

# Exerciții de seminar 1

## Metoda verosimilității maxime și testare de ipoteze statistice

Obiectivul acestui seminar este de a prezenta câteva exerciții de calcul cu metode utile atunci când vrem să verificăm dacă eșantionul provine dintr-o populație normală.

## 1 Estimare prin metoda verosimilității maxime

### 1.1 Metoda verosimilității maxime și repartiția Geometrică



Fie  $X_1, X_2, \dots, X_n$  un eșantion de talie  $n$  dintr-o populație Geometrică a cărei funcție de masă este dată de

$$f_\theta(x) = \mathbb{P}_\theta(X = x) = \theta(1 - \theta)^{x-1}, \quad \forall x \in \{1, 2, 3, \dots\}$$

unde  $\theta \in (0, 1)$  este necunoscut. Presupunem că

$$\mathbb{E}[X] = \frac{1}{\theta}, \quad \text{Var}(X) = \frac{1 - \theta}{\theta^2}$$

- Scrieți logaritmul funcției de verosimilitate pentru eșantionul dat.
- Determinați estimatorul de verosimilitate maximă  $\hat{\theta}_n$  pentru  $\theta$ .
- Arătați că estimatorul de verosimilitate maximă este consistent.
- Folosind proprietățile asimptotice ale estimatorilor de verosimilitate maximă, derivați repartiția asimptotică a lui  $\hat{\theta}_n$ .
- Folosind *Teorema Limită Centrală* și *metoda Delta*, derivați repartiția asimptotică a lui  $\hat{\theta}_n$ .
- Determinați marginea Rao-Cramer.
- Generați un eșantion de talie  $n = 1000$  dintr-o populație Geometrică de parametru  $\theta = 0.345$ . Estimați parametru  $\theta$  prin metoda verosimilității maxime folosind funcția `optim()` (sau `optimize()`).

- a) Din definiția funcției de verosimilitate avem

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f_\theta(x_i) = \prod_{i=1}^n \theta(1 - \theta)^{x_i-1} = \theta^n (1 - \theta)^{\sum_{i=1}^n x_i - n}$$

de unde logaritmul funcției de verosimilitate este

$$l(\theta|\mathbf{x}) = \sum_{i=1}^n \log f_\theta(x_i) = n \log \theta + \left( \sum_{i=1}^n x_i - n \right) \log(1 - \theta).$$

- b) Estimatorul de verosimilitate maximă pentru  $\theta$  este definit prin

$$\hat{\theta}_n = \arg \max_{0 < \theta < 1} L(\theta|\mathbf{x}) = \arg \max_{0 < \theta < 1} l(\theta|\mathbf{x})$$

iar pentru determinarea acestuia (sub anumite condiții de regularitate) trebuie să rezolvăm ecuația de verosimilitate  $\frac{\partial l(\theta|\mathbf{x})}{\partial \theta} = 0$  (condiție de ordin unu). Trebuie remarcat că în cazul în care  $\theta = (\theta_1, \theta_2, \dots, \theta_k)$  condiția se scrie sub forma

$$\nabla L(\theta|\mathbf{x}) = \frac{\partial L(\theta|\mathbf{x})}{\partial \theta} = \begin{pmatrix} \frac{\partial L(\theta|\mathbf{x})}{\partial \theta_1} \\ \dots \\ \frac{\partial L(\theta|\mathbf{x})}{\partial \theta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ \dots \\ 0 \end{pmatrix}.$$

Soluțiile acestei ecuații ne dau punctele critice (din interiorul domeniului) iar pentru determinarea maximului este necesară verificarea unor condiții de ordin doi: matricea Hessiană

$$\frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta \partial \theta^\top} = \begin{pmatrix} \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_1^2} & \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_1 \partial \theta_k} \\ \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_2^2} & \dots & \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_2 \partial \theta_k} \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_k \partial \theta_1} & \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_k \partial \theta_2} & \dots & \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta_k^2} \end{pmatrix}$$

evaluată în  $\hat{\theta}_n$  trebuie să fie negativ definită, adică

$$\mathbf{x}^\top \frac{\partial^2 L(\theta|\mathbf{x})}{\partial \theta \partial \theta^\top} \mathbf{x} < 0, \quad \forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\}.$$

În cazul problemei noastre obținem

$$\frac{\partial l(\theta|\mathbf{x})}{\partial \theta} = \frac{n}{\theta} - \left( \sum_{i=1}^n x_i - n \right) \frac{1}{1-\theta}$$

și rezolvând ecuația  $\frac{\partial l(\theta|\mathbf{x})}{\partial \theta} = 0$  găsim că

$$\frac{n}{\theta} - \left( \sum_{i=1}^n x_i - n \right) \frac{1}{1-\theta} \iff \frac{1-\theta}{\theta} = \frac{1}{n} \sum_{i=1}^n x_i - 1 \iff \frac{1}{\theta} = \frac{1}{n} \sum_{i=1}^n x_i$$

de unde  $\hat{\theta}_n = \frac{1}{\bar{x}_n}$ . Pentru a vedea că  $\hat{\theta}_n$  este într-adevăr valoarea care maximizează funcția de verosimilitate, avem

$$\left. \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \theta^2} \right|_{\hat{\theta}_n} = -\frac{n}{\hat{\theta}_n^2} - \left( \frac{1}{1-\hat{\theta}_n} \right)^2 \left( \sum_{i=1}^n x_i - n \right)$$

și cum  $\hat{\theta}_n = \frac{1}{\bar{x}_n}$  deducem că  $\sum_{i=1}^n x_i - n = n \left( \frac{1}{\hat{\theta}_n} - 1 \right)$  iar

$$\begin{aligned} \left. \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \theta^2} \right|_{\hat{\theta}_n} &= -\frac{n}{\hat{\theta}_n^2} - \left( \frac{1}{1-\hat{\theta}_n} \right)^2 n \left( \frac{1}{\hat{\theta}_n} - 1 \right) = -n \left( \frac{1}{\hat{\theta}_n^2} + \frac{1}{\hat{\theta}_n(1-\hat{\theta}_n)} \right) \\ &= -\frac{n}{\hat{\theta}_n^2(1-\hat{\theta}_n)} < 0 \end{aligned}$$

ceea ce arată că  $\hat{\theta}_n = \frac{1}{\bar{x}_n}$  este estimatorul de verosimilitate maximă.

c) Aplicând *Legea numerelor mari* (varianta slabă) avem că

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mathbb{E}[X_1] = \frac{1}{\theta}$$

Cum  $\hat{\theta}_n = \frac{1}{\bar{X}_n}$  putem aplica *Teorema aplicațiilor continue* pentru funcția  $g(x) = \frac{1}{x}$ ,  $0 < x < 1$  și găsim că

$$\hat{\theta}_n = g(\bar{X}_n) \xrightarrow{\mathbb{P}} g\left(\frac{1}{\theta}\right) = \theta$$

ceea ce arată că  $\hat{\theta}_n$  este consistent.

d) Observăm că funcția de masă verifică condițiile de regularitate<sup>1</sup> prin urmare are loc

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, I_1^{-1}(\theta_0))$$

unde  $\theta_0$  este valoarea adevărată a parametrului iar  $I_1^{-1}(\theta_0)$  este informația lui Fisher pentru o observație. În general *Informația lui Fisher* pentru eșantion este

$$\begin{aligned} I_n(\theta) &= \text{Var}_{\theta}(\nabla l(\theta|\mathbf{X})) = \text{Var}_{\theta}\left(\frac{\partial \log f_{\theta}(\mathbf{X})}{\partial \theta}\right) \\ &= \mathbb{E}_{\theta}\left[\frac{\partial \log f_{\theta}(\mathbf{X})}{\partial \theta} \times \frac{\partial \log f_{\theta}(\mathbf{X})}{\partial \theta}^{\top}\right] \\ &= \mathbb{E}_{\theta}\left[-\frac{\partial^2 \log f_{\theta}(\mathbf{X})}{\partial \theta \partial \theta^{\top}}\right]. \end{aligned}$$

Pentru cazul nostru găsim că informația lui Fisher este

$$I_1(\theta) = \mathbb{E}_{\theta}\left[-\frac{\partial^2 \log f_{\theta}(X_i)}{\partial \theta^2}\right] = \mathbb{E}_{\theta}\left[\frac{1}{\theta^2} - \left(\frac{1}{1-\theta}\right)^2 (X_i - 1)\right] = \frac{1}{\theta^2(1-\theta)}$$

și astfel

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \theta_0^2(1-\theta_0))$$

sau echivalent  $\hat{\theta}_n \approx \mathcal{N}\left(\theta_0, \frac{\theta_0^2(1-\theta_0)}{n}\right)$ .

e) Știind că  $\mathbb{E}[X] = \frac{1}{\theta_0}$  și  $\text{Var}(X) = \frac{1-\theta_0}{\theta_0^2}$  și aplicând *Teorema Limită Centrală* avem că

$$\sqrt{n}\left(\bar{X}_n - \frac{1}{\theta_0}\right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}\left(0, \frac{1-\theta_0}{\theta_0^2}\right).$$

Estimatorul de verosimilitate maximă este  $\hat{\theta}_n = \frac{1}{\bar{X}_n}$  și considerând  $g(x) = \frac{1}{x}$ ,  $x \in (0, 1)$  ( $g$  este derivabilă cu derivata continuă) putem aplica metoda Delta care conduce la

$$\sqrt{n}\left(g(\bar{X}_n) - g\left(\frac{1}{\theta_0}\right)\right) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}\left(0, g'\left(\frac{1}{\theta_0}\right)^2 \frac{1-\theta_0}{\theta_0^2}\right)$$

<sup>1</sup>e.g. Suportul  $\{x \mid f_{\theta}(x) > 0\}$  nu depinde de  $\theta$ ;  $f_{\theta}(x)$  este de cel puțin 3 ori derivabilă în raport cu  $\theta$  și derivatele sunt continue; Valoarea adevărată  $\theta$  se află într-o mulțime compactă.

și cum  $g'(x) = -\frac{1}{x^2}$  obținem același rezultat ca și în cazul punctului anterior

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \theta_0^2(1 - \theta_0)).$$

f) Marginea inegalității Rao-Cramer (MIRC) este  $I_n^{-1}(\theta_0)$  și cum

$$I_n(\theta_0) = \mathbb{E}_\theta \left[ - \frac{\partial^2 \log f_\theta(\mathbf{X})}{\partial \theta \partial \theta^\top} \Big|_{\theta_0} \right] = n I_1(\theta_0) = \frac{n}{\theta_0^2(1 - \theta_0)}$$

găsim

$$MIRC = I_n^{-1}(\theta_0) = \frac{\theta_0^2(1 - \theta_0)}{n}.$$

g) Pentru a genera eșantionul  $X_1, X_2, \dots, X_n$  vom folosi funcția `rgeom()`. Atenție, această funcție permite generarea de observații repartizate Geometric de parametru  $\theta$ , cu funcția de masă

$$\mathbb{P}_\theta(X = x) = \theta(1 - \theta)^x, \quad \forall x \in \{0, 1, 2, 3, \dots\}$$

deci trebuie să adăugăm 1 la fiecare observație pentru a fi în contextul din exercițiu.

```
theta = 0.345
n = 1000

x = rgeom(n, theta) + 1

# EVM gasit este
EVM = 1/mean(x)
EVM
[1] 0.3489184
```

Vom crea o funcție care să calculeze estimatorul de verosimilitate maximă plecând de la logaritmul funcției de verosimilitate (îi determinăm maximum cu ajutorul funcției `optimize()`):

```
EVM_geom = function(theta, n, init = 0.5, seed = NULL){

  if (!is.null(seed)){
    set.seed(seed)
  }

  x = rgeom(n, theta)+1

  loglik_geom = function(param){

    l = n*log(param) + (sum(x) - n)*log(1-param)
    # intoarcem -l pentru ca vrem maximumul
    return(-l)
  }

  # folosim functia optimize
  # a se vedea ?optimize
  return(optimize(loglik_geom, c(0,1))[[1]])
}

# exemple
```

```
EVM_geom(0.345, 1000)
[1] 0.33886
EVM_geom(0.478, 1000)
[1] 0.4854295
EVM_geom(0.222, 1000)
[1] 0.2206884
```

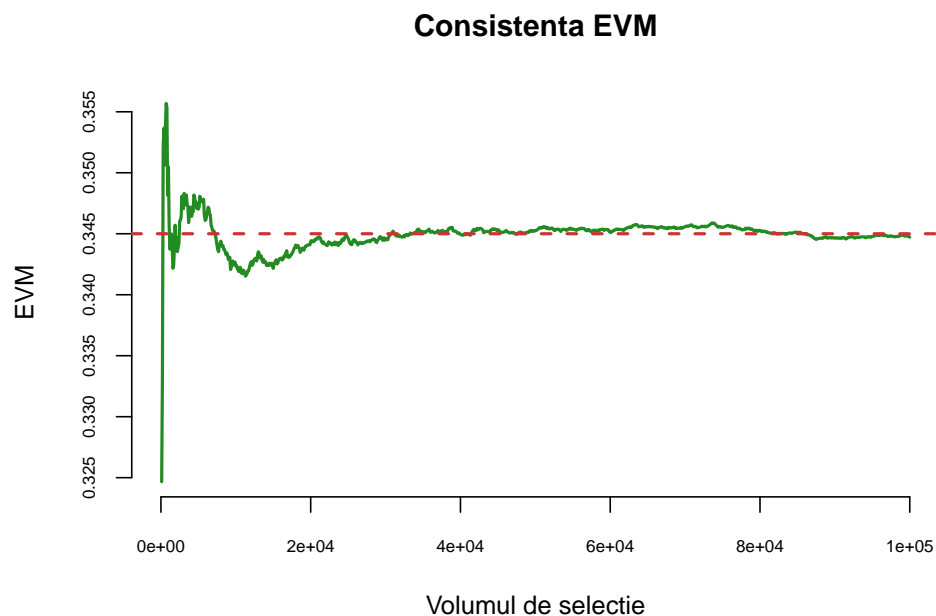
În figura de mai jos este ilustrată proprietatea de consistență a estimatorului de verosimilitate maximă, pentru  $\theta = 0.345$ :

```
theta = 0.345
t = seq(100, 100000, 100)

y = sapply(t, function(x){
  r = EVM_geom(theta, x, seed = 2018)
  return(r)
})

plot(t, y, type = "l",
     col = "forestgreen",
     xlab = "Volumul de selectie",
     ylab = "EVM",
     main = "Consistenta EVM",
     bty = "n",
     cex.axis = 0.7,
     lwd = 2)

abline(h = theta, col = "brown3",
       lty = 2, lwd = 2)
```



## 1.2 Exemplu de EVM determinat prin soluții numerice



Fie  $X_1, X_2, \dots, X_n$  un eșantion de talie  $n$  dintr-o populație logistică a cărei densitate este dată de formula

$$f_{\theta}(x) = \frac{e^{-(x-\theta)}}{(1 + e^{-(x-\theta)})^2}, \quad x \in \mathbb{R}, \theta \in \mathbb{R}$$

Determinați estimatorul de verosimilitate maximă  $\hat{\theta}_n$  pentru  $\theta$ .

Densitatea de repartiție și funcția de repartiție a repartiției logistice sunt ilustrate mai jos (în R se folosesc funcțiile: `rlogis`, `dlogis`, `plogis` și respectiv `qlogis`):

```
# Generam graficele
pars = c(2, 4, 6, 9)

x = seq(-8, 15, length.out = 250)

set.seed(1234)
cols = sample(colors(), length(pars))

par(mfrow = c(1, 2))
# densitatile
plot(x, dlogis(x, location = pars[1]),
     xlab = "x",
     ylab = TeX("$f_{\\theta}(x)$"),
     # ylim = c(0,1),
     col = "brown3",
     lwd = 2, type = "l",
     bty = "n",
     main = "Densitatea")

for (i in seq(length(pars)-1)){
  location = pars[i+1]

  y = dlogis(x, location = location)

  lines(x, y, lwd = 2,
        col = cols[i])
}

legend("topright",
      legend = TeX(paste0("$\\theta = ", pars, "$")),
      col = cols,
      lwd = rep(2, length(pars)),
      bty = "n",
      cex = 0.7,
      seg.len = 1.5)

# functiile de repartitie
plot(x, plogis(x, location = pars[1]),
     xlab = "x",
```

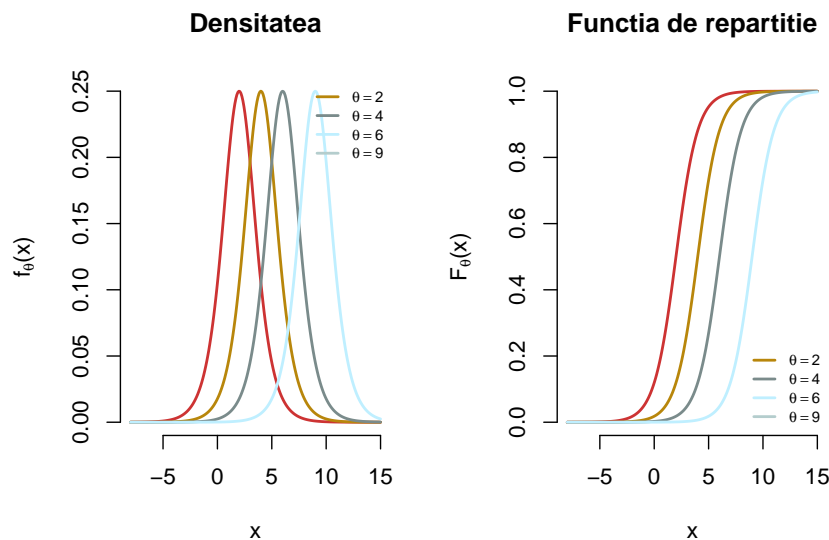
```
ylab = TeX("$F_{\\theta}(x)$"),
ylim = c(0,1),
col = "brown3",
lwd = 2, type = "l",
bty = "n",
main = "Funcția de repartiție")

for (i in seq(length(pars)-1)){
  location = pars[i+1]

  y = plogis(x, location = location)

  lines(x, y, lwd = 2,
        col = cols[i])
}

legend("bottomright",
      legend = TeX(paste0("$\\theta = ", pars, "$")),
      col = cols,
      lwd = rep(2, length(pars)),
      bty = "n",
      cex = 0.7,
      seg.len = 1.5)
```



Observăm că funcția de verosimilitate este dată de

$$L(\theta|\mathbf{x}) = \prod_{i=1}^n f_{\theta}(x_i) = \prod_{i=1}^n \frac{e^{-(x_i - \theta)}}{(1 + e^{-(x_i - \theta)})^2}$$

iar logaritmul funcției de verosimilitate este

$$l(\theta|\mathbf{x}) = \sum_{i=1}^n \log f_{\theta}(x_i) = n\theta - n\bar{x}_n - 2 \sum_{i=1}^n \log(1 + e^{-(x_i - \theta)}).$$

Pentru a găsi valoarea lui  $\theta$  care maximizează logaritmul funcției de verosimilitate și prin urmare a funcției de verosimilitate trebuie să rezolvăm ecuația  $l'(\theta|\mathbf{x}) = 0$ , unde derivata lui  $l(\theta|\mathbf{x})$  este

$$l'(\theta|\mathbf{x}) = n - 2 \sum_{i=1}^n \frac{e^{-(x_i-\theta)}}{1 + e^{-(x_i-\theta)}}$$

ceea ce conduce la ecuația

$$\sum_{i=1}^n \frac{e^{-(x_i-\theta)}}{1 + e^{-(x_i-\theta)}} = \frac{n}{2} \quad (\star)$$

Chiar dacă această ecuație nu se simplifică, se poate arăta că această ecuația admite soluție unică. Observăm că derivata parțială a membrului drept în  $(\star)$  devine

$$\frac{\partial}{\partial \theta} \sum_{i=1}^n \frac{e^{-(x_i-\theta)}}{1 + e^{-(x_i-\theta)}} = \sum_{i=1}^n \frac{e^{-(x_i-\theta)}}{(1 + e^{-(x_i-\theta)})^2} > 0$$

ceea ce arată că membrul stâng este o funcție strict crescătoare în  $\theta$ . Cum membrul stâng în  $(\star)$  tinde spre 0 atunci când  $\theta \rightarrow -\infty$  și spre  $n$  pentru  $\theta \rightarrow \infty$  deducem că ecuația  $(\star)$  admite soluție unică (vezi graficul de mai jos).

```
set.seed(112)
n = 20
x = rlogis(n, location = 7.5)

# derivata logaritmului functiei de verosimilitate
dLogLogistic = function(n, x, theta){
  sapply(theta, function(t){
    y = exp(-(x - t))
    n - 2*sum(y/(1+y))
  })
}

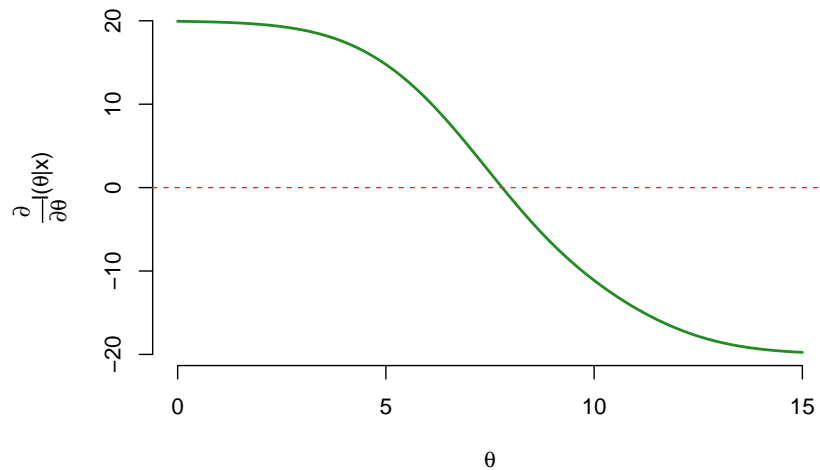
theta = seq(0, 15, length.out = 250)

mar.default <- c(5,4,4,2) + 0.1
par(mar = mar.default + c(0, 1.2, 0, 0))

plot(theta, dLogLogistic(n, x, theta), type = "l",
     col = "forestgreen", lwd = 2,
     bty = "n",
     xlab = TeX("$\\theta$"),
     ylab = TeX("$\\frac{\\partial}{\\partial \\theta} l(\\theta | x)$"))

abline(h = 0, col = "brown3",
       lty = 2)
```





Cum nu putem găsi o soluție a ecuației  $l'(\theta|\mathbf{x}) = 0$  sub formă compactă, este necesar să apelăm la metode numerice. O astfel de metodă numerică este binecunoscuta metodă a lui Newton-Raphson. Metoda presupune să începem cu o valoare (soluție) inițială  $\hat{\theta}^{(0)}$  și să alegem, plecând de la aceasta, o nouă valoare  $\hat{\theta}^{(1)}$  definită prin

$$\hat{\theta}^{(1)} = \hat{\theta}^{(0)} - \frac{l'(\hat{\theta}^{(0)})}{l''(\hat{\theta}^{(0)})},$$

adică  $\hat{\theta}^{(1)}$  este intersecția cu axa absciselor a tangentei în punctul  $(\hat{\theta}^{(0)}, l'(\hat{\theta}^{(0)}))$  la graficul funcției  $l'(\theta)$ . Ideea este de a itera procesul până când soluția converge, cu alte cuvinte pornind de la o valoare *rezonabilă* de start  $\hat{\theta}^{(0)}$  la pasul  $k + 1$  avem

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \frac{l'(\hat{\theta}^{(k)})}{l''(\hat{\theta}^{(k)})}$$

și oprim procesul atunco când  $k$  este suficient de mare și/sau  $|\hat{\theta}^{(k+1)} - \hat{\theta}^{(k)}|$  este suficient de mic. Următorul grafic ilustrează grafic algoritmul lui Newton:

```
set.seed(112)
n = 20
x = rlogis(n, location = 7.5)

# derivata logaritmului functiei de verosimilitate
dLogLogistic = function(n, x, theta){
  sapply(theta, function(t){
    y = exp(-(x - t))
    n - 2*sum(y/(1+y))
  })
}

theta = seq(0, 15, length.out = 250)
```

```
mar.default <- c(5,4,4,2) + 0.1
par(mar = mar.default + c(0, 1.2, 0, 0))

plot(theta, dLogLogistic(n, x, theta), type = "l",
     col = "forestgreen", lwd = 2,
     bty = "n",
     xlab = TeX("$\\theta$"),
     ylab = TeX("$\\frac{\\partial}{\\partial \\theta} l(\\theta | x)$"))

abline(h = 0, col = "brown3",
       lty = 2)

# ilustrarea metodei Newton

dl = function(theta) n - 2 * sum(exp(theta - x) / (1 + exp(theta - x)))
ddl = function(theta) {-2 * sum(exp(theta - x) / (1 + exp(theta - x))^2)}

x0 = 5 # punctul de start

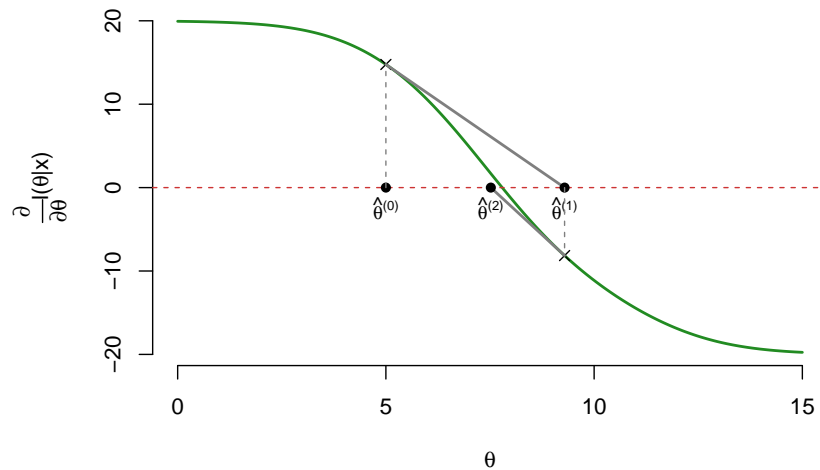
points(x0, 0, pch = 16, col = "black")
text(x0, 0, labels = TeX("$\\hat{\\theta}^{(0)}$"), pos = 1, cex = 0.8)
segments(x0, 0, x0, dl(x0), lty = 2, col = "grey50")
points(x0, dl(x0), pch = 4)

x1 = x0 - dl(x0)/ddl(x0)

segments(x0, dl(x0), x1, 0, lty = 1, lwd = 2, col = "grey50")
points(x1, 0, pch = 16, col = "black")
text(x1, 0, labels = TeX("$\\hat{\\theta}^{(1)}$"), pos = 1, cex = 0.8)
segments(x1, 0, x1, dl(x1), lty = 2, col = "grey50")
points(x1, dl(x1), pch = 4)

x2 = x1 - dl(x1)/ddl(x1)

segments(x1, dl(x1), x2, 0, lty = 1, lwd = 2, col = "grey50")
points(x2, 0, pch = 16, col = "black")
text(x2, 0, labels = TeX("$\\hat{\\theta}^{(2)}$"), pos = 1, cex = 0.8)
```



**Obs:** Singurul lucru care se schimbă atunci când trecem de la scalar la vector, este funcția  $l(\theta)$  care acum este o funcție de  $p > 1$  variabile,  $\theta = (\theta_1, \theta_2, \dots, \theta_p)^T \in \mathbb{R}^p$ . În acest context  $l'(\theta)$  este un vector de derivate parțiale iar  $l''(\theta)$  este o matrice de derivate parțiale de ordin doi. Prin urmare iterațiile din metoda lui Newton sunt

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - [l''(\hat{\theta}^{(k)})]^{-1} l'(\hat{\theta}^{(k)})$$

unde  $[\cdot]^{-1}$  este pseudoinversa unei matrici.

Funcția de mai jos implementează metoda lui Newton pentru cazul multidimensional:

```
# Metoda lui Newton

newton <- function(f, df, x0, eps=1e-08, maxiter=1000, ...) {
  # in caz ca nu e incarcat pachetul sa putem accesa pseudoinversa
  if(!exists("ginv")) library(MASS)

  x <- x0
  k <- 0

  repeat {
    k <- k + 1

    x.new <- x - as.numeric(ginv(df(x, ...)) %*% f(x, ...))

    if(mean(abs(x.new - x)) < eps | k >= maxiter) {
      if(k >= maxiter) warning("S-a atins numarul maxim de iteratii!")
      break
    }
    x <- x.new
  }
  out <- list(solution = x.new, value = f(x.new, ...), iter = k)

  return(out)
}
```

Să presupunem că am observat următorul eșantion de talie 20 din repartiția logistică:

```
[1] 6.996304 9.970107 12.304991 11.259549 6.326912 5.378941 4.299639
[8] 8.484635 5.601117 7.094335 6.324731 6.868456 9.753360 8.042095
[15] 8.227830 10.977982 7.743096 7.722159 8.562884 6.968356
```

```
set.seed(112)
x = rlogis(20, location = 7.5)

n = length(x)
dl = function(theta) n - 2 * sum(exp(theta - x) / (1 + exp(theta - x)))
ddl = function(theta) {-2 * sum(exp(theta - x) / (1 + exp(theta - x))^2)}

logis.newton = newton(dl, ddl, median(x))
```

și aplicând metoda lui Newton găsim estimatorul de verosimilitate maximă  $\hat{\theta}_n = 7.7933$  după numai 3 iterații (datele au fost simulate folosind  $\theta = 7.5$ ).

### 1.3 Metoda verosimilității maxime și procese autoregresive $AR(r)$



Se numește proces autoregresiv de ordin 1  $AR(1)$ , un proces Gaussian staționar definit prin

$$Y_t = c + \rho Y_{t-1} + \epsilon_t$$

cu  $\epsilon_t$  variabile aleatoare i.i.d. repartizate  $\mathcal{N}(0, \sigma^2)$  și  $|\rho| < 1$ .

Observăm că din condiția de staționaritate<sup>2</sup> rezultă că

$$\mathbb{E}[Y_t] = \frac{c}{1 - \rho}, \quad \text{Var}[Y_t] = \frac{\sigma^2}{1 - \rho^2}.$$



Fie  $\theta = (c, \rho, \sigma^2)^\top$  vectorul parametrilor modelului. Scrieți funcția de verosimilitate și logaritmul funcției de verosimilitate pentru o observație,  $y_1$ .

Cum variabila aleatoare  $Y_1$  are media și varianța date de

$$\mathbb{E}[Y_1] = \frac{c}{1 - \rho}, \quad \text{Var}[Y_1] = \frac{\sigma^2}{1 - \rho^2}.$$

iar  $\epsilon_t$  sunt i.i.d. repartizate  $\mathcal{N}(0, \sigma^2)$ , deducem că  $Y_1$  este repartizată tot normal, cu  $Y_1 \sim \mathcal{N}\left(\frac{c}{1 - \rho}, \frac{\sigma^2}{1 - \rho^2}\right)$ . Astfel funcția de verosimilitate pentru  $y_1$  este

$$L(\theta; y_1) = \frac{1}{\sqrt{2\pi} \sqrt{\frac{\sigma^2}{1 - \rho^2}}} e^{-\frac{1}{2} \frac{\left(y_1 - \frac{c}{1 - \rho}\right)^2}{\frac{\sigma^2}{1 - \rho^2}}}$$

iar logaritmul funcției de verosimilitate pentru  $y_1$  este

<sup>2</sup>Aici ne referim la proprietatea de staționaritate în sens larg (wide-sense stationary) care presupune că  $\forall t_1, t_2 \in \mathbb{N}$  și  $\forall \tau \in \mathbb{N}$  avem  $\mathbb{E}[Y_{t_1}] = \mathbb{E}[Y_{t_2}]$  și  $\mathbb{E}[Y_{t_1} Y_{t_2}] = \mathbb{E}[Y_{t_1 + \tau} Y_{t_2 + \tau}]$ .

$$l(\theta; y_1) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log \left( \frac{\sigma^2}{1-\rho^2} \right) - \frac{1}{2} \frac{\left( y_1 - \frac{c}{1-\rho} \right)^2}{\frac{\sigma^2}{1-\rho^2}}.$$



Care este repartiția condiționată a lui  $Y_2$  la  $Y_1 = y_1$ ? Scrieți funcția de verosimilitate și logaritmul funcției de verosimilitate (condiționată) pentru a doua observație  $y_2$ .

Observăm că pentru  $t = 2$  avem

$$Y_2 = c + \rho Y_1 + \epsilon_2,$$

unde  $\epsilon_2 \sim \mathcal{N}(0, \sigma^2)$ . Prin urmare repartiția condiționată a lui  $Y_2$  dat fiind  $Y_1 = y_1$  este

$$Y_2|Y_1 = y_1 \sim \mathcal{N}(c + \rho y_1, \sigma^2)$$

de unde funcția de verosimilitate (condiționată) pentru  $y_2$  este

$$L(\theta; y_2|y_1) = f_{Y_2|Y_1}(y_2|y_1; \theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(y_2 - c - \rho y_1)^2}{\sigma^2}}$$

iar logaritmul funcției de verosimilitate (condiționată) pentru  $y_2$  este

$$l(\theta; y_2|y_1) = \log f_{Y_2|Y_1}(y_2|y_1; \theta) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2} \frac{(y_2 - c - \rho y_1)^2}{\sigma^2}.$$



Considerați eșantionul  $\{y_1, y_2\}$  de talie 2. Scrieți funcția de verosimilitate (completă) și logaritmul funcției de verosimilitate a modelului  $AR(1)$  pentru acest eșantion. Extindeți rezultatul pentru un eșantion  $y_1, y_2, \dots, y_T$  de talie  $T$ .

Reamintim că dacă avem două variabile aleatoare continue (absolut continue)  $X$  și  $Y$  atunci densitatea cuplului  $(X, Y)$  este

$$f_{(X,Y)}(x, y) = f_{Y|X}(y|x) f_X(x),$$

prin urmare funcția de verosimilitate (completă) pentru eșantionul  $\{y_1, y_2\}$  este

$$L(\theta; y_1, y_2) = f_{(Y_1, Y_2)}(y_1, y_2; \theta) = f_{Y_2|Y_1}(y_2|y_1; \theta) f_{Y_1}(y_1; \theta)$$

sau echivalent

$$L(\theta; y_1, y_2) = L(\theta; y_2|y_1) L(\theta; y_1) = \frac{\sqrt{1-\rho^2}}{2\pi\sigma^2} e^{-\frac{1}{2} \frac{(1-\rho^2)(y_1 - \frac{c}{1-\rho})^2}{\sigma^2} - \frac{1}{2} \frac{(y_2 - c - \rho y_1)^2}{\sigma^2}}.$$

În mod similar, logaritmul funcției de verosimilitate este

$$l(\theta; y_1, y_2) = l(\theta; y_2|y_1) + l(\theta; y_1) = \frac{1}{2} \log(1-\rho^2) - \log(2\pi\sigma^2) - \frac{1}{2} \frac{(1-\rho^2)(y_1 - \frac{c}{1-\rho})^2}{\sigma^2} - \frac{1}{2} \frac{(y_2 - c - \rho y_1)^2}{\sigma^2}.$$

Observăm că densitatea lui  $Y_3$  condiționată la primele două variabile este

$$f_{Y_3|Y_2,Y_1}(y_3|y_2,y_1;\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(y_3 - c - \rho y_2)^2}{\sigma^2}}$$

de unde

$$\begin{aligned} f_{Y_3,Y_2,Y_1}(y_3,y_2,y_1;\theta) &= f_{Y_3|Y_2,Y_1}(y_3|y_2,y_1;\theta) f_{Y_2,Y_1}(y_2,y_1;\theta) \\ &= f_{Y_3|Y_2,Y_1}(y_3|y_2,y_1;\theta) f_{Y_2|Y_1}(y_2|y_1;\theta) f_{Y_1}(y_1;\theta). \end{aligned}$$

În general, valoarea lui  $Y_1, Y_2, \dots, Y_{t-1}$  influențează valoarea lui  $Y_t$  doar prin valoarea lui  $Y_{t-1}$  ceea ce arată că densitatea lui  $Y_t$  condiționată la celelalte  $t-1$  variabile este

$$f_{Y_t|Y_{t-1},Y_{t-2},\dots,Y_1}(y_t|y_{t-1},y_{t-2},\dots,y_1;\theta) = f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\theta) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(y_t - c - \rho y_{t-1})^2}{\sigma^2}}.$$

Astfel, pentru un eșantion  $y_1, y_2, \dots, y_T$  de talie  $T$  avem

$$\begin{aligned} L(\theta; y_1, y_2, \dots, y_T) &= L(\theta; y_1) \times \prod_{t=2}^T L(\theta; y_t|y_{t-1}) \\ l(\theta; y_1, y_2, \dots, y_T) &= l(\theta; y_1) + \sum_{t=2}^T l(\theta; y_t|y_{t-1}) \end{aligned}$$

ceea ce conduce la

$$L(\theta; y_1, y_2, \dots, y_T) = \frac{1}{\sqrt{2\pi} \sqrt{\frac{\sigma^2}{1-\rho^2}}} e^{-\frac{1}{2} \frac{(y_1 - \frac{c}{1-\rho})^2}{\frac{\sigma^2}{1-\rho^2}}} \times \prod_{t=2}^T \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2} \frac{(y_t - c - \rho y_{t-1})^2}{\sigma^2}}$$

și respectiv la

$$\begin{aligned} l(\theta; y_1, y_2, \dots, y_T) &= -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log\left(\frac{\sigma^2}{1-\rho^2}\right) - \frac{1}{2} \frac{\left(y_1 - \frac{c}{1-\rho}\right)^2}{\frac{\sigma^2}{1-\rho^2}} \\ &\quad + \sum_{t=2}^T \left( -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - \frac{1}{2} \frac{(y_t - c - \rho y_{t-1})^2}{\sigma^2} \right) \\ &= -\frac{T}{2} \log(2\pi) - \frac{T}{2} \log(\sigma^2) + \frac{1}{2} \log(1-\rho^2) + \\ &\quad + \frac{1}{2\sigma^2} \left[ (1-\rho^2) \left(y_1 - \frac{c}{1-\rho}\right)^2 + \sum_{t=2}^T (y_t - c - \rho y_{t-1})^2 \right] \end{aligned}$$

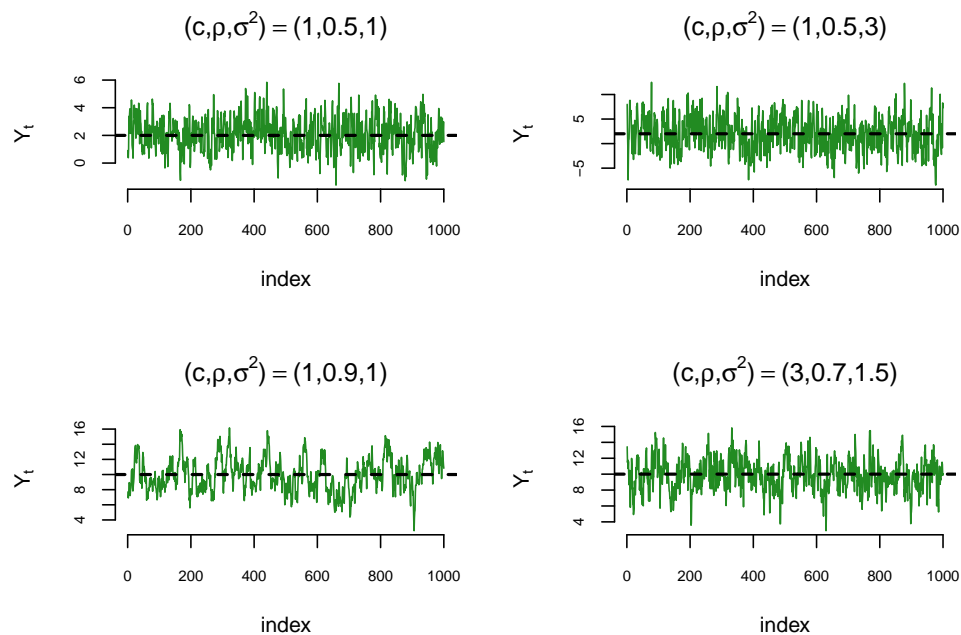


Funcția de verosimilitate este o funcție neliniară în parametrii  $\theta$ , prin urmare estimatorul de verosimilitate maximă  $\hat{\theta} = (\hat{c}, \hat{\rho}, \hat{\sigma}^2)^\top$  va fi determinat prin metode numerice. Scrieți o funcție în R care să permită generarea unui eșantion dintr-un proces  $AR(1)$ . Pentru  $c = 1$ ,  $\rho = 0.5$  și  $\sigma^2 = 1$  generați un eșantion de talie  $T = 1000$  și calculați estimatorul de verosimilitate maximă.

Avem următoarea funcție care generează procesul autoregresiv  $AR(1)$ :

```
genAR1 = function(n, c, rho, sigma){  
  # n - marimea esantionului  
  # c - termenul constant  
  # rho - parametrul autoregresiv  
  # sigma - abaterea standard a erorii  
  
  # generam Y_1 repartizat normal  
  y1 = rnorm(1, mean = c/(1-rho), sd = sqrt(sigma^2/(1-rho^2)))  
  
  # nitializam  
  y = rep(1, n)*y1  
  
  # vectorul de erori  
  epsilon = rnorm(n-1, 0, sigma)  
  
  for (i in 2:n){  
    y[i] = c + rho*y[i-1] + epsilon[i-1]  
  }  
  
  return(y)  
}
```

Ilustrăm grafic traiectoriile procesului  $AR(1)$  pentru diverse seturi de parametrii  $(c, \rho, \sigma^2)$ :



Considerăm setul de parametrii  $(c, \rho, \sigma^2) = (1, 0.5, 1)$  și calculăm estimatorul de verosimilitate maximă plecând de la logaritmul funcției de verosimilitate (utilizăm funcția `optim()`):

```
y = genAR1(1000, 1, 0.5, 1)  
  
loglik_AR1 = function(param){  
  # pentru a folosi functia optim trebuie sa avem un singur argument
```

```
# parametrii
c = param[1]
rho = param[2]
sigma = param[3]

# esantionul
ly = length(y) # talia esantionului

# prima observatie
l1 = log(dnorm(y[1], mean = c/(1-rho), sd = sqrt(sigma^2/(1-rho^2))))

# celelalte observatii
dif = y[2:ly] - c - rho*y[1:(ly-1)]
l2 = log(dnorm(dif, 0, sigma))

# logarithmul verosimilitatii
l = l1 + sum(l2)

# intoarcem -l pentru ca vrem maximul
return(-l)
}

# determinam MLE
param = c(0.6, 0.6, 0.6)
MLE = optim(param, loglik_AR1)$par
```

Obținem următoarele rezultate

	Theta	MLE
c	1.0	0.9261046
rho	0.5	0.5211693
sigma	1.0	1.0104216

care sunt apropiate de valorile reale.



Acum considerăm că prima observație  $y_1$  este dată (deterministă) și avem  $f_{Y_1}(y_1; \theta)$ . Scrieți logarithmul funcției de verosimilitate a modelului  $AR(1)$  pentru eșantionul  $y_1, y_2, \dots, y_T$ .

Funcția de verosimilitate condiționată este definită prin

$$L(\theta; y_2, \dots, y_T | y_1) = \prod_{t=2}^T f_{Y_t | Y_{t-1}, Y_1=y_1}(y_t | y_{t-1}, y_1; \theta) \times \underbrace{f_{Y_1}(y_1; \theta)}_{=1} = \prod_{t=2}^T f_{Y_t | Y_{t-1}}(y_t | y_{t-1}; \theta)$$

iar logarithmul funcției de verosimilitate condiționată devine

$$l(\theta; y_2, \dots, y_T | y_1) = \sum_{t=2}^T l_t(\theta; y_t | y_{t-1})$$

cu  $l_t(\theta; y_t | y_{t-1}) = \log(f_{Y_t | Y_{t-1}}(y_t | y_{t-1}; \theta))$ . Găsim că



$$l(\theta; y_2, \dots, y_T | y_1) = -\frac{T-1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=2}^T (y_t - c - \rho y_{t-1})^2.$$

## 2 Testarea ipotezelor statistice

### 2.1 Teste parametrice și Lema Neyman-Pearson

Să presupunem că ne aflăm în contextul următoarei probleme:



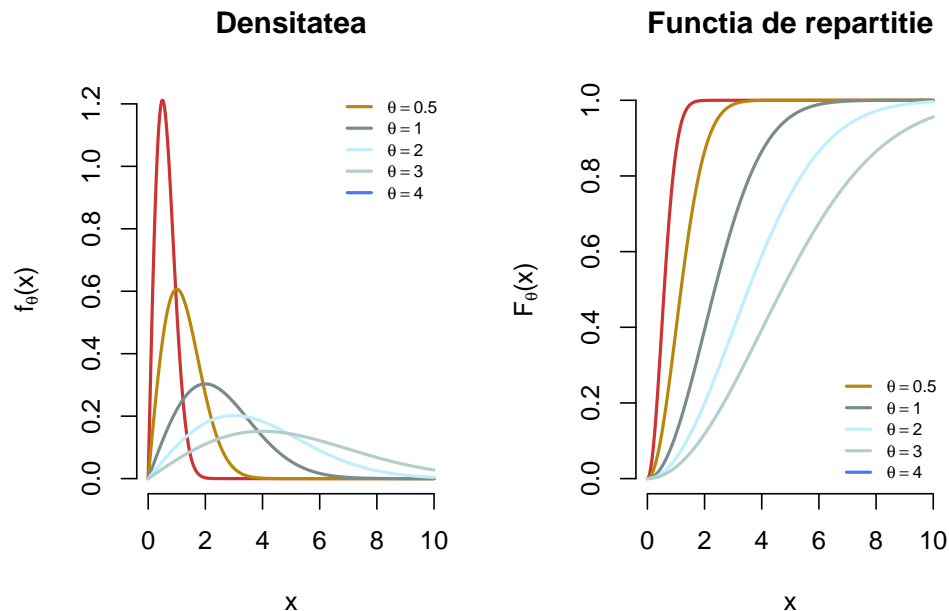
Fie  $U$  și  $V$  două variabile aleatoare independente și repartizate  $\mathcal{N}(0, \theta)$ . Variabila aleatoare  $X$  definită prin

$$X = \sqrt{U^2 + V^2}$$

este repartizată *Rayleigh* de parametru  $\theta$ ,  $X \sim \text{Rayleigh}(\theta)$ , și are densitatea

$$f_{\theta}(x) = \frac{x}{\theta} e^{-\frac{x^2}{2\theta}}, \quad \forall x \in [0, \infty]$$

Pentru mai multe detalii privind repartiția Rayleigh se poate consulta pagina [https://en.wikipedia.org/wiki/Rayleigh\\_distribution](https://en.wikipedia.org/wiki/Rayleigh_distribution) sau monografia [Merran Evans, 2000]. Densitatea de repartiție și funcția de repartiție a repartiției Rayleigh sunt ilustrate mai jos (pentru a folosi în R funcțiile: `rrayleigh`, `drayleigh`, `prayleigh` și respectiv `qrayleigh` trebuie instalat pachetul `VGAM`):



În cele ce urmează, ne propunem să răspundem la o serie de întrebări:



Fie  $X_1, X_2, \dots, X_n$  un eșantion de talie  $n$  dintr-o populație Rayleigh de parametru  $\theta$ . Determinați estimatorul de verosimilitate maximă pentru  $\theta$ .

Logaritmul funcției de verosimilitate este dat de

$$l(\theta|\mathbf{x}) = \sum_{i=1}^n \log f_{\theta}(x_i) = \sum_{i=1}^n \log(x_i) - n \log(\theta) + \frac{1}{2\theta} \sum_{i=1}^n x_i^2$$

iar estimatorul de verosimilitate verifică

$$\hat{\theta}_n = \arg \max_{\theta > 0} l(\theta|\mathbf{x}) = \arg \max_{\theta > 0} \sum_{i=1}^n \log(x_i) - n \log(\theta) + \frac{1}{2\theta} \sum_{i=1}^n x_i^2.$$

Rezolvând ecuația de verosimilitate (condiția de ordin 1)

$$\left. \frac{\partial l(\theta|\mathbf{x})}{\partial \theta} \right|_{\hat{\theta}_n} = -\frac{n}{\hat{\theta}_n} + \frac{1}{2\hat{\theta}_n^2} \sum_{i=1}^n x_i^2 = 0$$

găsim că

$$\hat{\theta}_n = \frac{1}{2n} \sum_{i=1}^n X_i^2.$$

Pentru a vedea că într-adevăr  $\hat{\theta}_n$  este estimatorul de verosimilitate maximă trebuie să verificăm condiția de ordin 2

$$\left. \frac{\partial^2 l(\theta|\mathbf{x})}{\partial \theta^2} \right|_{\hat{\theta}_n} = \frac{n}{\hat{\theta}_n^2} - \frac{1}{\hat{\theta}_n^3} \sum_{i=1}^n x_i^2 = \frac{n}{\hat{\theta}_n^2} - \frac{2n\hat{\theta}_n}{\hat{\theta}_n^3} = -\frac{n}{\hat{\theta}_n^2} < 0$$

unde am folosit faptul că  $\sum_{i=1}^n x_i^2 = 2n\hat{\theta}_n$ . Prin urmare  $\hat{\theta}_n$  este estimatorul de verosimilitate maximă.



Determinați repartiția asimptotică a EVM  $\hat{\theta}_n$  a lui  $\theta$ .

Știm că dacă  $\hat{\theta}_n$  este estimatorul de verosimilitate maximă pentru  $\theta$  și  $f_{\theta}(x)$  verifică o serie de condiții de regularitatea atunci are loc

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, I_1^{-1}(\theta_0))$$

unde  $\theta_0$  este valoarea adevărată a parametrului iar  $I_1^{-1}(\theta_0)$  este informația lui Fisher pentru o observație. În cazul problemei noastre, informația lui Fisher este

$$I_n(\theta) = \mathbb{E}_{\theta} \left[ -\frac{\partial^2 \log f_{\theta}(\mathbf{X})}{\partial \theta^2} \right] = \mathbb{E}_{\theta} \left[ -\frac{n}{\theta^2} + \frac{1}{\theta^3} \sum_{i=1}^n X_i^2 \right] = -\frac{n}{\theta^2} + \frac{1}{\theta^2} \sum_{i=1}^n \mathbb{E}_{\theta} \left[ \frac{X_i^2}{\theta} \right]$$

Cum  $\frac{X^2}{\theta} = \frac{U^2}{\theta} + \frac{V^2}{\theta}$  iar  $\frac{U}{\sqrt{\theta}}$  și  $\frac{V}{\sqrt{\theta}}$  sunt variabile aleatoare independente repartizate  $\mathcal{N}(0, 1)$  deducem că  $\frac{X^2}{\theta}$  este repartizată  $\chi^2(2)$  prin urmare

$$\mathbb{E}_\theta \left[ \frac{X_i^2}{\theta} \right] = 2.$$

Astfel  $I_n(\theta) = -\frac{n}{\theta^2} + \frac{2n}{\theta^2} = \frac{n}{\theta^2}$  de unde  $I_1(\theta) = \frac{1}{\theta^2}$ . Avem

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, I_1^{-1}(\theta))$$

sau echivalent

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow \infty]{d} \mathcal{N}(0, \theta^2).$$



Considerăm testul pentru ipotezele

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

unde  $\theta_1 > \theta_0$ . Determinați regiunea critică pentru UMP test de mărime  $\alpha$  pentru ipotezele  $H_0$  și  $H_1$ .

Din lema Neyman-Pearson avem că regiunea critică a testului UMP este dată de

$$C = \left\{ (x_1, x_2, \dots, x_n) \mid \frac{L_{\theta_0}(x_1, x_2, \dots, x_n)}{L_{\theta_1}(x_1, x_2, \dots, x_n)} < k \right\}$$

unde constanta  $k$  se determină din mărimea testului  $\alpha$

$$\mathbb{P}_{H_0}((x_1, x_2, \dots, x_n) \in C) = \alpha.$$

Prin logaritmare avem

$$\begin{aligned} l(\theta_0|\mathbf{x}) - l(\theta_1|\mathbf{x}) &< \log(k) \\ \iff n(\log(\theta_1) - \log(\theta_0)) + \frac{1}{2} \left( \frac{1}{\theta_1} - \frac{1}{\theta_0} \right) \sum_{i=1}^n x_i^2 &< \log(k) \\ \iff \frac{1}{2} \left( \frac{1}{\theta_1} - \frac{1}{\theta_0} \right) \sum_{i=1}^n x_i^2 &< k_1 = \log(k) - n(\log(\theta_1) - \log(\theta_0)) \end{aligned}$$

sau echivalent, ținând cont de faptul că  $\theta_1 > \theta_0$ , avem

$$\frac{1}{2n} \sum_{i=1}^n x_i^2 > c$$

unde  $c = \frac{k_1 \theta_0 \theta_1}{n(\theta_0 - \theta_1)}$ .

Regiunea critică pentru testul UMP de mărime  $\alpha$  cu ipotezele

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

unde  $\theta_1 > \theta_0$  este

$$C = \left\{ (x_1, x_2, \dots, x_n) \mid \hat{\theta}_n = \frac{1}{2n} \sum_{i=1}^n x_i^2 > c \right\}.$$

Constanta  $c$  se determină din condiția

$$\alpha = \mathbb{P}_{H_0}(C) = \mathbb{P}_{H_0}(\hat{\theta}_n > c).$$

Sub ipoteza nulă, dacă  $n$  este suficient de mare, am văzut că

$$\hat{\theta}_n \underset{H_0}{\approx} \mathcal{N}\left(\theta_0, \frac{\theta_0^2}{n}\right)$$

prin urmare

$$1 - \alpha = \mathbb{P}_{H_0}\left(\sqrt{n} \frac{\hat{\theta}_n - \theta_0}{\theta_0} < \sqrt{n} \frac{c - \theta_0}{\theta_0}\right)$$

de unde  $c = \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\alpha}$  cu  $z_{1-\alpha} = \Phi^{-1}(1 - \alpha)$ .

Regiunea critică a testului UMP cu ipotezele  $H_0$  vs  $H_1$  devine

$$C = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) > \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\alpha} \right\}.$$



Considerăm testul pentru ipotezele

$$H_0 : \theta = 2 \quad \text{vs} \quad H_1 : \theta > 2$$

Știind că pentru un eșantion de talie  $n = 100$  avem  $\sum_{i=1}^n x_i^2 = 470$  care este concluzia testului pentru un prag de semnificație de 10%?

Considerăm testul cu ipotezele

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1$$

unde  $\theta_1 > \theta_0$ . Am văzut că regiunea critică a testului UMP de mărime  $\alpha$  este

$$C = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) > \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\alpha} \right\}.$$

Cum regiunea critică nu depinde de  $\theta_1$  (în plus raportul de verosimilitate verifică proprietatea de monotonicitate), ea corespunde și la testul unilateral UMP de mărime  $\alpha$ :

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

Pentru  $\theta_0 = 2$ ,  $n = 100$  și  $\alpha = 0.1$  obținem

$$C = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) > 2 + \frac{2}{10} z_{0.9} \right\} = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) > 2.2563 \right\}$$

Cum, din ipoteză avem că  $\sum_{i=1}^n x_i^2 = 470$ , pentru  $n = 100$  deducem că

$$\hat{\theta}_n(\mathbf{x}) = \frac{1}{2n} \sum_{i=1}^n x_i^2 = \frac{470}{200} = 2.35$$

ceea ce arată că pentru pragul de semnificație de  $\alpha = 10\%$  respingem ipoteza nulă  $H_0 : \theta = 2$ .



Determinați puterea testului unilateral UMP de mărime  $\alpha$  pentru ipotezele

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

Ilustrați grafic în R pentru  $n = 100$ ,  $\theta_0 = 2$  și  $\alpha = 0.1$ .

Am văzut că regiunea critică este dată de

$$C = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) > \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\alpha} \right\}$$

iar din definiția funcției putere avem că

$$\text{pow}(\theta) = \mathbb{P}_{H_1}(\mathbf{x} \in C) = \mathbb{P}_{H_1}(\hat{\theta}_n(\mathbf{x}) > a)$$

cu  $a = \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\alpha}$ .

Sub ipoteza alternativă avem că

$$\hat{\theta}_n \underset{H_1}{\approx} \mathcal{N}\left(\theta, \frac{\theta^2}{n}\right), \quad \theta > \theta_0$$

prin urmare puterea este

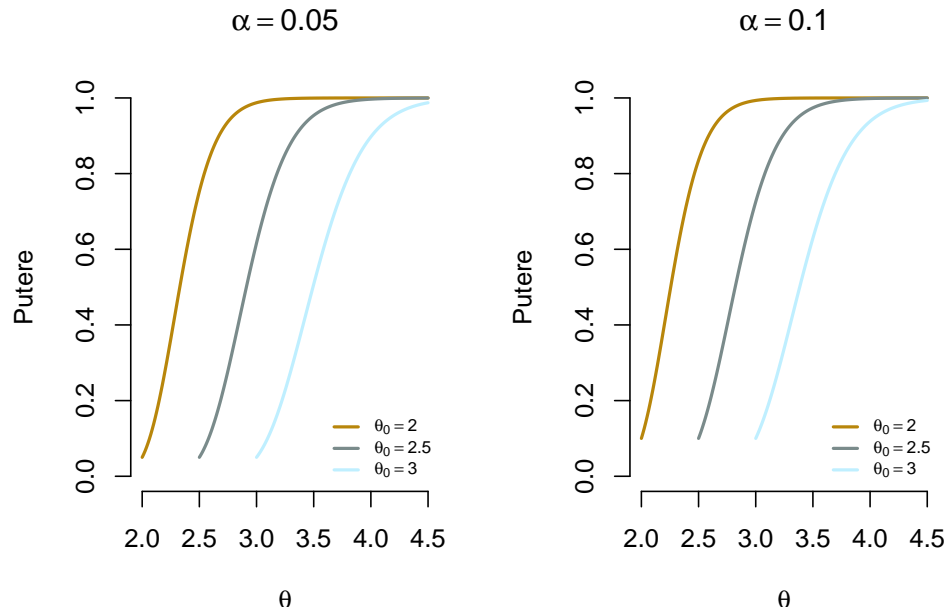
$$\text{pow}(\theta) = 1 - \mathbb{P}_{H_1}\left(\sqrt{n} \frac{\hat{\theta}_n - \theta}{\theta} < \sqrt{n} \frac{a - \theta}{\theta}\right) = 1 - \Phi\left(\sqrt{n} \frac{a - \theta}{\theta}\right) = 1 - \Phi\left(\sqrt{n} \frac{\theta_0 - \theta}{\theta} + \frac{\theta_0}{\theta} z_{1-\alpha}\right), \quad \theta > \theta_0.$$

Particularizând, pentru  $\theta_0 = 2$ ,  $n = 100$  și  $\alpha = 0.1$  obținem

$$\text{pow}(\theta) = 1 - \Phi\left(10 \frac{2 - \theta}{\theta} + \frac{2.5631}{\theta}\right), \quad \theta > 2.$$

Ilustrăm în R funcția putere a testului pentru ipotezele  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta > \theta_0$ :

```
pow_graf = function(theta, theta0, alpha, n){  
  z = qnorm(1-alpha)  
  pow = 1 - pnorm(sqrt(n)*(theta0 - theta)/theta + theta0/theta*z)  
  
  return(pow)  
}
```



Considerăm testul bilateral pentru ipotezele

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

Care este regiunea critică a testului de mărime  $\alpha$ ?

Considerăm testele unilaterale

$$\text{Testul A} \quad H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta < \theta_0$$

$$\text{Testul B} \quad H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0$$

Regiunile critice ale celor două teste unilaterale UMP de mărime  $\frac{\alpha}{2}$  sunt, după cum am văzut la întrebările precedente, date de

$$C_A = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) < \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{\frac{\alpha}{2}} \right\}$$

$$C_B = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) > \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\}$$

iar regiunea critică a testului bilateral este dată de reuniunea acestora

$$C = C_A \cup C_B = \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) < \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{\frac{\alpha}{2}} \right\} \cup \left\{ \mathbf{x} \mid \hat{\theta}_n(\mathbf{x}) > \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\}.$$

Știind că  $z_{\frac{\alpha}{2}} = -z_{1-\frac{\alpha}{2}}$  această regiune critică se poate scrie sub forma

$$C = \left\{ \mathbf{x} \mid \left| \hat{\theta}_n(\mathbf{x}) - \theta_0 \right| > \frac{\theta_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\}.$$



Determinați puterea testului bilateral de mărime  $\alpha$  pentru ipotezele

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

Ilustrați grafic în R pentru  $n = 100$ ,  $\theta_0 = 2$  și  $\alpha = 0.1$ .

Pentru a determina puterea testului avem, conform definiției, că

$$\text{pow}(\theta) = \mathbb{P}_{H_1}(\mathbf{x} \in C) = 1 - \mathbb{P}_{H_1}(\mathbf{x} \in C^c).$$

Regiunea de acceptare  $C^c$  este dată de

$$C^c = \left\{ \mathbf{x} \mid \theta_0 - \frac{\theta_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}} < \hat{\theta}_n(\mathbf{x}) < \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right\}$$

de unde funcția putere devine

$$\text{pow}(\theta) = 1 - \mathbb{P}_{H_1} \left( \hat{\theta}_n(\mathbf{x}) < \theta_0 + \frac{\theta_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right) + \mathbb{P}_{H_1} \left( \hat{\theta}_n(\mathbf{x}) < \theta_0 - \frac{\theta_0}{\sqrt{n}} z_{1-\frac{\alpha}{2}} \right).$$

Sub ipoteza alternativă,  $H_1$ , am văzut că estimatorul de verosimilitate maximă  $\hat{\theta}_n(\mathbf{x})$  este repartizat asimptotic

$$\hat{\theta}_n \underset{H_1}{\approx} \mathcal{N} \left( \theta, \frac{\theta^2}{n} \right), \quad \theta \neq \theta_0$$

prin urmare funcția putere se scrie

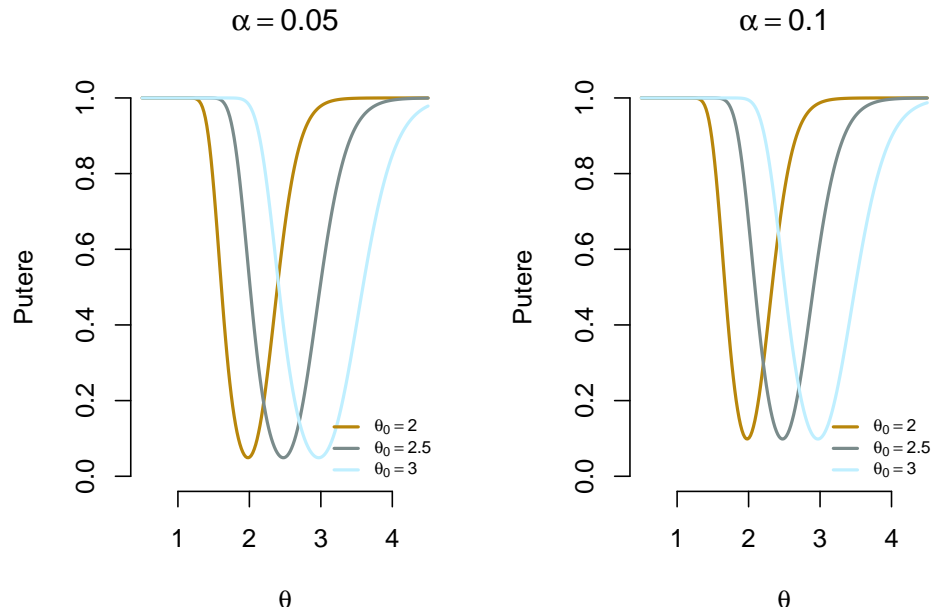
$$\text{pow}(\theta) \approx 1 - \Phi \left( \sqrt{n} \frac{\theta_0 - \theta}{\theta} + \frac{\theta_0}{\theta} z_{1-\frac{\alpha}{2}} \right) + \Phi \left( \sqrt{n} \frac{\theta_0 - \theta}{\theta} - \frac{\theta_0}{\theta} z_{1-\frac{\alpha}{2}} \right), \quad \theta \neq \theta_0$$


Particularizând, pentru  $\theta_0 = 2$ ,  $n = 100$  și  $\alpha = 0.1$  obținem

$$\text{pow}(\theta) = 1 - \Phi \left( 10 \frac{2 - \theta}{\theta} + \frac{2.5631}{\theta} \right) + \Phi \left( 10 \frac{2 - \theta}{\theta} - \frac{2.5631}{\theta} \right), \quad \theta \neq 2.$$

Ilustrăm în R funcția putere a testului bilateral pentru ipotezele  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta \neq \theta_0$ :

```
pow_graf_bilateral = function(theta, theta0, alpha, n){  
  z = qnorm(1-alpha/2)  
  pow = 1 - pnorm(sqrt(n)*(theta0 - theta)/theta + theta0/theta*z) +  
    pnorm(sqrt(n)*(theta0 - theta)/theta - theta0/theta*z)  
  
  return(pow)  
}
```



 Arătați că testul bilateral este nedeplasat și consistent.

Am văzut că puterea testului bilateral este dată de

$$pow(\theta) \approx 1 - \Phi\left(\sqrt{n}\frac{\theta_0 - \theta}{\theta} + \frac{\theta_0}{\theta}z_{1-\frac{\alpha}{2}}\right) + \Phi\left(\sqrt{n}\frac{\theta_0 - \theta}{\theta} - \frac{\theta_0}{\theta}z_{1-\frac{\alpha}{2}}\right), \theta \neq \theta_0$$

Pentru  $\theta < \theta_0$  obținem

$$\lim_{n \rightarrow \infty} pow(\theta) = 1 - \Phi(\infty) + \Phi(\infty) = 1 - 1 + 1 = 1$$

iar pentru  $\theta > \theta_0$

$$\lim_{n \rightarrow \infty} pow(\theta) = 1 - \Phi(-\infty) + \Phi(-\infty) = 1 - 0 + 0 = 1$$

deci testul este consistent.

Pentru a vedea dacă testul este nedeplasat trebuie să calculăm minimul funcției putere. Se poate observa că acest minim se atinge pentru  $\theta \rightarrow \theta_0$ , și cum

$$\lim_{\theta \rightarrow \theta_0} pow(\theta) = 1 - \Phi\left(z_{1-\frac{\alpha}{2}}\right) + \Phi\left(-z_{1-\frac{\alpha}{2}}\right) = 1 - \left(1 - \frac{\alpha}{2}\right) + \frac{\alpha}{2} = \alpha$$

deducem că testul este nedeplasat.

## 2.2 Test bazat pe raportul de verosimilități

Presupunem că  $Y$  este o variabilă aleatoare care ia valori în mulțimea  $\{y_1, y_2, \dots, y_c\}$  iar repartiția ei este dată de



$$\mathbb{P} \circ Y^{-1} = \sum_{j=1}^c p_j \delta_{y_j},$$

unde  $\mathbb{P}(Y = y_j) = p_j$ ,  $j \in \{1, 2, \dots, c\}$ .

Fie  $Y_1, Y_2, \dots, Y_n$  un eșantion de talie  $n$  din populația  $\mathbb{P} \circ Y^{-1}$  și

$$N_i = \sum_{k=1}^n \mathbf{1}_{y_i}(Y_k)$$

numărul de observații care categoria  $y_i$ . Observăm că variabilele aleatoare  $N_1, N_2, \dots, N_c$  verifică

$$N_1 + N_2 + \dots + N_c = n.$$

Putem modela o observație  $Y_k$  dintr-o variabilă discretă cu  $c$  categorii cu ajutorul unui vector elemente de  $\{0, 1\}$ ,  $(X_1^{(k)}, X_2^{(k)}, \dots, X_c^{(k)})$ , pentru care componenta  $j$  ia valoarea 1 dacă  $Y_k = y_j$  și 0 altfel. Funcția de masă a  $c$ -uplului este

$$\mathbb{P}((X_1^{(k)}, X_2^{(k)}, \dots, X_c^{(k)}) = (x_1, x_2, \dots, x_c)) = p_1^{x_1} p_2^{x_2} \dots p_c^{x_c}$$

unde  $x_j \in \{0, 1\}$  cu  $\sum_{j=1}^c x_j = 1$ . Pentru un eșantion de talie  $n$ ,  $\{(X_1^{(k)}, X_2^{(k)}, \dots, X_c^{(k)}), k = 1, 2, \dots, n\}$  avem

$$\mathbb{P}((X_1^{(k)}, X_2^{(k)}, \dots, X_c^{(k)}) = (x_1^{(k)}, x_2^{(k)}, \dots, x_c^{(k)}), k = 1, 2, \dots, n) = \prod_{k=1}^n p_1^{x_1^{(k)}} p_2^{x_2^{(k)}} \dots p_c^{x_c^{(k)}} = p_1^{n_1} p_2^{n_2} \dots p_c^{n_c}$$

unde  $n_j$  reprezintă numărul de observații din categoria  $y_j$  iar  $\sum_{j=1}^c n_j = n$ .

În următorul exercițiu ne propunem să aplicăm testul bazat pe raportul de verosimilități pentru efectuarea unui test asupra parametrilor unei repartiții multinomiale.



Spunem că vectorul  $(N_1, N_2, \dots, N_c)$  este repartizat multinomial  $\mathcal{M}(n; p_1, p_2, \dots, p_c)$  dacă

$$\mathbb{P}(N_1 = n_1, N_2 = n_2, \dots, N_c = n_c) = \frac{n!}{n_1! n_2! \dots n_c!} p_1^{n_1} p_2^{n_2} \dots p_c^{n_c}$$

unde  $n_1 + n_2 + \dots + n_c = n$ .

Determinați repartiția marginală a lui  $N_i$ .

Observăm că

$$\begin{aligned}
 \mathbb{P}(N_i = n_i) &= \sum_{n_1} \cdots \sum_{n_{i-1}} \sum_{n_{i+1}} \cdots \sum_{n_c} \mathbb{P}(N_1 = n_1, \dots, N_{i-1} = n_{i-1}, N_i = n_i, N_{i+1} = n_{i+1}, \dots, N_c = n_c), \quad n_1 + \cdots + n_c = n - n_i \\
 &= \sum_{n_1} \cdots \sum_{n_{i-1}} \sum_{n_{i+1}} \cdots \sum_{n_c} \frac{n!}{n_1! \cdots n_c!} p_1^{n_1} \cdots p_c^{n_c}, \quad n_1 + \cdots + n_c = n - n_i \\
 &= \frac{n! p_i^{n_i}}{n_i! (n - n_i)!} \sum_{n_1} \cdots \sum_{n_{i-1}} \sum_{n_{i+1}} \cdots \sum_{n_c} \frac{(n - n_i)!}{n_1! \cdots n_{i-1}! n_{i+1}! \cdots n_c!} p_1^{n_1} \cdots p_{i-1}^{n_{i-1}} p_{i+1}^{n_{i+1}} \cdots p_c^{n_c}, \quad n_1 + \cdots + n_c = n - n_i \\
 &= \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n - n_i} \underbrace{\sum_{n_1} \cdots \sum_{n_{i-1}} \sum_{n_{i+1}} \cdots \sum_{n_c} \binom{n - n_i}{n_1, \dots, n_c} \left( \frac{p_1}{1 - p_i} \right)^{n_1} \cdots \left( \frac{p_{i-1}}{1 - p_i} \right)^{n_{i-1}} \left( \frac{p_{i+1}}{1 - p_i} \right)^{n_{i+1}} \cdots \left( \frac{p_c}{1 - p_i} \right)^{n_c}}_{=1} \\
 &= \binom{n}{n_i} p_i^{n_i} (1 - p_i)^{n - n_i}
 \end{aligned}$$

prin urmare  $N \sim \mathcal{B}(n, p_i)$ .



Considerăm ipotezele

$$\begin{aligned}
 H_0 &: \{(p_1, p_2, \dots, p_c) = (\pi_1, \pi_2, \dots, \pi_c)\} \\
 H_1 &: \{\exists i \text{ astfel incat } p_i \neq \pi_i\}
 \end{aligned}$$

unde  $(\pi_1, \pi_2, \dots, \pi_c)$  sunt probabilități specificate în avans. Construim testul bazat pe raportul de verosimilități corespunzător.

Testul bazat pe raportul de verosimilitate este

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\mathbf{x})}{\sup_{\theta \in \Theta} L(\theta|\mathbf{x})},$$

unde  $\Theta$  este spațiul parametrilor modelului,  $\Theta_0$  este spațiul parametrilor corespunzător ipotezei nule iar  $L(\theta|\mathbf{x})$  este funcția de verosimilitate.

Observăm că spațiul parametrilor corespunzător modelului este

$$\Theta = \left\{ p_1, p_2, \dots, p_c \mid p_j \in (0, 1), \sum_{j=1}^c p_j = 1 \right\},$$

cu  $\dim \Theta = c - 1$ , cel corespunzător ipotezei nule este

$$\Theta_0 = \{(p_1, p_2, \dots, p_c) = (\pi_1, \pi_2, \dots, \pi_c)\}$$

cu  $\dim \Theta_0 = 0$  iar funcția de verosimilitate este

$$L(p_j, j = 1, \dots, c; \mathbf{x}) = \mathbb{P}(N_j = n_j, j = 1, \dots, c) = \frac{n!}{\prod_{j=1}^c n_j!} \prod_{j=1}^c p_j^{n_j}.$$

Observăm că

$$\sup_{\theta \in \Theta_0} L(\theta|\mathbf{x}) = \mathbb{P}_{H_0}(N_1 = n_1, N_2 = n_2, \dots, N_c = n_c) = \frac{n!}{\prod_{j=1}^c n_j!} \prod_{j=1}^c \pi_j^{n_j}.$$

Pentru a determina estimatorul de verosimilitate maximă pe  $\Theta$  trebuie să rezolvăm problema de optimizare:

$$\begin{cases} \max_{\theta \in \Theta} \log L(\theta|\mathbf{x}) = \max \log \left( \frac{n!}{\prod_{j=1}^c n_j!} \prod_{j=1}^c p_j^{n_j} \right) \\ \sum_{j=1}^c p_j = 1 \end{cases}$$

Cum logaritmul funcției de verosimilitate este

$$\log \left( \frac{n!}{\prod_{j=1}^c n_j!} \prod_{j=1}^c p_j^{n_j} \right) = \log \left( \frac{n!}{\prod_{j=1}^c n_j!} \right) + \sum_{j=1}^c n_j \log p_j$$

iar  $p_c = 1 - p_1 - \dots - p_{c-1}$ , rezolvând ecuația de verosimilitate  $\frac{\partial \log L}{\partial p_j} = 0$  deducem

$$\begin{cases} \frac{n_1}{p_1} - \frac{n_c}{1-p_1-p_2-\dots-p_{c-1}} = 0 \\ \frac{n_2}{p_2} - \frac{n_c}{1-p_1-p_2-\dots-p_{c-1}} = 0 \\ \dots\dots\dots \\ \frac{n_{c-1}}{p_{c-1}} - \frac{n_c}{1-p_1-p_2-\dots-p_{c-1}} = 0 \end{cases}$$

de unde

$$\frac{n_1}{p_1} = \frac{n_2}{p_2} = \dots = \frac{n_c}{p_c} = \frac{\sum_{j=1}^c n_j}{\sum_{j=1}^c p_j} = n,$$

deci  $\hat{p}_j = \frac{n_j}{n}$ .

Raportul de verosimilitate devine

$$\Lambda(\mathbf{x}) = \frac{\sup_{\theta \in \Theta_0} L(\theta|\mathbf{x})}{\sup_{\theta \in \Theta} L(\theta|\mathbf{x})} = \frac{\prod_{j=1}^c \pi_j^{n_j}}{\prod_{j=1}^c \left(\frac{n_j}{n}\right)^{n_j}} = \prod_{j=1}^c \left(\frac{n\pi_j}{n_j}\right)^{n_j}$$

și aplicând Teorema lui Wilks găsim

$$-2 \log \Lambda(\mathbf{x}) = 2 \sum_{j=1}^c n_j \log \left( \frac{n_j}{n \times \pi_j} \right) \xrightarrow[n \rightarrow \infty]{d} \chi^2(\underbrace{\dim \Theta - \dim \Theta_0}_{(c-1)-0}) = \chi^2(c-1).$$

Prin urmare, regiunea critică a testului asimptotic de nivel  $\alpha$  bazat pe raportul de verosimilitate este

$$C = \{\mathbf{x} \mid -2 \log \Lambda(\mathbf{x}) > \chi_{1-\alpha}^2(c-1)\}.$$

O metodă alternativă este bazată pe statistica  $\chi^2$  a lui Pearson, [Pearson, 1900], care este dată de

$$X^2 = \sum_{j=1}^c \frac{(O_j - E_j)^2}{E_j}$$

unde  $O_j = n_j$  sunt efectivele observate iar  $E_j$  sunt efectivele pe care ne așteptăm să le observăm dacă ipoteza nulă ar fi adevărată (sub  $H_0$ , repartiția condiționată a lui  $N_j$  la  $\sum_{j=1}^c N_j = n$  este  $\mathcal{B}(n, p_j)$ ),

$$E_j = \mathbb{E}_{H_0}[N_j] = n\pi_j$$

ceea ce implică

$$X^2 = \sum_{j=1}^c \frac{(n_j - n\pi_j)^2}{n\pi_j}.$$

Karl Pearson a arătat că această statistică este asimptotic repartizată

$$X^2 = \sum_{j=1}^c \frac{(n_j - n\pi_j)^2}{n\pi_j} \xrightarrow[n \rightarrow \infty]{d} \chi^2(c-1)$$

prin urmare testul  $\chi^2$  a lui Pearson de nivel  $\alpha$ , pentru ipotezele  $H_0$  vs  $H_1$  conduce la aceeași regiune critică ca și testul bazat pe raportul de verosimilitate (cu toate acestea se poate arăta că statistica lui Pearson  $X^2$  converge mai repede decât statistica  $-2 \log \Lambda(\mathbf{x})$ , [Agresti, 2012]).



Un exercițiu constă în extragerea la întâmplare, de către o persoană, a unei cărți de joc dintr-un pachet amestecat în prealabil, notarea culorii acesteia (inimă roșie, inimă neagră, romb și treflă) și cărții în pachet tot aleator. Să presupunem că în urma efectuării exercițiului pe 200 de persoane s-au obținut următoarele rezultate: 35 cărți de treflă, 51 cărți de romb, 64 cărți de inimă roșie și respectiv 50 cărți de inimă neagră. Ne propunem să testăm ipotezele

$$H_0 : \{\text{Cele patru culori sunt egal probabile}\}$$

$$H_1 : \{\text{Cel puțin una din cele patru culori este diferită de 0.25}\}$$

Observăm că dacă notăm cu  $Y$  variabila discretă care ia valori în mulțimea  $\{y_1, y_2, y_3, y_4\} = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$  și cu  $p_j = \mathbb{P}(Y = y_j)$  atunci ipotezele se scriu

$$H_0 : \{\text{Cele patru culori sunt egal probabile}\} = \{(p_1, p_2, p_3, p_4) = (\pi_1, \pi_2, \pi_3, \pi_4) \mid \pi_j = 0.25\}$$

$$H_1 : \{\text{Cel puțin una din cele patru culori este diferită de 0.25}\} = \{\exists j, \text{ astfel ca } p_j \neq 0.25\}$$

Avem că tabelul efectivelor observate este

$$\begin{array}{c|c|c|c} \clubsuit & \diamondsuit & \heartsuit & \spadesuit \\ \hline O_1 & O_2 & O_3 & O_4 \end{array} = \begin{array}{c|c|c|c} \clubsuit & \diamondsuit & \heartsuit & \spadesuit \\ \hline 35 & 51 & 64 & 50 \end{array}$$

care în R este

```
tab_observed = c(35, 51, 64, 50)
```

iar tabelul efectivelor pe care ne așteptăm să le observăm dacă ipoteza nulă este adevărată este

$$\begin{array}{c|c|c|c} \clubsuit & \diamondsuit & \heartsuit & \spadesuit \\ \hline E_1 = n\pi_1 & E_2 = n\pi_2 & E_3 = n\pi_3 & E_4 = n\pi_4 \end{array} = \begin{array}{c|c|c|c} \clubsuit & \diamondsuit & \heartsuit & \spadesuit \\ \hline 50 & 50 & 50 & 50 \end{array}$$

care în R se scrie

```
n = 200
prob = rep(0.25, 4)

tab_expected = n*prob
```

Putem calcula acum statistica lui Pearson

$$X^2 = \sum_{j=1}^c \frac{(O_j - E_j)^2}{E_j} = \sum_{j=1}^c \frac{(n_j - n\pi_j)^2}{n\pi_j}$$

care devine

```
alpha = 0.05
X2 = sum((tab_observed - tab_expected)^2/tab_expected)

# p - valoarea
1 - pchisq(X2, df = 3)
[1] 0.03774185
```

Acelați rezultat îl obținem și dacă aplicăm funcția `chisq.test()`

```
chisq.test(tab_observed, p = prob)

Chi-squared test for given probabilities

data:  tab_observed
X-squared = 8.44, df = 3, p-value = 0.03774
```

Pentru testul bazat pe raportul de verosimilitate maximă avem

$$-2 \log \Lambda(\mathbf{x}) = -2 \sum_{j=1}^c n_j \log \left( \frac{n_j}{n \times \pi_j} \right)$$

care devine

```
LRT = 2*sum(tab_observed*log(tab_observed/tab_expected))

# p - valoarea
1 - pchisq(LRT, df = 3)
[1] 0.03431406
```

Ambele proceduri conduc la respingerea ipotezei nule pentru un prag de semnificație  $\alpha = 0.05$ .

## 2.3 Testul raportului de verosimilități, testul lui Wald și testul de scor



Considerăm două variabile aleatoare  $X$  și  $Y$  astfel încât repartiția condiționată a lui  $Y|X = x$  este dată de

$$f_{Y|X}(y|x; \beta) = \frac{1}{\beta + x} e^{-\frac{y}{\beta + x}}$$

Vom nota pentru simplitate cu  $\beta_i = \frac{1}{\beta + x_i}$ . Repartiția condiționată de mai sus este o repartiție exponențială, care la rândul ei poate fi privită ca un caz particular de repartiție Gamma, cu  $\rho = 1$ ,

$$f_{Y|X}(y|x; \beta, \rho) = \frac{\beta_i^\rho}{\Gamma(\rho)} y^{\rho-1} e^{-y\beta_i}$$

Vrem să testăm ipotezele

$$H_0 : \rho = 1 \quad \text{vs} \quad H_1 : \rho \neq 1$$

Începem prin a reaminti că funcția  $\Gamma(p)$  este definită prin (vezi și [https://en.wikipedia.org/wiki/Gamma\\_function](https://en.wikipedia.org/wiki/Gamma_function))

$$\Gamma(p) = \int_0^\infty t^{p-1} e^{-t} dt, \quad p > 0,$$

verifică relația de recurență

$$\Gamma(p) = (p-1)\Gamma(p-1)$$

și  $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ . Mai mult, dacă  $n \in \mathbb{N}$  atunci  $\Gamma(n+1) = n!$ .

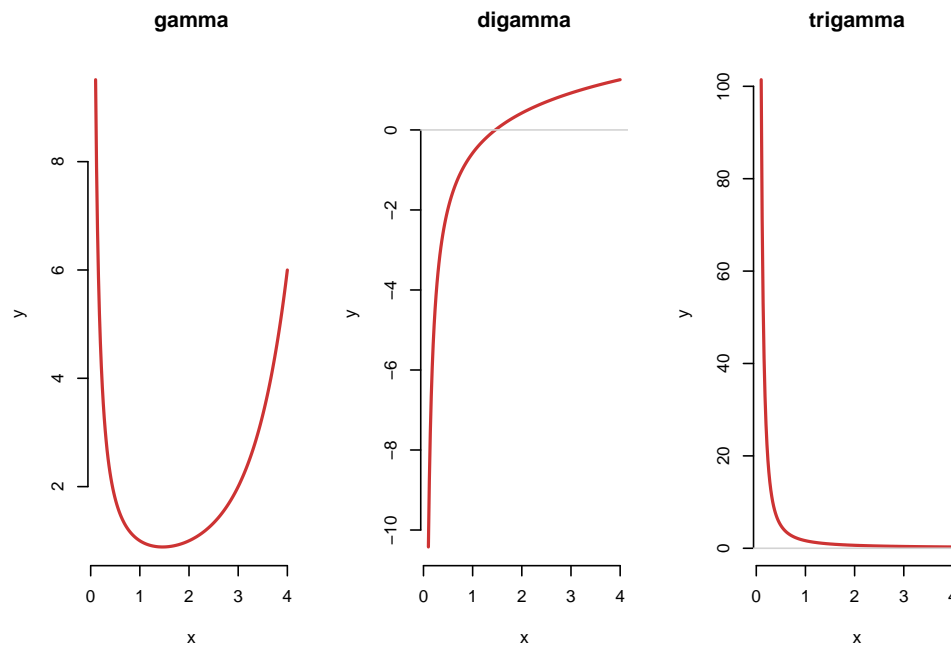
De asemenea, derivatele funcției  $\Gamma$  sunt

$$\frac{d^k \Gamma(p)}{dp^k} = \int_0^\infty (\log(t))^k t^{p-1} e^{-t} dt$$

iar primele două derivate ale funcției  $\log(\Gamma(p))$ , cunoscute și ca funcțiile digamma și respectiv trigamma, se notează cu  $\Psi$  și respectiv  $\Psi'$  și sunt definite prin

$$\Psi = \frac{\partial \log(\Gamma(p))}{\partial p} = \frac{\Gamma'}{\Gamma}$$
$$\Psi' = \frac{\partial^2 \log(\Gamma(p))}{\partial p^2} = \frac{\Gamma \Gamma'' - \Gamma'^2}{\Gamma^2}$$

Aceste funcții sunt implementate în R cu ajutorul funcțiilor `digamma()` și respectiv `trigamma()`.



Fie  $\{X_i, Y_i\}$  un eșantion de talie  $n$  din populația  $f_{Y|X}$  și scrieți logaritmul funcției de verosimilitate pentru modelul necondiționat și respectiv sub  $H_0$  (modelul condiționat).

Considerăm parametrul  $\theta = (\beta, \rho)^\top$  și avem

$$f_{Y_i|X_i}(y|x; \theta) = \frac{\beta_i^\rho}{\Gamma(\rho)} y_i^{\rho-1} e^{-y_i \beta_i}, \quad \text{cu} \quad \beta_i = \frac{1}{\beta + x_i}$$

Cum logaritmul funcției de verosimilitate este

$$l(y|x; \theta) = \sum_{i=1}^n \log f_{Y_i|X_i}(y|x; \theta)$$

deducem că, sub modelul necondiționat,

$$l(y|x; \theta) = \rho \sum_{i=1}^n \beta_i - n \log \Gamma(\rho) + (\rho - 1) \sum_{i=1}^n y_i - \sum_{i=1}^n y_i \beta_i.$$

Sub  $H_0$ ,  $\rho = 1$  avem

$$f_{Y_i|X_i}(y|x; \theta) = \beta_i e^{-y_i \beta_i}, \quad \text{cu} \quad \beta_i = \frac{1}{\beta + x_i}$$

și cum

$$l(y|x; \theta) = \sum_{i=1}^n \log f_{Y_i|X_i}(y|x; \theta)$$

găsim că logaritmul funcției de verosimilitate, sub  $H_0$ , este

$$l(y|x; \theta) = \sum_{i=1}^n \beta_i - \sum_{i=1}^n y_i \beta_i.$$



Scrieți vectorii gradient și matricele Hessiene pentru logaritmul funcției de verosimilitate asociat modelului necondiționat și respectiv modelului condiționat (sub  $H_0$ ).

Am văzut la punctul anterior că logaritmul funcției de verosimilitate pentru modelul necondiționat este

$$l(y|x; \theta) = \rho \sum_{i=1}^n \beta_i - n \log \Gamma(\rho) + (\rho - 1) \sum_{i=1}^n y_i - \sum_{i=1}^n y_i \beta_i.$$

Din definiția lui  $\beta_i = \frac{1}{\beta + x_i}$  avem că

$$\begin{aligned} \frac{\partial \beta_i}{\partial \beta} &= \frac{\partial \left( \frac{1}{\beta + x_i} \right)}{\partial \beta} = -\frac{1}{(\beta + x_i)^2} = -\beta_i^2 \\ \frac{\partial \log \beta_i}{\partial \beta} &= \frac{\partial (-\log \beta + x_i)}{\partial \beta} = -\frac{1}{\beta + x_i} = -\beta_i \end{aligned}$$

iar  $\frac{\partial \log \Gamma(\rho)}{\partial \rho} = \Psi(\rho)$ .

Prin urmare găsim că

$$\begin{aligned} \frac{\partial l(y|x; \theta)}{\partial \beta} &= -\rho \sum_{i=1}^n \beta_i + \sum_{i=1}^n y_i \beta_i^2 \\ \frac{\partial l(y|x; \theta)}{\partial \rho} &= \sum_{i=1}^n \log \beta_i - n \Psi(\rho) + \sum_{i=1}^n \log y_i \end{aligned}$$

astfel, vectorul gradient sub modelul necondiționat este

$$\frac{\partial l(y|x; \theta)}{\partial \theta} = \begin{pmatrix} \frac{\partial l(y|x; \theta)}{\partial \beta} \\ \frac{\partial l(y|x; \theta)}{\partial \rho} \end{pmatrix} = \begin{pmatrix} -\rho \sum_{i=1}^n \beta_i + \sum_{i=1}^n y_i \beta_i^2 \\ \sum_{i=1}^n \log \beta_i - n \Psi(\rho) + \sum_{i=1}^n \log y_i \end{pmatrix}.$$

Pentru matricea Hessiană avem

$$H(y|x; \theta) = \frac{\partial^2 l(y|x; \theta)}{\partial \theta \partial \theta^T} = \begin{pmatrix} \frac{\partial^2 l(y|x; \theta)}{\partial \beta^2} & \frac{\partial^2 l(y|x; \theta)}{\partial \beta \partial \rho} \\ \frac{\partial^2 l(y|x; \theta)}{\partial \rho \partial \beta} & \frac{\partial^2 l(y|x; \theta)}{\partial \rho^2} \end{pmatrix}$$

și cum



$$\begin{aligned}\frac{\partial^2 l(y|x; \theta)}{\partial \beta^2} &= \rho \sum_{i=1}^n \beta_i^2 - 2 \sum_{i=1}^n y_i \beta_i^3 \\ \frac{\partial^2 l(y|x; \theta)}{\partial \beta \partial \rho} &= \frac{\partial^2 l(y|x; \theta)}{\partial \rho \partial \beta} = - \sum_{i=1}^n \beta_i \\ \frac{\partial^2 l(y|x; \theta)}{\partial \rho^2} &= -n \Phi'(\rho)\end{aligned}$$

găsim

$$H(y|x; \theta) = \begin{pmatrix} \rho \sum_{i=1}^n \beta_i^2 - 2 \sum_{i=1}^n y_i \beta_i^3 & \sum_{i=1}^n \beta_i \\ \sum_{i=1}^n \beta_i & -n \Phi'(\rho) \end{pmatrix}.$$

Sub ipoteza nulă  $H_0$ , avem  $\rho = 1$  ( $\theta = \beta$ ) deci vectorul gradient (care acum e scalar) este

$$\frac{\partial l(y|x; \theta)}{\partial \theta} = \frac{\partial l(y|x; \beta)}{\partial \beta} = - \sum_{i=1}^n \beta_i + \sum_{i=1}^n y_i \beta_i^2$$

iar matricea Hessiană (care este tot scalară) este

$$H(y|x; \theta) = \frac{\partial^2 l(y|x; \beta)}{\partial \beta^2} = \sum_{i=1}^n \beta_i^2 - 2 \sum_{i=1}^n y_i \beta_i^3.$$



Scrieți matricea informațională medie a lui Fisher pentru modelul necondiționat și respectiv modelul condiționat (sub  $H_0$ ).

## Referințe

Alan Agresti. *Categorical Data Analysis*. Wiley, 3 edition, 2012. URL <http://gen.lib.rus.ec/book/index.php?md5=640290AD3F29427A30C48D483E420C96>.

Brian Peacock Merran Evans, Nicholas Hastings. *Statistical Distributions*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2000. URL <http://gen.lib.rus.ec/book/index.php?md5=EED460A11523229329F6DA85E3FF7936>.

K. Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, 50:157–175, 1900.