervals $A_1, \dots, A_k$ for X and *m* – grouping intervals $B_1, \dots, B_m$ for Y and frequencies v_{ij} – number $(X,Y) \in A_i \times B_j$, $v_{.j}$ – number $(X,Y) \in B_j$, $v_{i.}$ – number $(X,Y) \in A_i$ are used in the table:

$\vec{Y}$ / $\vec{X}$	$B_1 B_2 \dots B_m$	$\sum_{j=1}^{m}$
A_1 $\vdots$ A_k	v_{11} v_{12} v_{1m} $\dots$ $\dots$ $\dots$ v_{k1} v_{k2} v_{km}	$v_{1.}$ $\vdots$ $v_{k.}$
$\sum_{i=1}^{k}$	$v_{.1}$ $v_{.2}$ $v_{.m}$	

If H_1 – true and $(X,Y) \in A_i \times B_j$ then probability of $(X,Y) \in A_i \times B_j$ is equal to

$$p_{ij} = P\left((X,Y) \in A_i \text{ x } B_j\right) = P(X \in A_i)\, p\left(Y \in B_j\right) = p_i^X\, p_j^Y$$

This H really check; denoted it H_1'; according to LLN: $\frac{v_{i\cdot}}{n} \xrightarrow{P} p_i^X$, $\frac{v_{\cdot j}}{n} \xrightarrow{P} p_j^Y$, $\frac{v_{ij}}{n} \xrightarrow{p} p_{ij}$, so large difference between $\frac{v_{ij}}{n}$ and $\frac{v_{i\cdot}}{n}\,\frac{v_{\cdot j}}{n}$ is the reason for rejection H_1 (about independence). Let

$$\rho\left(\vec{X},\vec{Y}\right) = n\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\left(v_{ij} - \frac{\left(v_{i\cdot}\, v_{\cdot j}\right)}{n}\right)^2}{v_{i\cdot}\, v_{\cdot j}}.$$

Theorem. If H_1 – true, then if $n \to \infty$ $\rho\left(\vec{X},\vec{Y}\right) => H_{(k-1)(m-1)}$ and criterion built as usual: for given $\varepsilon > 0$, find C – quantile of level $1 - \varepsilon$ for distribution $H_{(k-1)(m-1)}$ and

$$\delta\left(\vec{X},\vec{Y}\right) = \begin{cases} H_1, & \text{if } \rho\left(\vec{X},\vec{Y}\right) < C \\ H_2, & \text{if } \rho\left(\vec{X},\vec{Y}\right) \geq C \end{cases}$$

Remark. *Number k and m enough large!*

VIII Simple hypotheses

(A): About mean of normal distributed random variable with known variance*:* Sample $\vec{X} = \{X_1, X_2, .., X_n\} \in N(a_1, \sigma^2)$, $H_1 = \{a = a_o\}$, $H_2 = \{a \neq a_0\}$ and clear (why?) that statistics

$\rho(\vec{X}) = \sqrt{n}\frac{\bar{x} - a_0}{\sigma}$ ($if\ H_1 - \text{true},\ \text{then}\ \rho(\vec{X}) \in N(0,1)$ and criterion of consent is

$$\delta\left(\vec{X}\right) = \begin{cases} H_1, \left|\rho(\vec{X})\right| < C \\ H_2, \left|\rho(\vec{X})\right| \geq C \end{cases} \qquad C = \tau_{1-\frac{\varepsilon}{2}} \qquad (*)$$

(B): About mean of normal distributed random variable with *unknown* variance*:*

$\rho(\vec{X}) = \sqrt{n}\frac{\bar{x}-a_0}{\sqrt{S_0^2}}$,

if $a = a_0$ then according to consequence of Fisher Lemma $\rho(\vec{X}) \in T_{n-1}$;

$\delta(\vec{X})$ – the same as (*) and $C = \tau_{1-\frac{\varepsilon}{2}}$ – quantile of Student distribution.