DATE STATISTICE

MODELE STOCASTICE

TESTE DE CONCORDANTA (goodness-of-fit)

$$\begin{array}{ccccc} & \swarrow & \text{fenomen aleator} & \searrow & \\ \text{date statistice} & & & & \text{model stocastic} \\ & \searrow & \text{test de concordanta} & \swarrow & \end{array}$$

Fenomene aleatoare

- prin natura lor; Exemple din biologie, medicina, finante
- prin modul de colectare a datelor; Exemple din sondaje statistice

(A) DATE STATISTICE

1. Valori calitative;

Exemplu: intrebare cu raspunsuri posibile "f. nemultumit", "nemultumit", "indiferent", "multumit", "foarte multumit"
n indivizi independenti, alesi in mod aleator dintr-o aceeasi categorie, raspund la intrebare

```
> rasp=c("fnem","nem","ind","mul","fmul")
> p=c(0.2,0.3,0.1,0.3,0.1)
> x<-sample(rasp,50,replace=T,prob=p)
> x
"fmul" "ind" "mul" "mul" "nem" "nem" "fmul" "nem" "nem"
"nem" "fnem" "fnem" "nem" "nem" "nem" "mul" "fnem" "fnem"
"fnem" "nem" "fnem" "mul" "fnem" "fnem" "mul" "nem" "nem"
"mul" "nem" "mul" "mul" "ind" "fmul" "mul" "fmul" "fnem" "nem"
"nem" "fmul" "nem" "mul" "fnem" "mul" "nem" "nem" "fnem"
"nem" "fnem" "ind" "nem"
```

2. Valori cantitative

- apartinand unei multimi cel mult numarabile de numere reale
- apartinand lui R sau unui interval inclus in R

Exemplu: nota obtinuta la un examen (0 = absent)
n indivizi independenti, alesi in mod aleator dintr-o aceeasi categorie

```
> nota=c(0:10)
> p=c(0.05,0,0,0,0.3,0.2,0.15,0.1,0.05,0.1,0.05)
> y<-sample(nota,25,replace=T,prob=p)
> y
4 6 8 4 4 6 5 5 9 7 8 4 6 9 4 8 4 4 4 7 5 5 6 5 7
```

Exemplu: tensiunea arteriala sistolica
n indivizi independenti, alesi in mod aleator dintr-o aceeasi categorie

```
> z<-c(rnorm(50,13,1.5))
> z
```

11.4, 14.2, 14.9, 12.5, 12.8, 13.8, 10.7, 13.1, 15.1, 11.4, 11.6, 15.5, 11.8, 12.9, 15.3, 13.7, 13.5, 11.8, 11.9, 12.9, 13.3, 14.2, 14.5, 12.7, 12.4, 13.7, 10.9, 15.4, 14.1, 9.4, 12.5, 11.7, 13.2, 14.9, 14.5, 13.5, 12.5, 13.8, 13.3, 12.8, 10.5, 12.1, 13.5, 14.6, 10.7, 12.1, 10.9, 11.5, 11.7, 11.1

Statistica descriptiva (pt datele statistice)

1. Repartitia de frecvente

valori distincte x	"fnem"	"nem"	"ind"	"mul"	"fmul"
frecvente	$\frac{12}{50}$	$\frac{19}{50}$	$\frac{3}{50}$	$\frac{11}{50}$	$\frac{5}{50}$

valori distincte y	0	1	2	3	4	5	6	7	8	9	10
frecvente	0	0	0	0	$\frac{8}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	0

2. Histograma

interv val z	[9,10)	[10,11)	[11,12)	[12.13)	[13,14)	[14.15)	[15,16]
frecv cum	$\frac{1}{50}$	$\frac{5}{50}$	$\frac{10}{50}$	$\frac{11}{50}$	$\frac{11}{50}$	$\frac{8}{50}$	$\frac{4}{50}$

package:........graphics.........R Documentation

Description: The generic function 'hist' computes a histogram of the given data values. If 'plot=TRUE', the resulting object of 'class "histogram"' is plotted by 'plot.histogram', before it is returned.

Usage: hist(x, ...)

Arguments: x: a vector of values for which the histogram is desired.

3. Indicatori de pozitie (date cantitative)

Datele $(x_1, ..., x_n)$

Datele ordonate $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$

Minim, maxim, cuartile

$$\begin{aligned} x_{(1)} &= \min_i x_i \\ x_{(n)} &= \max_i x_i \end{aligned}$$

$$Q_2 = Me = \begin{cases} x_{(k+1)}, & n = 2k+1 \\ \frac{1}{2}\left(x_{(k)} + x_{(k+1)}\right), & n = 2k \end{cases}$$

$$Q_1 = \text{mediana pt. } x_{(1)} \leq ... \leq Me$$

$$Q_3 = \text{ mediana pt. } Me \leq ... \leq x_{(n)}$$

Media (de selectie)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

> x

3, 4, 6, 5, 5, 7, 3, 5, 6, 4, 5, 7, 4, 3, 2, 4, 4, 5, 7, 5, 6, 4, 5, 2, 6, 4, 8, 6, 7, 5, 7, 4, 4, 2, 3, 2, 0, 1, 4, 4, 3, 7, 5, 7, 4, 3, 7, 2, 5, 5, 7, 5, 7, 7, 5, 4, 4, 7, 3, 8, 5, 6, 5, 6, 5, 6, 4, 5, 8, 2, 6, 4, 6, 5, 5, 5, 3, 5, 4, 3, 7, 7, 2, 4, 5, 4, 6, 5, 3, 1, 5, 7, 4, 5, 3, 3, 10, 6, 7, 6

> summary(x)

Min........1st Qu...... Median....... Mean....... 3rd Qu......... Max.

0.00 ....4.00 ....5.00 ....4.81 ....6.00 ....10.00

4. Indicatori de variabilitate (date cantitative)

Amplitudinea

$$a = x_{(n)} - x_{(1)}$$

Dispersia de selectie, abaterea standard

$$\begin{aligned} s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ s &= \sqrt{s^2} \end{aligned}$$

Functii din R

```
> mean(x)
[1] 4.81
```

```
> var(x)
[1] 3.165556
```

```
> sd(x)
[1] 1.779201
```

5. Indicatori ai formei (date cantitative)

Notam momentele de selectie centrate, de ordin 3 si 4 cu

$$\begin{aligned} \overline{m_3} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3 \\ \overline{m_4} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4 \end{aligned}$$

Coeficient de asimetrie (skewness)

$$\beta_1 = \frac{\overline{m_3}}{\sqrt{(s^2)^3}}$$

Coeficient de aplatizare (kurtosis)

$$\beta_2 = \frac{\overline{m_4}}{(s^2)^2} - 3$$

(B) MODELE STOCASTICE (variabile aleatoare)

$(\Omega, \mathcal{K}, P_\theta)$, $\theta \in \Theta \sqsubseteq R^k$, $k \geq 1$;
Spatiul starilor (al valorilor) $(S, \mathcal{S})$
$S = A \subset R$, A cel mult numarabila;.....$(A, \mathcal{P}(A))$
$S = R; ..... (R, \mathcal{B})$
Variabila aleatoare = functie masurabila $X : \Omega \longrightarrow S$

1. Repartitia lui X

$$P_\theta \circ X^{-1} : \mathcal{S} \longrightarrow [0, 1]$$

Variabila aleatoare cu repartitie discreta

$$\begin{aligned} \left(P_\theta \circ X^{-1}\right)(\{x\}) &= p(x; \theta) \in [0, 1], \quad x \in A \\ P_\theta \circ X^{-1} &= \sum_{x \in A} p(x; \theta) \cdot \delta_{\{x\}} \\ \sum_{x \in A} p(x; \theta) &= 1 \end{aligned}$$

Exemple:

- $X \sim U\{1, ..., r\}$, $r \in N$, $r \geq 2$, $A = \{1, 2, ..., r\}$ (ex: numarul de puncte la aruncarea unui zar),

$$P_\theta \circ X^{-1} = \sum_{x=1}^{r} \frac{1}{r} \cdot \delta_{\{x\}}$$

- $X \sim B(1, \theta)$, $\theta \in (0, 1)$, $A = \{0, 1\}$ (ex: aparitia unui "succes" intr-o proba cu doua rezultate posibile),

$$P_\theta \circ X^{-1} = \sum_{x=0}^{1} \theta^x (1 - \theta)^{1-x} \cdot \delta_{\{x\}}$$

- $X \sim B(r, \theta)$, $\theta \in (0, 1)$, $A = \{0, 1, ..., r\}$ (ex: numarul de "succese" in r probe independente, cu cate doua rezultate posibile),

$$P_\theta \circ X^{-1} = \sum_{x=0}^{r} C_r^x \cdot \theta^x (1 - \theta)^{r-x} \cdot \delta_{\{x\}}$$

- $X \sim Po(\theta),\ \theta \in (0,\infty),\ A = N$ (ex: numarul de defecte ce pot fi identificate la piesele dintr-un lot de volum mare),

$$P_\theta \circ X^{-1} = \sum_{x=0}^{\infty} \frac{\theta^x}{x!} \exp(-\theta) \cdot \delta_{\{x\}}$$

Variabila aleatoare cu repartitie continua si cu densitate de repartitie

$$\begin{aligned}
\left(P_\theta \circ X^{-1}\right)(\{x\}) &= 0,\ \forall x \in R \\
\left(P_\theta \circ X^{-1}\right)(B) &= \int_B f(x;\theta)\,dx, \\
f(x;\theta) &\geq 0,\ \forall x \in R \\
\int_R f(x;\theta)\,dx &= 1
\end{aligned}$$

Exemple:

- $X \sim U(0,\theta),\ \theta \in (0,\infty),$

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}, & x \in [0,\theta] \\ 0, & x \notin [0,\theta] \end{cases}$$

- $X \sim Expo(\theta),\ \theta \in (0,\infty),$

$$f(x;\theta) = \begin{cases} \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right), & x \in [0,\infty) \\ 0, & x \in (-\infty,0) \end{cases}$$

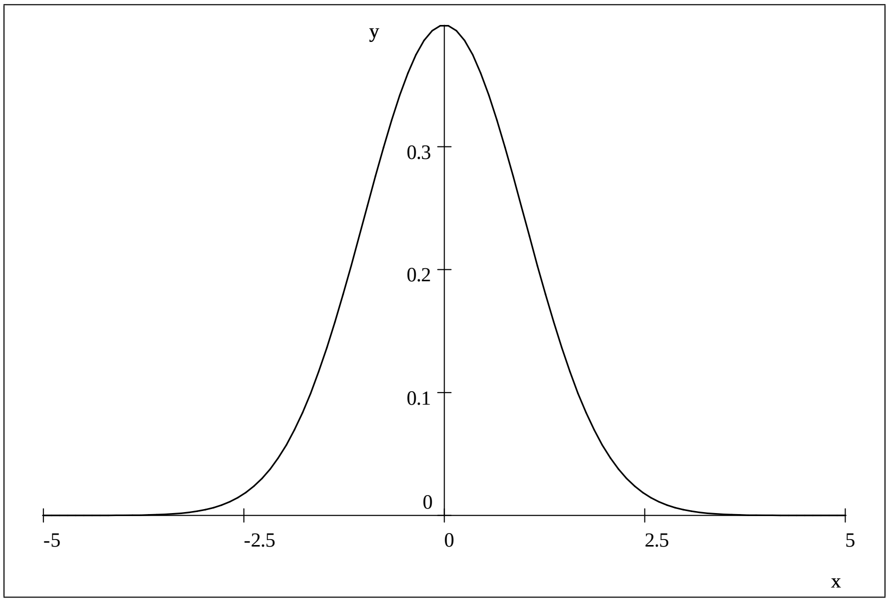
- $X \sim Gamma(\alpha,\theta),\ \alpha \in (0,\infty),\ \theta \in (0,\infty),$

$$f(x;\alpha,\theta) = \begin{cases} \frac{1}{\Gamma(\alpha)\cdot\theta^\alpha} \cdot x^{\alpha-1} \cdot \exp\left(-\frac{x}{\theta}\right), & x \in [0,\infty) \\ 0, & x \in (-\infty,0) \end{cases}$$

- $X \sim N\left(\mu,\sigma^2\right),\ \theta = \left(\mu,\sigma^2\right) \in R \times (0,\infty),$

$$f\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right), \quad x \in R$$

densitatea $N(0,1)$
$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$



2. Functia de repartitie a lui X

$$\begin{aligned} F_\theta &: R \longrightarrow [0,1] \\ F_\theta(y) &= \left(P_\theta \circ X^{-1}\right)((-\infty, y)) = P_\theta(X < y) \end{aligned}$$

$$F_\theta(y) = \sum_{\substack{x \in A \\ x < y}} p(x;\theta), \ y \in R, \ \text{(functie in scara)}$$

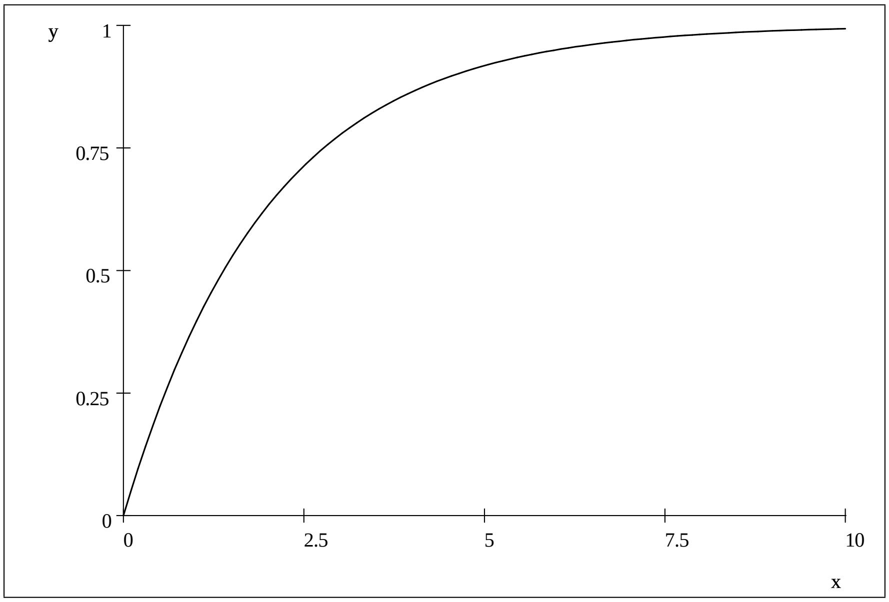
$$F_\theta(y) = \int_{-\infty}^{y} f(x;\theta)\,dx, \ y \in R$$

Exemplu:
$X \sim Expo(2)$

$$f(x) = \begin{cases} \frac{1}{2}\exp\left(-\frac{x}{2}\right), & x \in [0,\infty) \\ 0, & x \in (-\infty, 0) \end{cases}$$

$$F_\theta(y) = \begin{cases} 0, & x \in (-\infty,0) \\ \int_0^y \frac{1}{2}\exp\left(-\frac{x}{2}\right)dx, & x \in [0,\infty) \end{cases} = \begin{cases} 0, & x \in (-\infty,0) \\ 1 - \exp\left(-\frac{x}{2}\right), & x \in [0,\infty) \end{cases}$$

$1 - \exp\left(-\frac{x}{2}\right)$



3. Cuantila de rang α a lui X

Fie $\alpha \in (0,1)$ fixat.
Notam $q_\alpha \in S$ cu proprietatea

$$\begin{aligned} P_\theta(X < q_\alpha) &\leq \alpha \\ P_\theta(X \leq q_\alpha) &\geq \alpha \end{aligned}$$

Pentru modelele cu repartitie continua,

$$P_\theta(X < q_\alpha) = P_\theta(X \leq q_\alpha) = \alpha$$

4. Medie, momente; dispersie

$$M_\theta(X) = \int_\Omega X dP_\theta = \begin{cases} \sum\limits_{x \in A} x \cdot p(x;\theta), & (<\infty), \quad \text{pt. rep. discreta} \\ \int\limits_R x \cdot f(x;\theta)\,dx, & (<\infty), \quad \text{pt. rep. continua} \end{cases}$$

$$M_\theta(X^r) = \int_\Omega X^r dP_\theta = \begin{cases} \sum\limits_{x \in A} x^r \cdot p(x;\theta), & (<\infty), \quad \text{pt. rep. discreta} \\ \int\limits_R x^r \cdot f(x;\theta)\,dx, & (<\infty), \quad \text{pt. rep. continua} \end{cases}, \quad r \in N^*$$

$$D_\theta^2(X) = M_\theta\left((X - M_\theta(X))^2\right) = M_\theta(X^2) - (M_\theta(X))^2$$

Exemple:

- $X \sim U\{1, ..., r\},\ r \in N,\ r \geq 2,$

$$\begin{aligned} M(X) &= \sum_{x=1}^{r} x \cdot \frac{1}{r} = \frac{r+1}{2} \\ D^2(X) &= \frac{r^2-1}{12} \end{aligned}$$

- $X \sim B(1, \theta),\ \theta \in (0,1),$

$$\begin{aligned} M_\theta(X) &= \sum_{x=0}^{1} x \cdot \theta^x (1-\theta)^{1-x} = \theta \\ D_\theta^2(X) &= \theta(1-\theta) \end{aligned}$$

- $X \sim B(r, \theta),\ \theta \in (0,1),$

$$\begin{aligned} M_\theta(X) &= \sum_{x=0}^{r} x \cdot C_r^x \cdot \theta^x (1-\theta)^{r-x} = r\theta \\ D_\theta^2(X) &= r\theta(1-\theta) \end{aligned}$$

- $X \sim Po(\theta),\ \theta \in (0,\infty),$

$$\begin{aligned} M_\theta(X) &= \sum_{x=0}^{\infty} x \cdot \frac{\theta^x}{x!} \exp(-\theta) = \theta \\ D_\theta^2(X) &= \theta \end{aligned}$$

- $X \sim U(0, \theta),\ \theta \in (0,\infty),$

$$\begin{aligned} M_\theta(X) &= \int_0^\theta x \cdot \frac{1}{\theta} dx = \frac{\theta}{2} \\ D_\theta^2(X) &= \frac{\theta^2}{12} \end{aligned}$$

- $X \sim Expo(\theta), \theta \in (0,\infty),$

$$\begin{aligned} M_\theta(X) &= \int_0^\infty x \cdot \frac{1}{\theta} \exp\left(-\frac{x}{\theta}\right) dx = \theta \\ D_\theta^2(X) &= \theta^2 \end{aligned}$$

- $X \sim Gamma(\alpha, \theta)$, $\alpha \in (0, \infty)$, $\theta \in (0, \infty)$,

$$\begin{aligned} M_\theta(X) &= \frac{1}{\Gamma(\alpha) \cdot \theta^\alpha} \int_0^\infty x \cdot x^{\alpha-1} \cdot \exp\left(-\frac{x}{\theta}\right) dx = \alpha\theta \\ D_\theta^2(X) &= \alpha\theta^2 \end{aligned}$$

- $X \sim N(\mu, \sigma^2)$, $\theta = (\mu, \sigma^2) \in R \times (0, \infty)$,

$$\begin{aligned} M_\theta(X) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \cdot \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) dx = \mu \\ D_\theta^2(X) &= \sigma^2 \end{aligned}$$

5. Functie generatoare, functie caracteristica

Fie $P_\theta \circ X^{-1} = \sum_{x=0}^{\infty} p(x;\theta) \cdot \delta_{\{x\}}$. Functia generatoare asociata este

$$G_X : [-1, 1] \longrightarrow R$$
$$G_X(t) = \sum_{x=0}^{\infty} p(x;\theta) \cdot t^x$$

Pentru variabile cu medie (dispersie) finita se verifica relatiile

$$\begin{aligned} M_\theta(X) &= G_X'(1) \\ D_\theta^2(X) &= G_X''(1) + G_X'(1) - (G_X'(1))^2 \end{aligned}$$

Fie variabila aleatoare X, cu valori in R. Functia caracteristica asociata este

$$\varphi_X : R \longrightarrow C$$
$$\varphi_X(t) = M_\theta\left(e^{itX}\right)$$

Daca repartitia $P_\theta \circ X^{-1}$ are densitatea de repartitie $f(x;\theta)$, atunci

$$\varphi_X(t) = \int_R e^{itx} \cdot f(x;\theta)\, dx$$

Pentru variabile cu medie (dispersie) finita se verifica relatiile

$$\begin{aligned} M_\theta(X) &= \frac{1}{i} \cdot \varphi_X'(0) \\ D_\theta^2(X) &= -\varphi_X''(0) + (\varphi_X'(0))^2 \end{aligned}$$

6. Transformata Laplace

Fie variabila aleatoare X, cu valori in R_+. Transformta Laplace asociata este

$$\psi : R_+ \longrightarrow R_+$$
$$\psi(\lambda) = M\left(e^{-\lambda X}\right)$$

Daca repartitia $P_\theta \circ X^{-1}$ pe $(R_+, \mathcal{B}_+)$ are densitatea de repartitie $f(x;\theta)$ pentru $x \geq 0$, atunci

$$\psi(\lambda) = \int_0^\infty e^{-\lambda x} f(x;\theta)\, dx$$

(C) CONCORDANTA DINTRE DATE STATISTICE / MODEL STOCASTIC

Datele statistice sunt valori observate ale unor variabile aleatoare independente, identic repartizate, cu repartitia data de un model stocastic.

Analiza de statistica descriptiva ne permite sa alegem un model stocastic - drept sursa posibila a datelor statistice.

Consideram modelul stocastic reprezentat de variabila aleatoare X cu repartitia $P_\theta \circ X^{-1}$ complet specificata. Neglijam indicele θ, căci presupunem cunoscuta valoarea parametrului.

- Fie modelul stocastic dat de variabila aleatoare X cu repartitia $P \circ X^{-1}$ si functia de repartitie $F(y)$.
- Fie "observatiile" $X_1, ..., X_n$, care sunt variabile aleatoare independente, identic repartizate, cu repartitia $P \circ X^{-1}$
- Fie datele statistice $(x_1, ..., x_n) = (X_1, ..., X_n)(\omega)$

Problema: Putem confirma ipoteza ca datele statistice $(x_1, ..., x_n)$ furnizate de un beneficiar provin intr-adevar din modelul considerat?

Vom compara functia de repartitie "teoretica" $F(y)$ cu o functie construita din datele statistice $(x_1, ..., x_n)$.

Spatiul de selectie n−dimensional

Fie modelul stocastic $P_\theta \circ X^{-1}$ cu multimea valorilor lui X egala cu $S = A$ (cel mult numarabila) sau cu $S = R$.

Fie observatiile $X_1, ..., X_n$ v.a.i.i.r. $(P_\theta \circ X^{-1})$.

Spatiul de selectie n−dimensional este campul de probabilitate construit pe multimea valorilor lui $(X_1, ..., X_n)$:

$$\left(A^n, (\mathcal{P}(A))^n, \bigotimes_{i=1}^{n} P_\theta \circ X_i^{-1}\right)$$

$$\left(R^n, \mathcal{B}^n, \bigotimes_{i=1}^{n} P_\theta \circ X_i^{-1}\right)$$

Functia de repartitie de selectie (empirica)

Fie functia de repartitie complet specificata, $F(y)$, pentru variabila aleatoare $X:\Omega \longrightarrow S$.
Fie observatiile $X_1,...,X_n$ v.a.i.i.r. ca si X.

DEFINITIE: Functia de repartitie de selectie

$$F_n(\cdot,\cdot): R\times\Omega \longrightarrow [0,1]$$

$$F_n(y,\omega) = \frac{1}{n}\cdot card\{i \mid i\in\{1,...,n\}, x_i = X_i(\omega) < y\}$$

Observatie:

$$F_n(y,\omega) = \frac{1}{n}\cdot\sum_{i=1}^{n} I_{\{X_i<y\}}(\omega)$$

PROPRIETATEA 1
Pentru ω arbitrar fixat, $F_n(\cdot,\omega)$ este functia de repartitie a unei repartitii Uniforme discrete

$$\sum_{i=1}^{n}\frac{1}{n}\cdot\delta_{\{x_i\}}$$

Demonstratie:
Notam $(X_1,...,X_n)(\omega) = (x_1,...,x_n)$ valori fixate (pentru ω fixat).
Notam cu Z o variabila aleatoare cu repartitia uniforma data de

$$P(Z=x_i) = \frac{1}{n},\quad i=1,...,n$$

$$F_Z(y) = P(Z<y) = \sum_{x_i<y}\frac{1}{n} = \frac{1}{n}\cdot\sum_{i=1}^{n} I_{\{x_i<y\}} = F_n(y,\omega)$$

■

PROPRIETATEA 2
Pentru y arbitrar fixat, $F_n(y,\cdot)$ este variabila aleatoare cu proprietatea

$$n\cdot F_n(y,\cdot)\sim B(n,F(y))$$

Demonstratie:
Pentru $\forall i,\ I_{\{X_i<y\}}$ este v.a. cu valori in $\{0,1\}$ si cu

$$P\left(I_{\{X_i<y\}}=1\right) = P(X_i<y) = F(y)$$

adica

$$I_{\{X_i<y\}} \sim B(1, F(y))$$

Avem $\{I_{\{X_i<y\}}, i=1,...,n\}$ v.a. indep, id. rep $B(1, F(y))$.
Rezulta

$$\begin{aligned} \sum_{i=1}^{n} I_{\{X_i<y\}} &\sim B(n, F(y)) \\ n \cdot F_n(y,\cdot) &\sim B(n, F(y)) \end{aligned}$$

■

COROLAR

$$\begin{aligned} M(F_n(y,\cdot)) &= F(y) \\ D^2(F_n(y,\cdot)) &= \frac{1}{n} F(y)(1-F(y)) \end{aligned}$$

PROPRIETATEA 3

Pentru y arbitrar fixat, sirul de var. al. $\{F_n(y,\cdot), n=1,2,...\}$ are proprietatea

$$F_n(y,\cdot) \overset{P-a.s.}{\longrightarrow} F(y) \text{ pentru } n \longrightarrow \infty$$

Demonstratie

Avem sirul $\{I_{\{X_i<y\}}, i=1,...,n\}$ de v.a. indep, id. rep $B(1, F(y))$, avand $M\left(I_{\{X_1<y\}}\right) = F(y)$. Aplicam legea tare a numerelor mari:

$$\frac{1}{n} \cdot \sum_{i=1}^{n} I_{\{x_i<y\}} \overset{P-a.s.}{\longrightarrow} M\left(I_{\{X_1<y\}}\right) = F(y) \text{ pentru } n \longrightarrow \infty$$

■

Spunem ca functia de repartitie de selectie este un estimator consistent si nedeplasat la functiei de repartitie pt modelul din care provin datele statistice.

Functii din R: functia $ecdf$ ploteaza functia de repartitie de selectie

$> data < -c(x_1, ..., x_n)$
$> ecdf(data)$

"Distanta" Kolmogorov dintre functia de repartitie de selectie si functia de repartitie a modelului

$$D_n(\omega) = \sqrt{n} \cdot \sup_{y \in R} |F_n(y, \omega) - F(y)|$$

Pentru datele statistice $(X_1, ..., X_n)(\omega) = (x_1, ..., x_n)$, se poate calcula valoarea

$$\widetilde{D_n} = \sqrt{n} \cdot \max_{1 \le i \le n} |F_n(x_i, \omega) - F(x_i)|$$

TEOREMA LUI KOLMOGOROV

Fie modelul probabilist dat de o variabila aleatoare X, cu functia de repartitie $F(y)$ continua. Daca $\{X_n, n \geq 1\}$ este un sir de variabile aleatoare independente, identic repartizate ca si X pentru care notam $\{F_n(y, \omega), n \geq 1\}$ sirul functiilor de repartitie de selectie atunci, pentru orice $z \in R$, are loc convergenta

$$\lim_{n \to \infty} P(D_n < z) = K(z),$$

unde $K(z)$ este functia de repartitie Kolmogorov,

$$K(z) = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j-1} \exp\left(-2j^2 z^2\right)$$

Pentru demonstratie:
PARTHASARATHY, K., R., Probability measures on metric spaces, Academic Press, 1967.

TESTUL LUI KOLMOGOROV DE CONCORDANTA (R:.....ks.test for one sample)

Fie datele statistice $(x_1, ..., x_n)$ si fie modelul stocastic dat de variabila aleatoare X cu functia de repartitie $F(y)$ continua.

Pentru $\alpha \in (0, 1)$ arbitrar fixat, notam $z_{1-\alpha}$ cuantila de rang $(1 - \alpha)$ a repartitiei Kolmogorov,

$$K(z_{1-\alpha}) = 1 - \alpha$$

Formulam ipoteza H :{variabilele aleatoare independente si identic repartizate $X_1, ..., X_n$ care au generat datele statistice au functia de repartitie $F(y)$}

Algoritm:

- Se ordoneaza datele statistice, $x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$
- Se calculeaza $F(x_{(i)})$ si $F_n(x_{(i)}, \omega)$, $i = 1, ..., n$
- Se calculeaza $\widetilde{D_n} = \sqrt{n} \cdot \max\limits_{1 \leq i \leq n} \left| F_n(x_{(i)}, \omega) - F(x_{(i)}) \right|$
- Regula de decizie: Daca $\widetilde{D_n} \geq z_{1-\alpha}$, decidem sa respingem ipoteza H (nu avem concordanta intre model si datele statistice)

Comentariu: Testul se bazeaza pe teorema lui Kolmogorov (este un test asimptotic), deci n trebuie sa fie mare $(n \geq 100)$
======================

APLICATIE: TESTAREA NORMALITATII DATELOR

Input : $(x_1, ..., x_n) = (X_1, ..., X_n)(\omega)$

H : { variabilele aleatoare independente $X_1, ..., X_n$ au repartitie normala }

(a) Partea exploratorie

> data $\longleftarrow c(x_1, ..., x_n)$
> mean(data)
> var(data)
> hist(data)

qq - line (quantile - quantile line)

$$X \sim N\left(\mu, \sigma^2\right) \quad \Leftrightarrow \quad \frac{X-\mu}{\sigma} \sim N(0,1)$$

$$F_{N(\mu,\sigma^2)}\left(x_\alpha\right) = \alpha \quad \Leftrightarrow \quad F_{N(0,1)}\left(\frac{x_\alpha - \mu}{\sigma}\right) = \alpha$$

$$z_\alpha = \frac{1}{\sigma}\left(x_\alpha - \mu\right), \quad \alpha \in (0,1)$$

> qqnorm(data)
> qqline(data)

(b) Test de concordanta

Pentru a utiliza ks.test (for one sample) trebuie sa specificam valorile (μ, σ^2)

> ks.test(data)

$$p - value = 1 - K\left(\widetilde{D_n}\right)$$

$p - value \leq 0.05 \quad \longrightarrow \quad$ respingem ipoteza H (respingem normalitatea)

Observatie: Exista o varianta a testului, testul Lilliefors, in care programul isi alege singur valorile

$$\begin{aligned} \mu &= mean(data) \\ \sigma &= sd(data) \end{aligned}$$

Alt test de concordanta este "Testul Chi Patrat", construit pentru modele stocastice $P \circ X^{-1}$ avand functia de repartitie $F(y)$ continua sau nu.

AUXILIAR: Convergenta in repartitie

Notam cu $\{\mu_n,\ n \geq 1\}$ si μ probabilitati pe $(R, \mathcal{B})$ (repartitii)

Notam cu $\{F_n,\ n \geq 1\}$ si F functiile de repartitie corespunzatoare,

$$\begin{aligned} F_n(y) &= \mu_n(-\infty, y) \\ F(y) &= \mu(-\infty, y) \end{aligned}$$

Notam cu $\{\varphi_n,\ n \geq 1\}$ si φ functiile caracteristice corespunzatoare,

$$\begin{aligned} \varphi_n(t) &= \int\limits_R e^{itx} d\mu_n(x) \\ \varphi(t) &= \int\limits_R e^{itx} d\mu(x) \end{aligned}$$

Pentru cazul cand $\{\mu_n,\ n \geq 1\}$ si μ sunt probabilitati pe $(R_+, \mathcal{B}_+)$, notam cu $\{\psi_n,\ n \geq 1\}$ si ψ transformatele Laplace corespunzatoare,

$$\begin{aligned} \psi_n(\lambda) &= \int\limits_{(0,\infty)} e^{-\lambda x} d\mu_n(x) \\ \psi(\lambda) &= \int\limits_{(0,\infty)} e^{-\lambda x} d\mu(x) \end{aligned}$$

DEFINITIE (convergenta slaba, sau convergenta in repartitie)

$$\mu_n \Longrightarrow \mu$$

daca

$$\int\limits_R h d\mu_n \underset{n\to\infty}{\longrightarrow} \int\limits_R h d\mu$$

pentru orice functie h continua si marginita, definita pe R cu valori in R.

TEOREMA 1

O conditie necesara si suficienta ca $\mu_n \Longrightarrow \mu$ este ca $F_n(y) \underset{n\to\infty}{\longrightarrow} F(y)$ pentru orice y care este punct de continuitate al lui F.

TEOREMA 2 (PAUL LEVY)

a) Daca $\mu_n \Longrightarrow \mu$, atunci $\varphi_n \underset{n\to\infty}{\longrightarrow} \varphi$ uniform pe orice compact din R.

b) Notam cu $\{\varphi_n,\ n \geq 1\}$ functiile caracteristice corespunzatoare repartitiilor $\{\mu_n,\ n \geq 1\}$. Daca $\varphi_n(t) \underset{n\to\infty}{\longrightarrow} \varphi(t)$ pentru orice t si φ este continua in origine, atunci exista o repartitie μ asa incat $\mu_n \Longrightarrow \mu$, iar φ este functia caracteristica pt μ.

TEOREMA 3

Fie $\{\mu_n,\ n \geq 1\}$ si μ probabilitati pe $(R_+, \mathcal{B}_+)$.

a) Daca $\mu_n \Longrightarrow \mu$, atunci $\psi_n(\chi) \underset{n\to\infty}{\longrightarrow} \psi(\lambda)$ pentru orice $\lambda \geq 0$.

b) Notam cu $\{\psi_n,\ n \geq 1\}$ transformatele Laplace corespunzatoare repartitiilor $\{\mu_n,\ n \geq 1\}$. Daca $\psi_n(\chi) \underset{n\to\infty}{\longrightarrow} \psi(\lambda)$ pentru orice $\lambda > 0$ si $\lim\limits_{\lambda\to 0} \psi(\lambda) = 1$, atunci exista o repartitie μ asa incat $\mu_n \Longrightarrow \mu$, iar ψ este transformata Laplace pt μ.

TEOREMA LIMITA CENTRALA (LINDEBERG - LEVY)

Fie $\{X_n, n \geq 1\}$ un sir de variabile aleatoare independente, identic repartizate, cu $M(X_n) = \mu\ \forall n$ si $D^2(X_n) = \sigma^2 < \infty$ $\forall n$. Notam

$$Y_n = \frac{1}{\sqrt{n\sigma^2}} \left(\sum_{i=1}^{n} X_i - n\mu \right)$$

Atunci sirul $(P \circ Y_n^{-1})_n$ converge slab la repartitia $N(0,1)$.(spunem ca sirul $\{Y_n, n \geq 1\}$ converge in repartitie la o variabila aleatoare cu repartitia $N(0,1)$)

Pentru demonstratii:

CIUCU G., TUDOR C., Teoria probabilitatilor si aplicatii, Editura Stiintifica si Enciclopedica, 1983

=================

Repartitia "CHI Patrat" cu d grade de libertate $(d \in N^*)$

$$X \sim \chi^2(d) \quad \Leftrightarrow \quad f(x) = \frac{1}{2^{d/2} \cdot \Gamma\left(\frac{d}{2}\right)} x^{d/2-1} \exp\left(-\frac{x}{2}\right), \quad x \geq 0$$

$$\varphi_{\chi^2(d)}(t) = (1-2it)^{-d/2}$$

$$\psi_{\chi^2(d)}(\lambda) = (1+2\lambda)^{-d/2}$$

Repartitia Multinomiala $M(r; p_1, ..., p_d)$

Definitie
$\mathbf{X} = (X_1, ..., X_d)' \sim M(r; p_1, ..., p_d)$ daca

$$P \circ \mathbf{X}^{-1} = \sum_{\substack{x_1,\dots,x_d=0 \\ x_1+\dots+x_d=r}}^{r} \frac{r!}{x_1!\dots x_d!} (p_1)^{x_1} \dots (p_d)^{x_d} \cdot \delta_{(x_1,\dots,x_d)}$$

unde $r \in N^*,\ p_i \in [0,1]$ pentru $i = 1, ..., d$ si $\sum_{i=1}^{d} p_i = 1$

Experiment: O urna cu bile de d culori, din care se fac r extrageri cu revenire. Vectorul aleator $\mathbf{X} = (X_1, ..., X_d)$ inregistreaza numarul de bile de fiecare culoare care au fost extrase.

Bibliografie:

Dumitrescu M, Florea D, Tudor C, Probleme de teoria probabilitatilor si statistica matematica, Editura Tehnica, 1985

======================

TEOREMA LUI PEARSON

Pentru $r \in N^*$ consideram urmatoarele variabile aleatoare:

$$\mathbf{Y}_r = (Y_{r1}, ..., Y_{rd})' \sim M(r; p_1, ..., p_d), \quad \text{cu} \quad p_i \in [0,1], \forall i, \ \sum_{i=1}^{d} p_i = 1$$

$$X_r^2 = \sum_{j=1}^{d} \frac{(Y_{rj} - rp_j)^2}{rp_j}$$

Notam repartitia lui X_r^2 cu $G_r = P \circ (X_r^2)^{-1}$. Atunci

$$G_r \underset{r \to \infty}{\Longrightarrow} \chi^2(d-1)$$

(spunem ca sirul $\{X_r^2, r \geq 1\}$ converge in repartitie la o variabila repartizata CHI Patrat cu $(d-1)$ grade de libertate).

Demonstratie (prof. Ioan Cuculescu)

In schema multinomiala (d culori, r extrageri independente) apar r partitii independente, corespunzatoare celor r extrageri,

$$\left\{A_j^{(k)}, j = 1, ..., d\right\}, \quad k = 1, ..., r$$

Notam

$$Y_{rj} = \sum_{k=1}^{r} I_{A_j^{(k)}}, \quad j = 1, ..., d$$

$$\mathbf{Z}_r = \left(\frac{Y_{r1} - rp_1}{\sqrt{rp_1}}, ..., \frac{Y_{rd} - rp_d}{\sqrt{rp_d}}\right)'$$

Atunci

$$X_r^2 = \|\mathbf{Z}_r\|^2$$

$$\psi_{X_r^2}(\lambda) = M\left(\exp\left(-\lambda \|\mathbf{Z}_r\|^2\right)\right)$$

Vom arata ca

$$\psi_{X_r^2}(\lambda) \underset{r \to \infty}{\longrightarrow} (1 + 2\lambda)^{-(d-1)/2}$$

Notam

$$\begin{aligned} \mathbf{v} &= (v_1, ..., v_d)' \\ \mathbf{t} &= (t_1, ..., t_d)' \end{aligned}$$

$$\exp\left(-\lambda\left\|\mathbf{v}\right\|^2\right)=\prod_{j=1}^{d}\exp\left(-\lambda v_j^2\right)$$

Dar

$$\exp\left(-\lambda v_j^2\right)=\varphi_{N(0,2\lambda)}\left(v_j\right)=\frac{1}{\sqrt{4\pi\lambda}}\int\limits_{-\infty}^{\infty}\exp\left(iv_jt_j\right)\cdot\exp\left(-\frac{1}{4\lambda}t_j^2\right)dt_j$$

Notand cu $<\mathbf{v},\mathbf{t}>$ produsul scalar, putem scrie

$$\exp\left(-\lambda\left\|\mathbf{v}\right\|^2\right)=\frac{1}{\left(4\pi\lambda\right)^{d/2}}\int\limits_{-\infty}^{\infty}...\int\limits_{-\infty}^{\infty}\exp\left(i<\mathbf{v},\mathbf{t}>\right)\cdot\exp\left(-\frac{1}{4\lambda}\left\|\mathbf{t}\right\|^2\right)dt_1...dt_d$$

Putem scrie

$$\psi_{X_r^2}\left(\lambda\right)=\frac{1}{\left(4\pi\lambda\right)^{d/2}}\int\limits_{-\infty}^{\infty}...\int\limits_{-\infty}^{\infty}M\left(\exp\left(i<\mathbf{Z}_r,\mathbf{t}>\right)\cdot\exp\left(-\frac{1}{4\lambda}\left\|\mathbf{t}\right\|^2\right)\right)dt_1...dt_d$$

$$=\frac{1}{\left(4\pi\lambda\right)^{d/2}}\int\limits_{-\infty}^{\infty}...\int\limits_{-\infty}^{\infty}M\left(\varphi_{<\mathbf{Z}_r,\mathbf{t}>}\left(1\right)\cdot\exp\left(-\frac{1}{4\lambda}\left\|\mathbf{t}\right\|^2\right)\right)dt_1...dt_d$$

Identificam urmatorii vectori independenti, identic repartizati

$$\mathbf{f}_k=\left(\frac{1}{\sqrt{p_1}}I_{A_1^{(k)}},...,\frac{1}{\sqrt{p_d}}I_{A_d^{(k)}}\right)',\quad k=1,...,r$$

cu

$$M\left(\mathbf{f}_k\right)=\left(\frac{p_1}{\sqrt{p_1}},...,\frac{p_d}{\sqrt{p_d}}\right)'=\left(\sqrt{p_1},...,\sqrt{p_d}\right)',\quad k=1,...,r$$

$$<\mathbf{Z}_r,\mathbf{t}>=\frac{1}{\sqrt{r}}\left(<\mathbf{f}_1,\mathbf{t}>+...+<\mathbf{f}_r,\mathbf{t}>-rM\left(<\mathbf{f},\mathbf{t}>\right)\right)$$

Dar

$$M\left(<\mathbf{f},\mathbf{t}>\right)=<M\left(\mathbf{f}\right),\mathbf{t}>=\sum_{j=1}^{d}t_j\sqrt{p_j}$$

$$M\left(<\mathbf{f},\mathbf{t}>\right)^2=M\left(\sum_{j=1}^{d}\frac{t_j}{\sqrt{p_j}}I_{A_j^{(k)}}\right)^2=M\left(\sum_{j=1}^{d}\frac{t_j^2}{p_j}I_{A_j^{(k)}}\right)=\sum_{j=1}^{d}t_j^2$$

$$D^2\left(<\mathbf{f},\mathbf{t}>\right)=\sum_{j=1}^{d}t_j^2-\left(\sum_{j=1}^{d}t_j\sqrt{p_j}\right)^2$$

Consideram $\{u_1, ..., u_d\}$ o baza ortonormala a lui R^d, cu $u_1 = \left(\sqrt{p_1}, ..., \sqrt{p_d}\right)'$.

$$D^2\left(<\mathbf{f},\mathbf{t}>\right) = \|\mathbf{t}\|^2 - <\mathbf{t},\mathbf{u}_1>^2 = \sum_{j=2}^{d} <\mathbf{t},\mathbf{u}_j>^2$$

Pentru sirul de variabile aleatoare independente, identic repartizate

$$\{<\mathbf{Z}_r,\mathbf{t}>, \quad r = 1,2,...\},$$

de medie 0,aplicam teorema limita centrala si teorema lui Paul Levy (pentru $t = 1$) :

$$\varphi_{<\mathbf{Z}_r,\mathbf{t}>}(1) \underset{r\to\infty}{\longrightarrow} \varphi_{N(0,D^2(<\mathbf{f},\mathbf{t}>))}(1) = \exp\left(-\frac{1}{2}\sum_{j=2}^{d} <\mathbf{t},\mathbf{u}_j>^2\right)$$

Rezulta

$$\psi_{X_r^2}(\lambda) \underset{r\to\infty}{\longrightarrow} \frac{1}{(4\pi\lambda)^{d/2}} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{j=2}^{d} <\mathbf{t},\mathbf{u}_j>^2\right) \cdot \exp\left(-\frac{1}{4\lambda}\|\mathbf{t}\|^2\right) dt_1...dt_d$$

Dar trecerea de la coordonatele $\{t_1, ..., t_d\}$ la coordonatele $\{v_1 =<\mathbf{t},\mathbf{u}_1>, ..., v_d =<\mathbf{t},\mathbf{u}_d>\}$ este ortogonala, deci de determinant 1.

$$\lim_{r\to\infty}\psi_{X_r^2}(\lambda) =$$

$$\frac{1}{(4\pi\lambda)^{d/2}} \int_{-\infty}^{\infty} ... \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2}\sum_{j=2}^{d} v_j^2\right) \cdot \exp\left(-\frac{1}{4\lambda}\sum_{j=1}^{d} v_j^2\right) dv_1...dv_d =$$

$$\frac{1}{(4\pi\lambda)^{d/2}} \left(\int_{-\infty}^{\infty} \exp\left(-\frac{v^2}{4\lambda}\right) dv\right) \left(\int_{-\infty}^{\infty} \exp\left(-v^2\left(\frac{1}{4\lambda}+\frac{1}{2}\right)\right) dv\right)^{d-1} =$$

$$\frac{1}{(4\pi\lambda)^{d/2}} \cdot \sqrt{\pi} \cdot \sqrt{4\lambda} \cdot (\pi)^{(d-1)/2} \left(\frac{1}{4\lambda}+\frac{1}{2}\right)^{-(d-1)/2} =$$

$$\frac{1}{(4\lambda)^{(d-1)/2}} \left(\frac{1}{4\lambda}+\frac{1}{2}\right)^{-(d-1)/2} = (1+2\lambda)^{-(d-1)/2}$$

Am demonstrat deci ca

$$\psi_{X_r^2}(\lambda) \underset{r\to\infty}{\longrightarrow} (1+2\lambda)^{-(d-1)/2}$$

si cum $(1+2\lambda)^{-(d-1)/2}$ este transformata Laplace corespunzatoare repartitiei $\chi^2(d-1)$, am obtinut c.t.d.
■

Testul Chi Patrat pentru concordanta dintre modelul stocastic si datele statistice

Fie datele statistice $(x_1,...,x_n)$. Din interpretarea lor, plus elementele de statistica descriptiva, alegem un posibil model stocastic din care ar proveni aceste date (ca valori ale unor observatii independente, identic repartizate).

- Notam $P\circ X^{-1}$ modelul ales si cu $S=X(\Omega)$ spatiul starilor.
- Partitionam $X(\Omega)$ in d submultimi masurabile $\{A_1,...,A_d\}$, $A_i\cap A_j=\Phi$ pentru $i\neq j$, $\bigcup_{i=1}^d A_i=X(\Omega)$.
- Calculam

$$p_j=P(X\in A_j),\ j=1,...,d,\ p_j\in[0,1]\ \forall j,\ \ \sum_{j=1}^d p_j=1$$

- Formulam ipoteza ca observatiile independente, identic repartizate $X_1,...,X_n$ care au produs datele statistice $(x_1,...,x_n)$ au repartitia $P\circ X^{-1}$

$$H:\{X_1,...,X_n \text{ sunt identic repartizate ca si } X\}$$

- Daca ipoteza H este adevarata, atunci functioneaza teorema lui Pearson.
- Calculam

$$n_j=card\{i\mid i=1,...,n,\ x_i\in A_j\}=\sum_{i=1}^n I_{A_j}(x_i),\ \ j=1,...,d$$

$$\sum_{j=1}^d n_j=n$$

- Calculam "distanta Pearson" dintre $(p_1, ..., p_d)$ si $\left(\frac{n_1}{n}, ..., \frac{n_d}{n}\right)$

$$S_n^2 = \sum_{j=1}^{d} \frac{n}{p_j}\left(\frac{n_j}{n} - p_j\right)^2 = \sum_{j=1}^{d} \frac{(n_j - np_j)^2}{np_j}$$

- Fie $\alpha \in (0,1)$ arbitrar fixat valoarea acceptata a probabilitatii de eroare (respingerea ipotezei H cand aceasta este adevarata).
- Fie $h_{d-1;1-\alpha}$ cuantila de rang $(1-\alpha)$ a repartitiei $\chi^2(d-1)$.
- REGULA DE DECIZIE: Daca $S_n^2 \geq h_{d-1;1-\alpha}$, decidem sa respingem ipoteza H

Comentarii:

- Testul se bazeaza pe teorema lui Pearson (este un test asimptotic), deci n trebuie sa fie mare $(n \geq 100)$
- Recomandari pentru alegerea valorii d:

$$\begin{aligned} d &\simeq 1 + 3.322 \cdot \log n \\ d &= \left[\frac{n}{3}\right] \end{aligned}$$

- Recomandari pentru alegerea elementelor partitiei:

$$A_j \text{ asa incat } p_j \simeq \frac{1}{d}, \quad j = 1, .., d$$

- Pentru implementarea in R

$$p - value = F_{\chi^2(d-1)}\left(S_n^2\right)$$

$$\text{Daca } p - value \leq 0.05, \text{ decidem sa respingem ipoteza } H$$