Anexa 3. Statistică descriptivă şi inferențială

3.1 Măsuri statistice pentru populații şi eşantioane

Tabelul 1. Măsuri statistice pentru caracterizarea variabilelor cantitative

Măsură	Referă	Expresie	Interpretare
Suma valorilor	Un şir de numere	$\Sigma(\cdot)$	-
Numărul de valori		$\lvert\cdot\rvert$	-
Valoarea medie		$E(\cdot) = \Sigma(\cdot)/\lvert\cdot\rvert$	Valoarea aşteptată
Moment central de ordin k, k>1		$E_k(\cdot) = E((X-E(X))^k)$	-
Media caracteristicii X	O populație	$\mu = \mu(X) = E(X)$	Tendința centrală
Media observabilei Y	Un eşantion	$m = m(Y) = E(Y)$	
Estimatorul mediei caracteristicii X	O populație	$M(Y) = m(Y)$	
Varianța caracteristicii X	O populație	$Var(X) = E((X-\mu)^2)$	Împrăştierea
Deviația standard a caracteristicii X		$\sigma = \sigma(X) = \sqrt{Var(X)}$	Dispersia
Varianța observabilei Y	Un eşantion	$var = var(Y) = E((Y-E(Y))^2)$	Împrăştierea
Deviația standard a observabilei Y		$s = s(Y) = \sqrt{var(Y)}$	Dispersia
Estimatorul varianței caracteristicii X	O populație	$VAR(Y) = \frac{\lvert Y\rvert}{\lvert Y\rvert - 1} var(Y)$	Împrăştierea
Estimatorul deviației standard a caracteristicii X		$S = S(Y) = \sqrt{\frac{\lvert Y\rvert}{\lvert Y\rvert - 1}} s(Y)$	Dispersia

Tabelul 2. Statistici pentru caracterizarea depărtării de normalitate a variabilelor cantitative

Simbol şi măsură	Referă	Expresie	Mărimi care intervin
γ_1, Asimetria caracteristicii X	O populație	$\gamma_1 = \mu_3/\mu_2^{3/2}$	$\mu_k = E_k(X)$, k>1
β_2, Boltirea caracteristicii X		$\beta_2 = \mu_4/\mu_2^2$	
γ_2, Excesul de boltire al caracteristicii X		$\gamma_2 = \beta_2 - 3$	
g_1, Asimetria observabilei Y	Un eşantion	$g_1 = m_3/m_2^{3/2}$	$m_k = E_k(Y)$, k>1
b_2, Boltirea observabilei Y		$b_2 = m_4/m_2^2$	
g_2, Excesul de boltire al observabilei Y		$g_2 = b_2 - 3$	
Estimatorul asimetriei caracteristicii X	O populație	$G_1 = \frac{\sqrt{n_Y(n_Y-1)}}{(n_Y-2)} M_3/M_2^{3/2}$	$n_Y = \lvert Y\rvert$ $M_k = \frac{n_Y}{n_Y-1} E_k(Y)$, k>1
Estimatorul boltirii caracteristicii X		$B_2 = \frac{(n_Y-1)(n_Y+1)}{(n_Y-2)(n_Y-3)} M_4/M_2^2$	
Estimatorul excesului de boltire a caracteristicii X		$G_2 = B_2 - 3\cdot\frac{(n_Y-1)^2}{(n_Y-2)(n_Y-3)}$	

Extragerea repetată de eşantioane (de volum dat) dintr-o populație face ca valorile obținute să urmeze o distribuție, numită distribuția de eşantionare. Tabelul 3 prezintă rezultatele care se obțin pentru varianța mărimilor statistice prin extragerea repetată de eşantioane dintr-o populație.

Când valorile parametrilor statistici ai populației nu sunt cunoscute, dar se poate face presupunerea că distribuția populației se comportă suficient de bine [1], aceştia pot fi aproximați cu ajutorul estimatorilor acestora (Tabelul 1). Formulele de calcul aproximativ ale mediei şi variației pentru medie şi varianță sunt redate în Tabelul 4. Dacă se pot asuma ipoteze cu privire la distribuția caracteristicii X în populație, atunci se pot obține formule de calcul pentru parametrii statistici (ai populației) şi folosind relațiile din Tabelul 1 estimatorii parametrilor statistici ai populației din măsurătorile (statisticile) efectuate asupra eşantionului.

Tabelul 3. Medii şi variațe ale mediei şi variației observabilei Y ce rezultă din distribuția de eşantionare din populația cu caracteristica X

Mărime şi notație	Valoare
Media mediei, $\mu_{\bar{Y}}$	$\mu_{\bar{Y}} = \mu(m(Y)) = \mu(X)$
Varianța mediei, $\sigma_{\bar{Y}}^2$	$\sigma_{\bar{Y}}^2 = \sigma^2(m(Y)) = \frac{\sigma^2(X)}{n_Y}$
Media variației, $\mu(s^2)$	$\mu(s^2) = \mu(s^2(Y)) = \frac{(n_Y - 1)}{n_Y}\sigma^2(X)$
Varianța variației, $\sigma^2(s^2)$	$\sigma^2(s^2) = \sigma^2(s^2(Y)) = \frac{(n_Y - 1)^2}{n_Y^3}\mu_4(X) - \frac{(n_Y - 1)(n_Y - 3)}{n_Y^3}\mu_2^2(X)$

Tabelul 4. Valori aproximative pentru mediile şi variațele mediei şi variației observabilei Y în ipotezele teoremei limită centrale

Mărime şi notație	Aproximare
Media mediei, $\mu_{\bar{Y}}$	$\mu_{\bar{Y}} \cong m(Y)$
Varianța mediei, $\sigma_{\bar{Y}}^2$	$\sigma_{\bar{Y}}^2 \cong \frac{s^2(Y)}{(n_Y - 1)}$
Media variației, $\mu(s^2)$	$\mu(s^2) \cong s^2(Y)$
Varianța variației, $\sigma^2(s^2)$	$\sigma^2(s^2) \cong \frac{(n_Y - 1)}{n_Y^2}m_4(Y) - \frac{(n_Y - 3)}{n_Y(n_Y - 1)}m_2^2(Y)$

[1] Teorema Limită Centrală

÷ Cronologia contribuțiilor majore:

- Abraham DE MOIVRE. 1733. Approximatio ad Summam Terminorum Binomii $(a+b)^n$ in Seriem expansi. In: The Doctrine of Chance: or The Method of Calculating the Probability of Events in Play (Abraham DE MOIVRE). W. Pearforn 1738: 235-243.
- Joseph L. LAGRANGE. 1776. Mémoire sur l'utilité de la méthode de prendre le milieu entre les résultats de plusieurs observations; dans lequel on examine les avantages de cette méthode par le calcul des probabilités; et où l'on résoud différents problèmes relat ifs à cette matière. Miscellanea Taurinensia 5:167-232.
- Pierre S. LAPLACE. 1812. Théorie Analytique des Probabilités. Courcier, 465 p.
- Aleksandr M. LIAPUNOV. 1901. Nouvelle forme du théoreme sur la limite des probabilités. Mémoires de l'Académie Impériale des Sciences de St. Pétersbourg 12(5):1-24.

÷ Enunțul teoremei (fie $(X_n)_{n\geq 1}$ variabile independente şi $\exists\delta>0$ a.î. $\mu_{2+\delta}(X_n)<\infty$):

- dacă $\lim_{n\to\infty} \frac{\sum_{k=1}^{n}\mu_{2+\delta}(X_k)}{\left(\sum_{k=1}^{n}\sigma_k^2\right)^{(2+\delta)/2}} = 0$ atunci $\frac{\sum_{i=1}^{n}(X_n - \mu_1(X_n))}{\sqrt{\sum_{k=1}^{n}\sigma_k^2}} \xrightarrow[n\to\infty]{} N(0,1)$

3.2 Măsuri statistice pentru legi de distribuție

Tabelele 1-19 dau expresiile unor mărimi statistice (valabile pentru populație) în timp ce expresiile pentru estimatori se pot obține din Tabelul 1 Anexa 3-1.

Tabelul 1. Mărimi statistice ale distribuției discrete uniforme

Mărime statistică	Expresie de calcul
Suport	$k \in \{a, a+1, ..., b-1, b\}$
Minim; Maxim	a; b
Funcția de probabilitate	$1/(b-a+1)$
Funcția de repartiție	$([k]-a+1)/(b-a+1)$
Media şi mediana; varianța	$(a+b)/2$; $((b-a+1)^2-1)/12$
Asimetria; excesul de boltire	0; $-\frac{6((b-a+1)^2+1)}{5((b-a+1)^2-1)}$

Tabelul 2. Mărimi statistice ale distribuției discrete Bernoulli

Mărime statistică	Expresie de calcul	
Suport	$k \in \{0,1\}$; $p \in (0,1)$	
Minim; Maxim	0; 1	
Funcția de probabilitate	(1-p), k = 0	p, k = 1
Funcția de repartiție	(1-p), $k \in [0,1)$	1, $1 \le k$
Media; varianța	p; p(1-p)	
Asimetria; excesul de boltire	0; $(6p^2-6p+1)/(p(1-p))$	

Tabelul 3. Mărimi statistice ale distribuției discrete binomiale

Mărime statistică	Expresie de calcul
Suport	$k \in \{0, ..., n\}$; $p \in (0,1)$
Minim; Maxim	0; n
Funcția de probabilitate	$\frac{n!}{k!\cdot(n-k)!}p^k(1-p)^{n-k}$
Funcția de repartiție	$\sum_{i=0}^{k}\frac{n!}{i!(n-i)!}p^i(1-p)^{n-i}$
Media; varianța	np; np(1-p)
Asimetria; excesul de boltire	$(1-2p)/\sqrt{np(1-p)}$; $\frac{1-6p(1-p)}{np(1-p)}$

Tabelul 4. Mărimi statistice ale distribuției discrete Poisson

Mărime statistică	Expresie de calcul
Suport	$k = 0, 1, ...; \lambda \ge 0$
Minim; Maxim	$0; \infty$
Funcția de probabilitate	$e^{-\lambda}\lambda^k/k!$
Funcția de repartiție	$\sum_{i=0}^{k} e^{-\lambda}\lambda^i/i!$
Media; varianța	$\lambda; \lambda$
Asimetria; excesul de boltire	$1/\sqrt{\lambda}$; $1/\lambda$

Tabelul 5. Mărimi statistice ale distribuției continue uniforme

Mărime statistică	Expresie de calcul
Suport	$x \in [a, b]$
Minim; Maxim	a; b
Funcția de probabilitate	1/(b-a)
Funcția de repartiție	(x-a)/(b-a)
Media şi mediana; varianța	$(a+b)/2$; $(b-a)^2/12$
Asimetria; excesul de boltire	0; -6/5

Tabelul 6. Mărimi statistice ale distribuției continue Cauchy-Lorentz

Mărime statistică	Expresie de calcul
Suport	$x \in (-\infty,\infty)$; $x_0 \in (-\infty,\infty)$; $\gamma \in (0,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\dfrac{1}{\gamma\pi\left(1+\left(\dfrac{x-x_0}{\gamma}\right)^2\right)}$
Funcția de repartiție	$\dfrac{1}{\pi}\arctan\left(\dfrac{x-x_0}{\gamma}\right)+\dfrac{1}{2}$
Mediana şi moda	x_0

Tabelul 7. Mărimi statistice ale distribuției continue Student t

Mărime statistică	Expresie de calcul
Suport	$x \in (-\infty,\infty)$; $\nu \in (0,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\dfrac{\Gamma\left(\dfrac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\dfrac{\nu}{2}\right)}\left(1+\dfrac{t^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$\dfrac{1}{2}+x\Gamma\left(\dfrac{\nu+1}{2}\right)\cdot\sum_{n\geq 0}\dfrac{\left(-x^2/\nu\right)^n}{n!}\prod_{i=0}^{n-1}\dfrac{(1+2i)(\nu+1+2i)}{2(3+2i)}$
Media; mediana; moda; varianța	0 ($\nu > 1$); 0; 0; $\nu/(\nu-2)$, $\nu > 2$
Asimetria; excesul de boltire	0, $\nu > 3$; $6/(\nu-4)$, $\nu > 4$

Tabelul 8. Mărimi statistice ale distribuției continue Fisher-Snedecor F

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $d_1,d_2 \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$\dfrac{\Gamma\left((d_1+d_2)/2\right)}{\Gamma(d_1/2)\Gamma(d_2/2)}\dfrac{(d_1)^{d_1/2}(d_2)^{d_2/2}x^{d_1/2-1}}{(d_1x+d_2)^{(d_1+d_2)/2}}$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$IB\left(\dfrac{d_1x}{d_1x+d_2},\dfrac{d_1}{2},\dfrac{d_2}{2}\right)\Big/IB\left(1,\dfrac{d_1}{2},\dfrac{d_2}{2}\right)$, $IB(z,a,b)=\int_0^z t^{a-1}(1-t)^{b-1}dt$
Media; moda	$\dfrac{d_2}{d_2-2}$, $d_2 > 2$; $\dfrac{d_1-2}{d_1}\dfrac{d_2}{d_2+2}$, $d_1 > 2$
Varianța; asimetria	$\dfrac{2d_2{}^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$, $d_2 > 4$; $\dfrac{(2d_1+d_2-2)\sqrt{8(d_2-4)}}{(d_2-6)\sqrt{d_1(d_1+d_2-2)}}$, $d_2 > 6$

Excesul de boltire	$\frac{3d_2^{\,3}+(5d_1-8)d_2^{\,2}+(5d_1^{\,2}-32d_1+20)d_2-22d_1^{\,2}+44d_1-16}{d_1(d_2-6)(d_2-8)(d_1+d_2-2)/12}$, $d_2 > 8$

Tabelul 9. Mărimi statistice ale distribuției continue χ^2

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $d \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$(1/2)^{d/2}x^{d/2-1}e^{-x/2}/\Gamma(d/2)$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$\int_0^{x/2} t^{d/2-1}e^{-t}dt \Big/ \Gamma(d/2)$
Media; mediana; moda; varianța	d; $\cong d-2/3$; $d-2$, $d>2$; $2d$
asimetria; excesul de boltire	$\sqrt{8/d}$; $12/d$

Tabelul 10. Mărimi statistice ale distribuției continue exponențiale

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $\lambda \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$\lambda e^{-\lambda x}$
Funcția de repartiție	$1-e^{-\lambda x}$
Media; mediana; moda; varianța; asimetria; excesul de boltire	$1/\lambda$; $\ln(2)/\lambda$; 0; $1/\lambda^2$; 2; 6

Tabelul 11. Mărimi statistice ale distribuției continue Weibull

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; λ, $k \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate; funcția de repartiție	$kx^{k-1}e^{-(x/\lambda)^k}/\lambda^k$; $1-e^{-(x/\lambda)^k}$
Media; mediana; moda	$\mu=\lambda\Gamma(1+1/k)$; $\lambda(\ln(2))^{1/k}$; $\lambda((k-1)/k)^{1/k}$, $k>1$
Varianța; asimetria	$\sigma^2=\lambda^2\Gamma(1+2/k)-\mu^2$; $\gamma_1=\left(\Gamma(1+3/k)\lambda^3-3\mu\sigma^2-\mu^3\right)/\sigma^3$
Excesul de boltire	$\gamma_2=\left(\lambda^4\Gamma(1+4/k)-4\gamma_1\sigma^3\mu-6\mu^2\sigma^2-\mu^4\right)/\sigma^4$

Tabelul 12. Mărimi statistice ale distribuției continue Log-normale

Mărime statistică	Expresie de calcul
Suport	$x \in [0,\infty)$; $\mu \in (-\infty,\infty)$; $\sigma \in (0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}} \Big/ \left(x\sigma\sqrt{2\pi}\right)$
Funcția de repartiție	$\frac{1+\operatorname{erf}\left((\ln(x)-\mu)/(\sigma\sqrt{2})\right)}{2}$; $\operatorname{erf}(z)=2\int_0^z e^{-t^2}dt \Big/ \sqrt{\pi}$
Media; mediana; moda; varianța	$e^{\mu+\sigma^2/2}$; e^{μ} ; $e^{\mu-\sigma^2}$; $(e^{\sigma^2}-1)e^{2\mu+\sigma^2}$
Asimetria; excesul de boltire	$(e^{\sigma^2}+2)\sqrt{e^{\sigma^2}-1}$; $e^{4\sigma^2}+2e^{3\sigma^2}+3e^{2\sigma^2}-6$

Tabelul 13. Mărimi statistice ale distribuției continue Birnbaum-Saunders (a vieții obosite)

Mărime statistică	Expresie de calcul
Suport	$\mu, \beta, \gamma \in (0,\infty)$; $x \in (\mu,\infty)$
Minim; Maxim	μ; ∞
Funcția de probabilitate	$\frac{\sqrt{\frac{x-\mu}{\beta}}+\sqrt{\frac{\beta}{x-\mu}}}{2\gamma(x-\mu)} N_{0,1}\left(\left(\sqrt{\frac{x-\mu}{\beta}}-\sqrt{\frac{\beta}{x-\mu}}\right)\Big/\gamma\right)$
Funcția de probabilitate standard	$\frac{\sqrt{x}+\sqrt{1/x}}{2\gamma(x-\mu)} N_{0,1}\left((\sqrt{x}-\sqrt{1/x})/\gamma\right)$, $N_{0,1}(z)=\int_{-\infty}^{z}\frac{e^{-t^2/2}}{\sqrt{2\pi}}dt$
Funcția de repartiție standard	$N_{0,1}\left((\sqrt{x}-\sqrt{1/x})/\gamma\right)$
Media; varianța (standard)	$1+\gamma^2/2$; $\gamma\sqrt{1+5\gamma^2/4}$

Tabelul 14. Mărimi statistice ale distribuției continue Gamma

Mărime statistică	Expresie de calcul
Suport	$k, \theta \in (0,\infty)$; $x \in [0,\infty)$
Minim; Maxim	0; ∞
Funcția de probabilitate	$x^{k-1}e^{-x/\theta}\theta^{-k}/\Gamma(k)$, $\Gamma(z)=\int_0^\infty t^{z-1}e^{-t}dt$
Funcția de repartiție	$\int_0^{x/\theta} t^{k-1}e^{-t}dt \Big/ \int_0^\infty t^{k-1}e^{-t}dt$
Media; moda; varianța	$k\theta$; $(k-1)\theta$, $k > 1$; $k\theta^2$
Asimetria; excesul de boltire	$2/\sqrt{k}$; $6/k$

Tabelul 15. Mărimi statistice ale distribuției continue Laplace (dublu exponențială)

Mărime statistică	Expresie de calcul	
Suport	$b \in (0,\infty)$; $\mu, x \in (-\infty,\infty)$	
Minim; Maxim	$-\infty$; ∞	
Funcția de probabilitate	$e^{-\lvert x-\mu\rvert/b}/2b$	
Funcția de repartiție	$e^{(x-\mu)/b}/2$, $x < \mu$	$1-e^{-(x-\mu)/b}/2$, $\mu \le x$
Media; mediana; moda; varianța	μ; μ; μ; $2b^2$	
Asimetria; excesul de boltire	0; 3	

Tabelul 16. Mărimi statistice ale distribuției continue Gumbel (log-Weibull)

Mărime statistică	Expresie de calcul
Suport	$\beta \in (0,\infty)$; $\mu, x \in (-\infty,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\exp(-\exp(-(x-\mu)/\beta))\exp(-(x-\mu)/\beta)/\beta$
Funcția de repartiție	$\exp(-\exp(-(x-\mu)/\beta))$
Media; mediana; moda; varianța	$\mu+\beta\gamma$; $\mu-\beta\cdot\ln(\ln(2))$; μ; $\pi^2\beta^2/6$
Asimetria; excesul de boltire	$\frac{12\sqrt{6}\zeta(3)}{\pi^3}\cong 1.14$; 12/5

Tabelul 17. Mărimi statistice ale distribuției continue Beta

Mărime statistică	Expresie de calcul
Suport	$\alpha, \beta \in (0,\infty)$; $x \in [0,1]$
Minim; Maxim	0; 1
Funcția de probabilitate	$x^{\alpha-1}(1-x)^{\beta-1} / IB(1,\alpha,\beta)$; $IB(z,a,b) = \int_0^z t^{a-1}(1-t)^{b-1}dt$
Funcția de repartiție	$IB(x,\alpha,\beta) / IB(1,\alpha,\beta)$
Media; moda; varianța	$\frac{\alpha}{\alpha+\beta}$; $\frac{\alpha-1}{\alpha+\beta-2}$, $\alpha, \beta > 1$; $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$
Asimetria; excesul de boltire	$\frac{2(\beta-\alpha)\sqrt{\alpha+\beta+1}}{(\alpha+\beta+2)\sqrt{\alpha\beta}}$; $\frac{\alpha^3-(2\beta-1)\alpha^2-2\alpha\beta(\beta+2)+(\beta+1)\beta^2}{\alpha\beta(\alpha+\beta+2)(\alpha+\beta+3)/6}$

Tabelul 18. Mărimi statistice ale distribuției continue Gauss (normale)

Mărime statistică	Expresie de calcul
Suport	$\sigma \in (0,\infty)$; $\mu, x \in (-\infty,\infty)$
Minim; Maxim	$-\infty$; ∞
Funcția de probabilitate	$\exp\left(-((x-\mu)/\sigma)^2/2\right) / \left(\sigma\sqrt{2\pi}\right)$
Funcția de repartiție	$\left(1+\text{erf}\left((x-\mu)/\left(\sigma\sqrt{2}\right)\right)\right)/2$; $\text{erf}(z) = 2\int_0^z e^{-t^2}dt \Big/ \sqrt{\pi}$
Media; moda; varianța	μ; μ; μ; σ^2
Asimetria; excesul de boltire	0; 0

Tabelul 19. Alte mărimi statistice ale distribuției continue Gauss (normale)

Mărime	Populație (finită) de volum n_X	Eșantion de volum n_Y	Estimator
Media	$\mu_{\bar{X}} = \mu$; $\sigma_{\bar{X}}^2 = \sigma^2/n_X$	$\mu_{\bar{Y}} = m$; $\sigma_{\bar{Y}}^2 = s^2/n_Y$	m; $s^2/(n_Y-1)$
Varianța	$(n_X-1)\sigma^2/n_X$ $\frac{(n_X-1)^2}{n_X^3\mu_4^{-1}} - \frac{(n_X-1)\mu_2^2}{n_X^3(n_X-3)^{-1}}$	$(n_Y-1)s^2/n_Y$ $\frac{(n_Y-1)^2}{n_Y^3 m_4^{-1}} - \frac{(n_Y-1)m_2^2}{n_Y^3(n_Y-3)^{-1}}$	s^2 $\frac{(n_Y-1)m_4}{n_Y^2} - \frac{(n_Y-3)m_2^2}{n_Y(n_Y-1)}$ $= \frac{2\sigma^4(n_Y-1)}{n_Y^2} \cong \frac{2s^4}{n_Y-1}$
Var γ_1	$\frac{6n_X(n_X-1)}{(n_X-2)(n_X+1)(n_X+3)}$	$\frac{6n_Y(n_Y-1)}{(n_Y-2)(n_Y+1)(n_Y+3)}$	$c_4^2 \cdot \text{var}(g_1)$ c_4 - Vezi Tabelul 29
Var γ_2	$\frac{24n_X(n_X-1)^2(n_X-3)^{-1}}{(n_X-2)(n_X+3)(n_X+5)}$	$\frac{24n_Y(n_Y-1)^2(n_Y-3)^{-1}}{(n_Y-2)(n_Y+3)(n_Y+5)}$	$c_4^2 \cdot \text{var}(g_2)$ c_4 - Vezi Tabelul 29

3.3 Statistici

Statistica Benford

Testul Benford foloseşte distribuția Z (normală) pentru a verifica ipoteza că un şir de numere urmează distribuția Benford, frecvențele după care se distribuie o anume cifră a fiecărui număr din şir.

Un şir de numere urmează distribuția Benford dacă probabilitatea de distribuție a unei cifre (d_i) a numerelor ($d=d_0d_1\ldots$) reprezentate în baza de numerație b (uzual baza 10) urmează legea (Benford):

$p(d_0) = \log_b\left(1+\frac{1}{d_0}\right)$, $d_0 = 1..(b-1)$; $p(d_1) = \sum_{k=1}^{b-1}\log_b\left(1+\frac{1}{k\cdot b+d_1}\right)$, $d_1 = 0..(b-1)$ $p(d_2) = \sum_{j=1}^{b-1}\sum_{k=0}^{b-1}\log_b\left(1+\frac{1}{j\cdot b^2+k\cdot b+d_2}\right)$, $d_2 = 0..(b-1)$ …	(Benford)

Ipoteza acestei legi de distribuție este că valorile măsurătorilor rezultate din observație sunt frecvent distribuite logaritmic şi astfel logaritmul setului de măsurători este distribuit uniform. Legea de distribuție este numită după fizicianul Frank BENFORD care a formulat-o intuitiv în 1938 [2], dar demonstrația acesteia a fost dată mult mai târziu [3].

Acest rezultat intuitiv de numărare a aparițiilor a fost găsit aplicându-se la o mare varietate de seturi de date incluzând facturile la electricitate, adresele de străzi, prețurile acțiunilor, numerele populației, ratele de deces, lungimile râurilor, constante fizice şi matematice şi procesele descrise de legi putere (care sunt foarte comune în natură). Este foarte important de ştiut că rezultatul (odată observat într-o bază de numerație) are loc independent de baza de numerație în care se exprimă numerele, chiar dacă proporțiile de reprezentare se schimbă. De aici, ***acest rezultat poate fi folosit pentru a verifica datele în suspiciunea de alterare (mistificare) a acestora prin compararea frecvențelor teoretice cu cele observate pentru prima cifră a acestora***.

[2] BENFORD Frank. 1938. The law of anomalous numbers. Proceedings of the American Philosophical Society 78(4):551-572.
[3] HILL Theodore P. 1995. Base invariance implies Benford's Law. Proceedings of the American Mathematical Society 123(3):887-895.

Statistica Jarque-Bera

Testul Jarque-Bera [4,5] calculează şi atribuie probabilitatea statistică ca valorile unui eşantion ce provine din populație normal distribuită să îşi abată simultan asimetria şi excesul de boltire de la valorile teoretice corespunzătoare distribuției normale.

Statistica Jarque-Bera se calculează cu relația:

$$JB = n\frac{g_1^{\ 2} + g_2^{\ 2}/4}{6}$$

în care g_1 este asimetria, g_2 este excesul de boltire şi n este volumul eşantionului.

Statistica JB are o distribuție asimptotică către χ^2(df=2).

g_1, Asimetria observabilei Y	Un eşantion	$g_1 = m_3/m_2^{3/2}$	$m_k = E_k(Y)$, k>1
b_2, Boltirea observabilei Y		$b_2 = m_4/m_2^{\ 2}$	
g_2, Excesul de boltire al observabilei Y		$g_2 = b_2$-3	

Statistica Kolmogorov-Smirnov

Testul Kolmogorov-Smirnov [6] poate fi folosit pentru verificarea ipotezei că un eşantion de date urmează o anumită lege de distribuție (redat în continuare), precum şi pentru compararea legilor de distribuție ale populațiilor din care provin două eşantioane [7].

Statistica Kolmogorov-Smirnov verifică dacă observațiile independente $X=(X_i)_{1\le i\le n}$ provin dintr-o populație ce urmează legea de distribuție dată de funcția cumulativă de probabilitate $F_t(x)$ prin calcularea maximumului diferenței absolute între $F_t(x)$ şi funcția cumulativă de probabilitate observată $F_o(x)$ în toate punctele observației:

$D = \max_{1\le i\le n}\left\|F_t(X_i) - F_o(X_i)\right\|$	(K-S Stat)

Distribuția Kolmogorov

Legea de distribuție Kolmogorov se obține pentru variabila aleatoare K dată de:

$K = \max_{0\le t\le 1}\left\|B(t)\right\|$, unde B este puntea Browniană condiționată de: B(0) = B(1) = 0 M(B(t)) = 0 Var(B(t)) = t(t-1) $P(K \le x) = 1 - 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2x^2} = \frac{\sqrt{2\pi}}{x}\sum_{i=1}^{\infty}e^{-(2i-1)^2\pi^2/8x^2}$	(K-S Dist)

[4] Carlos M JARQUE, Anil K BERA. 1980. Efficient tests for normality, homoscedasticity and serial independence of regression residuals. *Econ Lett* 6(3):255-259.

[5] Carlos M JARQUE, Anil K BERA. 1981. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. *Econ Lett* 7(4):313-318.

[6] KOLMOGOROV Andrey. 1941. Confidence Limits for an Unknown Distribution Function. The Annals of Mathematical Statistics 12(4):461-463.

[7] SMIRNOV Nikolay V. 1948. Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics 19(2):279-281.

Testul Kolmogorov-Smirnov

Ipoteza testului este că următoarea convergență are loc în distribuție:

$D\sqrt{n} \underset{n\to\infty}{\to} \sup_{t\in[0,1]} \lvert B(F(t)) \rvert$ Ipoteza se respinge la nivelul de semnificație dacă: $D\sqrt{n} > K_\alpha$, unde K_α: $P(K \le K_\alpha) = 1-\alpha$	(K-S Test)

Pentru compararea a două distribuții observate:

$D = \max_{1\le i\le \max(n,m)} \lvert F_{o1}(X_i) - F_{o2}(X_i) \rvert$ Ipoteza se respinge la nivelul de semnificație dacă: $D\sqrt{\frac{mn}{m+n}} > K_\alpha$	(K-S Test)

Statistica Anderson-Darling

Testul Anderson-Darling [8] verifică dacă este o evidență statistică ca un eşantion să nu provină dintr-o funcție de probabilitate dată.

Statistica Anderson verifică dacă asupra observațiilor distincte ordonate crescător $(X_i)_{1\le i\le n}$, $X_i<X_{i+1}$ se poate respinge ipoteza că provin dintr-o distribuție dată de funcția cumulativă de probabilitate F calculând valoarea A dată de relația:

$$A^2 = -n - \sum_{k=1}^{n} \frac{2k-1}{n}\left(\ln(F(Y_k) + \ln(1 - F(Y_{n+1-k}))\right)$$

O aplicație de interes însă o reprezintă testul Anderson-Darling pentru mai multe eşantioane asupra cărora se poate verifica proveniența din aceeaşi populație, caz în care legea de distribuție a populației nu mai trebuie să fie specificată [9,10]. Formulele de calcul şi interpretarea testului pentru compararea de eşantioane se găsesc la adresa [11].

În cazul comparației unei legi de distribuție discrete cunoscute cu legea de distribuție observată în eşantion varianța statisticii A^2 se calculează cu formula ([12], n - numărul de observații din eşantion, π=3.14159265358979323846264338...):

$$Var(A^2) = \frac{2(\pi^2 - 9)}{3} + \frac{10 - \pi^2}{n}$$

---

[8] Theodore W ANDERSON, Donald A DARLING. 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics 23(2):193-212.

[9] Fritz W SCHOLZ, Michael A STEPHENS. 1987. K-sample Anderson-Darling Tests. Journal of the American Statistical Association 82(399):918-924.

[10] Department of Defense Handbook. 2002. Composite Materials Handbook. Volume 1. Polymer Matrix Composites Guidelines for Characterization of Structural Materials. Chapter 8. Statistical Methods. 8.3.2.2 The k-sample Anderson-Darling test MIL-HDBK-17-1F:8-17.

[11] Lorentz JÄNTSCHI. 2009. http://l.academicdirect.org/Statistics/tests/kAD/, k-sample Anderson-Darling.

[12] Fritz W SCHOLZ, Michael A STEPHENS. 1986. K-Sample Anderson-Darling Tests of Fit, for Continuous and Discrete Cases. Technical Report. University of Washington. GN-22:81.

În cazul verificării ipotezei de normalitate, este posibil să se aproximeze probabilitatea de observație asociată valorii statisticii A^2 [13]. Se aplică corecția de volum al eşantionului:

$$A^2{}_c = A^2(1 + 0.75/n + 2.25/n^2)$$

$$p = \begin{cases} 1 - \exp\left(-13.436 + 101.14 \cdot x - 223.73 \cdot x^2\right), & x < 0.2 \\ 1 - \exp\left(-8.318 + 42.796 \cdot x - 59.938 \cdot x^2\right), & 0.2 \le x < 0.34 \\ \exp\left(0.9177 - 4.279 \cdot x - 1.38 \cdot x^2\right), & 0.34 \le x < 0.6 \\ \exp\left(1.2937 - 5.709 \cdot x + 0.0186 \cdot x^2\right), & x \le 0.6 \end{cases}$$

Statistica Pearson-Fisher Chi Square

Experimentul variaţelor ce conduce la distribuţia χ^2

Distribuţia χ^2 a fost descoperită de Karl PEARSON [14] în urma încercării de a explica varianţa observată a numerelor care provin din distribuţia normală.

Astfel, dacă se consideră distribuţia normală standard N(0,1) şi variabila întâmplătoare X ce urmează această distribuţie (Figura 1), probabilitatea (dp) de a extrage valorile -x şi x din N(0,1) sunt ambele egale şi egale cu diferenţiala funcţiei de densitate de probabilitate a distribuţiei normale ($PDF_{N(0,1)}$).

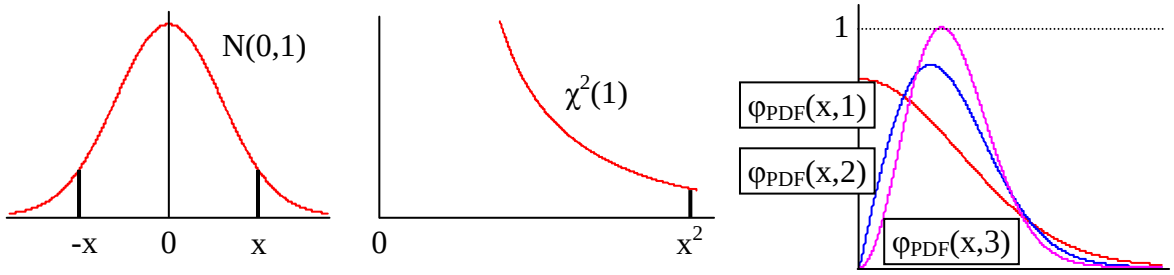


Figura 1. Funcţiile de densitate de probabilitate (PDFs) pentru N(0,1), $\chi^2(1)$ şi φ(k)

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

Distribuţia normală standard are media 0; astfel, pentru a exprima probabilitatea de observaţie pentru deviaţia x^2 trebuie adunate două probabilităţi (pentru -x şi x) date de relaţia (1):

$$dp(x^2) = 2 \cdot dPDF_{N(0,1)}(x) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \qquad (2)$$

Pentru a reconstitui PDF pentru x^2 trebuie să efectuăm o schimbare de variabilă $x^2 = t$;

[13] A. Trujillo-Ortiz, R. Hernandez-Walls, K. Barba-Rojo, A. Castro-Perez. 2007. AnDartest:Anderson-Darling test for assessing normality of a sample data. http://mathworks.com/matlabcentral/fileexchange/14807

[14] PEARSON Karl. 1900. 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 5th Ser 50:157-175.

atunci x = √t şi:

$$dp(t) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t}{2}\right) d\sqrt{t} = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{t}{2}\right) \frac{1}{2\sqrt{t}} dt \quad (3)$$

Este uşor de verificat că (3) este un caz particular al lui (4) când k = 1:

$$\chi^2_{PDF}(t,k) = \frac{1}{2^{k/2}\Gamma(k/2)} t^{k/2-1} \exp\left(-\frac{t}{2}\right) \quad (4)$$

Procedura descrisă mai sus corespunde pentru distribuția Chi Square cu un grad de libertate (extragerea lui X din distribuția normală). Dacă sunt extrase mai multe valori (k valori) din distribuția normală atunci se obține distribuția Chi Square cu k grade de libertate, şi demonstrația că ecuația (4) este adevărată poate fi găsită în [15].

Calea directă de la distribuția normală la distribuția χ^2 nu este reversibilă (Figura 1); astfel, definind variabila φ ca în relația (5) - ce reprezintă o expresie modificată a coeficientului de asociere definit de LIEBETRAU [16]:

$$\varphi = \varphi(X^2, k) = \sqrt{\frac{X^2}{k}} \quad (5)$$

obținerea distribuției lui φ se poate obține pe o cale similară cu cea descrisă mai sus; notând u=φ în (5) şi substituind $t = X^2 = ku^2$ în (4) se obține ($du^2 = 2u \cdot du$):

$$d\chi^2(ku^2, k) = \frac{1}{2^{k/2}\Gamma(k/2)} (ku^2)^{k/2-1} \exp\left(-\frac{ku^2}{2}\right) d(ku^2) \quad (6)$$

După rearanjarea termenilor:

$$d\varphi_{PDF}(u,k) = \frac{2u^{k-1}}{\Gamma(k/2)} \left(\frac{k}{2}\right)^{k/2} \exp\left(-\frac{u^2}{2/k}\right) du \quad (7)$$

Pornind de la densitatea de probabilitate (PDF) a distribuției Gamma:

$$\Gamma_{PDF}(x;a,b,c) = \frac{cx^{ca-1}}{b^{ca}\Gamma(a)} \exp\left(-(x/b)^c\right) \quad (8)$$

este uşor de verificat că:

$$\varphi_{PDF}(x,k) = \Gamma_{PDF}\left(x, \frac{k}{2}, \sqrt{\frac{2}{k}}, 2\right) \quad (9)$$

Relația (9) demonstrează că distribuția lui $\sqrt{\frac{X^2}{k}}$ este un caz particular al distribuției Gamma (Figura 1).

---

[15] FISHER Ronald A. 1935. The Mathematical Distributions Used in the Common Tests of Significance. Econometrica 3:353-365.

[16] LIEBETRAU Albert M. 1983. Measures of association. Newbury Park, CA: Sage Publications. Quantitative Applications in the Social Sciences 32:1-96 (p.13).

Testul χ^2 ca măsură a independenței, omogenității şi asocierii în distribuție

Distribuția χ^2 are 3 aplicații imediate:

÷ Testul Chi Square pentru verificarea independenței
  - o testează asocierea între două variabile cu valori grupate pe categorii;
  - o se poate aplica dacă au loc două condiții:
    - nici una din valorile aşteptate nu este mai mică decât 1;
    - nu mai mult de 20% din valorile aşteptate nu sunt mai mici de 5;
  - o ipotezele de lucru sunt: nu există nici o asociere între cele două variabile (ipoteza nulă) şi este o asociere între cele două variabile (ipoteza contrară);
  - o Când statistica Chi Square (X^2) este mai mare decâ valoarea funcției cumulative de probabilitate a distribuției Chi Square (χ^2) pentru numărul de grade de libaertate egal cu numărul de cazuri minus unu şi pentru riscul de a fi în eroare (nivelul de semnificație) ales, atunci există o diferență semnificativă de la ipoteza lipsei de asociere şi cele două variabile sunt asociate;

÷ Testul Chi Square pentru verificarea omogenității
  - o testează dacă mai multe populații sunt similare (sau omogene sau egale) în anumite caracteristici (acele caracteristici care sunt incluse în testare);
  - o ipotezele de lucru sunt: populațiile sunt similare (sau omogene sau egale) în caracteristica supusă observației (ipoteza nulă) şi populațiile sunt diferite în caracteristică (ipoteza contrară);
  - o uzual caracteristica supusă observației este un moment central (ex. valoare medie, varianță);

÷ Testul Chi Square pentru verificarea asocierii în distribuție
  - o testează dacă un model teoretic poate fi asociat observațiilor;
  - o ipotezele de lucru sunt: datele observate urmează distribuția dată de modelul teoretic (ipoteza nulă) şi datele observate nu provin dintr-o populație ce urmează modelul teoretic (ipoteza contrară);

Probleme frecvent întâlnite în aplicarea testului χ^2 ca măsură a asocierii în distribuție

Testul χ^2, propus ca măsură a depărtării întâmplătoare între observație şi modelul teoretic de Karl PEARSON [14] a fost corectat în interpretare de Ronald FISHER prin reducerea numărului de grade de libertate corespunzător cu o unitate (datorită estimării frecvenței teoretice din frecvența observată, [17]), şi cu numărul parametrilor necunoscuți ai

[17] FISHER Ronald A. 1922. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. Journal of the Royal Statistical Society 85:87-94.

distribuției teoretice estimați din observații din măsuri ale tendinței centrale ([18]).

Testarea agrementului între observație şi ipoteză se realizează prin divizarea observațiilor într-un număr definit de intervale (n), pentru care se calculează expresia X^2 (unde s este numărul de parametrii ai distribuției teoretice estimați din momente centrale, O_i este frecvența experimentală observată în clasa de frecvență i, E_i este frecvența aşteptată calculată din legea de distribuție teoretică pentru clasa de frecvență i, X^2 este valoarea statisticii chi square iar χ^2 este valoarea parametrului statistic chi square din distribuția cu acelaşi nume):

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \approx \chi^2(n-s-1) \qquad (10)$$

Pe baza distribuției teoretice χ^2 se calculează probabilitatea de respingere a ipotezei de agrement. Uzual ipoteza de agrement este acceptată dacă probabilitatea de respingere a ipotezei de agrement ($\chi^2_{CDF}(X^2,n\text{-}s\text{-}1)$) este mai mică de 5%.

În ciuda faptului că testul χ^2 este cea mai cunoscută statistică pentru verificarea agrementului între observație şi ipoteză, testarea independenței şi a omogenității, definirea cadrului de aplicare al acesteia este dintre cele mai complexe [19].

O serie de probleme la compararea unei distribuții observate cu o distribuție teoretică apar în calcularea statisticii X^2 şi în aplicarea testului χ^2.

O primă problemă este alegerea numărului de clase de frecvență şi există mai multe soluții, dintre care două sunt:

÷ calcularea prin rotunjire a numărului de clase de frecvență din entropia Hartley [20] a observației vs. expectație: $\log_2(2N)$, unde N este numărul de observații (EasyFit [21] foloseşte această procedură);

÷ calcularea numărului de clase de frecvență odată cu lărgimea clasei folosind histograma ca estimator al densității [22] şi alegerea pe baza acesteia a criteriului optimal pentru lărgimea clasei (Dataplot [23] generează automat clasele de frecvență folosind această regulă: lărgimea clasei de frecvență este 0.3·s unde s este deviația standard a eşantionului; limitele inferioară şi superioară sunt date de medie ±6·s şi clasele de frecvență observată 0 marginale sunt omise;

---

[18] FISHER Ronald A. 1924. The Conditions Under Which χ^2 Measures the Discrepancy Between Observation and Hypothesis. Journal of the Royal Statistical Society 87:442-450.
[19] SNEDECOR George W. and COCHRAN William G. 1989. Statistical Methods, Eighth Edition, Iowa State University Press.
[20] HARTLEY Ralph V L. 1928. Transmission of Information. Bell System Technical Journal 1928:535-563.
[21] Software. 2008. EasyFit v.5. MathWave Technologies. http://www.mathwave.com
[22] SCOTT David. 1992. Multivariate Density Estimation. John Wiley, Chapter 3.
[23] Software. 2005. Dataplot. National Institute for Standards and Technology. http://www.itl.nist.gov/div898/software/dataplot.html

O a doua problemă este lărgimea claselor de frecvență; şi aici există cel puțin două abordări:

÷ Datele pot fi grupate în clase de frecvență de probabilitate (teoretică sau observată) egală;

÷ Datele pot fi grupate în intervale de lărgime egală;

Prima abordare (probabilitatea egală) este mai frecvent adoptată deoarece este o soluție mai bună pentru observații foarte grupate.

O altă problemă este numărul de observații din interiorul fiecărei clase de frecvență. Fiecare clasă de frecvență trebuie să conțină cel puțin 5 observații, astfel încât în practică clase de frecvență alăturate se reunesc pentru a satisface această impunere.

Probleme frecvent întâlnite în aplicarea testului χ^2 ca măsură a omogenității

Statistica Chi Square operează în ipoteza în care o observabilă este rezultatul suprapunerii a doi (sau mai mulți, dintre care pentru doi dintre aceştia se realizează un experiment) factori. În acest caz se constituie un experiment menit să verifice dacă se poate accepta independența între aceşti doi factori. Se construieşte un tabel de contingență format din linii (reprezentând valorile primului factor) şi coloane (reprezentând valorile celui de-al doilea factor) în care se cumulează frecvențele sau valorile medii ale variabilei observate şi în care ipoteza independenței factorilor se translatează în ipoteza omogenității valorilor înregistrate în tabel.

Valoarea statisticii X^2 se calculează cu formula (unde $1\le i\le r$ reprezintă indicii observațiilor asociate primului factor, $1\le j\le c$ reprezintă indicii observațiilor asociate celui de-al doilea factor, $O_{i,j}$ reprezintă valori medii (pentru testul de omogenitate) sau frecvențe (pentru testul de independență) observate pentru perechea (i,j) de valori ale factorilor, $E_{i,j}$ este valoarea medie (pentru testul de omogenitate) sau frecvența (pentru testul de independență) aşteptată pentru perechea (i,j) de valori ale factorilor, X^2 este valoarea statisticii chi square iar χ^2 este valoarea parametrului statistic chi square din distribuția cu acelaşi nume):

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}} \approx \chi^2((r-1)(c-1)) \qquad (11)$$

Testarea individuală a omogenității valorilor dintr-o clasă (linie sau coloană în tabel) şi în acelaşi timp crearea unei ierarhii a iregularităților se obține descompunând expresia lui X^2 în:

$$X^2{}_c = \sum_{i=1}^{r}\frac{(O_{i,c}-E_{i,c})^2}{E_{i,c}} \approx \chi^2(r-1);\ X^2{}_r = \sum_{j=1}^{c}\frac{(O_{r,j}-E_{r,j})^2}{E_{r,j}} \approx \chi^2(c-1) \qquad (12)$$

Presupunerea naturală este că observațiile $O_{i,j}$ sunt rezultatul multiplicării celor doi factori, ceea ce face ca observațiile repetate să aproximeze tot mai bine efectul de

multiplicare, şi de aici rezultă o formulă de exprimare pentru frecvența aşteptată $E_{i,j}$ [24]:

$$E_{i,j} = \frac{\sum_{k=1}^{r} O_{i,k} \sum_{k=1}^{c} O_{k,j}}{\sum_{i=1}^{r} \sum_{j=1}^{c} O_{i,j}} \tag{13}$$

În acelaşi cadru al presupunerii naturale al efectului multiplicativ al celor doi factori asupra observabilei O din punct de vedere matematic se pot formula trei presupuneri cu privire la eroarea pătratică $(O_{i,j}-E_{i,j})^2$ produsă de observație:

÷ măsurătoarea este afectată de erori absolute întâmplătoare;

÷ măsurătoarea este afectată de erori relative întâmplătoare;

÷ măsurătoarea este afectată de erori întâmplătoare pe o scară intermediară între erori absolute şi erori relative;

Prima dintre ipoteze (erori absolute întâmplătoare) conduce din punct de vedere matematic la minimizarea varianței între model şi observație (relația 14), a doua dintre ipoteze conduce la minimizarea pătratului coeficientului de variație (relația 15) iar o soluție (una din mai multe soluții posibile) la cea de-a treia dintre ipoteze o reprezintă minimizarea statisticii X^2 (relația 16).

$$S^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} (O_{i,j} - a_i b_j)^2 = \min. \tag{14}$$

$$CV^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - a_i b_j)^2}{(a_i b_j)^2} = \min. \tag{15}$$

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{i,j} - a_i b_j)^2}{a_i b_j} = \min. \tag{16}$$

În relațiile (14)-(16) apar exprimați cei doi factori a căror independență se verifică prin intermediul efectului multiplicativ (a_i, $1 \le i \le r$ reprezintă contribuția primului factor la valoarea aşteptată $E_{i,j}$ iar b_j, $1 \le j \le c$ reprezintă contribuția celui de-al doilea factor la valoarea aşteptată $E_{i,j}$ şi expresia valorii aşteptate $E_{i,j}$ este dată, aşa cum presupunerea naturală a fost făcută de produsul celor două contribuții: $E_{i,j}=a_i \cdot b_j$).

Minimizarea cantităților date de relațiile (14)-(16) în scopul determinării contribuțiilor factorilor A ($A=(a_i)_{1 \le i \le r}$) şi B ($B=(b_j)_{1 \le j \le c}$) se face pe aceeaşi cale, dată generic de relația (17):

$$\left(\frac{\partial \cdot (a_i, b_j)}{\partial a_i} = 0 \right)_{1 \le i \le r} ; \quad \left(\frac{\partial \cdot (a_i, b_j)}{\partial b_j} = 0 \right)_{1 \le j \le c} \tag{17}$$

unde expresia de derivat $\cdot(a_i,b_j)$ este una din expresiile S^2, CV^2 şi X^2 date de relațiile (14)-

[24] FISHER Ronald A. 1923. Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties. Journal of Agricultural Science 13:311-320.

(16). În urma calculului se obține că relația (14) este verificată de acele valori $(a_i)_{1\le i\le r}$ şi $(b_i)_{1\le j\le c}$ care verifică de asemenea relația (18), relația (15) este verificată de acele valori $(a_i)_{1\le i\le r}$ şi $(b_i)_{1\le j\le c}$ care verifică de asemenea relația (19), iar relația (16) este verificată de acele valori $(a_i)_{1\le i\le r}$ şi $(b_i)_{1\le j\le c}$ care verifică de asemenea relația (20):

$$a_i = \frac{\sum_{j=1}^{c} b_j O_{i,j}}{\sum_{j=1}^{c} b_j^{\,2}},\ i=1..r;\quad b_j = \frac{\sum_{i=1}^{r} a_i O_{i,j}}{\sum_{i=1}^{r} a_i^{\,2}},\ j=1..c \tag{18}$$

$$a_i = \frac{\sum_{j=1}^{c} \frac{O_{i,j}^{\,2}}{b_j^{\,2}}}{\sum_{j=1}^{c} \frac{O_{i,j}}{b_j}},\ i=1..r;\ b_j = \frac{\sum_{i=1}^{r} \frac{O_{i,j}^{\,2}}{a_i^{\,2}}}{\sum_{i=1}^{r} \frac{O_{i,j}}{a_i}},\ j=1..c \tag{19}$$

$$a_i^{\,2} = \frac{\sum_{j=1}^{c} \frac{O_{i,j}^{\,2}}{b_j}}{\sum_{j=1}^{c} b_j},\ i=1..r;\ b_j^{\,2} = \frac{\sum_{i=1}^{r} \frac{O_{i,j}^{\,2}}{a_i}}{\sum_{i=1}^{r} a_i},\ j=1..c \tag{20}$$

Se poate de asemenea arăta matematic că relațiile (18)-(20) admit o infinitate de soluții şi că familiile de soluții ale relațiilor (18)-(20) se află în vecinătatea familiei de soluții date de relația (13), re-scrisă aici ca relația (21), exprimând explicit cei doi factori A şi B:

$$a_i \cdot b_j = \frac{\sum_{k=1}^{r} O_{i,k} \sum_{k=1}^{c} O_{k,j}}{\sum_{i=1}^{r}\sum_{j=1}^{c} O_{i,j}} \tag{21}$$

Calea directă de rezolvare a ecuațiilor (18)-(20) fără a face apel la ecuația (21) este ineficientă. De exemplu pentru r=2, c=3 substituțiile în relația (18) duc la:

$$\left(\frac{a_2}{a_1}\right)^2 + \frac{(O_{1,1}^{\,2} + O_{1,2}^{\,2} + O_{1,3}^{\,2}) - (O_{2,1}^{\,2} + O_{2,2}^{\,2} + O_{2,3}^{\,2})}{(O_{1,1}O_{2,1} + O_{1,2}O_{2,2} + O_{1,3}O_{2,3})}\left(\frac{a_2}{a_1}\right) - 1 = 0 \tag{22}$$

care este rezolvabilă în (a_2/a_1) care dovedeşte că există o infinitate de soluții (pentru orice valoare nenulă a lui a_1 există o valoare a_2 care să verifice ecuația 22) şi gradul ecuației (22) este dat de min(r,c). Ecuațiile ce se obțin pe calea substituției directe devin din ce în ce mai complicate cu creşterea lui `r` şi `c` şi cu coborârea dinspre relația (18) către relația (20). Astfel, de exemplu pentru acelaşi r=2 şi c=3 substituțiile în relația (20) conduc la:

$$O_{1,1}{}^2O_{1,2}{}^2(O_{1,1}{}^2-O_{1,2}{}^2)\left(\frac{a_2}{a_1}\right)^5+(O_{1,1}{}^4O_{2,2}{}^2-O_{1,2}{}^4O_{2,1}{}^2)\left(\frac{a_2}{a_1}\right)^4+$$
$$+2O_{1,1}{}^2O_{1,2}{}^2(O_{2,2}{}^2-O_{2,1}{}^2)\left(\frac{a_2}{a_1}\right)^3+2O_{2,1}{}^2O_{2,2}{}^2(O_{1,2}{}^2-O_{1,1}{}^2)\left(\frac{a_2}{a_1}\right)^2 \quad (23)$$
$$+(O_{1,2}{}^2O_{2,1}{}^4-O_{1,1}{}^2O_{2,2}{}^4)\left(\frac{a_2}{a_1}\right)+O_{2,2}{}^2O_{2,1}{}^2(O_{2,1}{}^2-O_{2,2}{}^2)=0$$

care este o ecuație de gradul 5 (r+c).

Calea indirectă de rezolvare a relațiilor (18)-(20) este prin aproximații succesive făcând apel la soluția aproximativă oferită de (21). Astfel, se folosește relația (21) pentru a obține prima aproximație (aproximația inițială) a soluției după care în fiecare succesiune de aproximații se înlocuiesc vechile valori ale aproximației în partea dreaptă a relațiilor (18)-(20) pentru a obține noile aproximații.

Metoda aproximațiilor succesive converge rapid către soluția optimală. Astfel pentru relația (18) trei iterații sunt suficiente pentru a obține (vezi Tabelul 1) o valoare reziduală de 282.11735 şi de la această iterație încolo valoarea reziduală îşi schimbă cifrele dincolo de a 5-a zecimală, în timp ce pentru relația (20) aceeaşi calitate a reprezentării soluției optimale este obținută după 4 iterații.

Folosind datele din [24] redate în Tabelul 1, valorile sugerate de ecuațiile (21) pentru produsele $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ sunt redate în Tabelul 2, valorile ce rezultă după rezolvarea iterativă a relațiilor (18)-(20) sunt redate în Tabelele 3-5, în timp ce Tabelul 6 centralizează rezultatele obținute pe cele 4 căi.

Tabelul 1. Valori experimentale în tratamentul cartofilor

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY	Suma
DS	25.3	28	23.3	20	22.9	20.8	22.3	21.9	18.3	14.7	13.8	10	**241.3**
DC	26	27	24.4	19	20.6	24.4	16.8	20.9	20.3	15.6	11	11.8	**237.8**
DB	26.5	23.8	14.2	20	20.1	21.8	21.7	20.6	16	14.3	11.1	13.3	**223.4**
US	23	20.4	18.2	20.2	15.8	15.8	12.7	12.8	11.8	12.5	12.5	8.2	**183.9**
UC	18.5	17	20.8	18.1	17.5	14.4	19.6	13.7	13	12	12.7	8.3	**185.6**
UB	9.5	6.5	4.9	7.7	4.4	2.3	4.2	6.6	1.6	2.2	2.2	1.6	**53.7**
Suma	**128.8**	**122.7**	**105.8**	**105**	**101.3**	**99.5**	**97.3**	**96.5**	**81**	**71.3**	**63.3**	**53.2**	**1125.7**

Legendă:
- ÷ T_V: Tratament vs. Varietate
- ÷ UD, KK, KP, TP, ID, GS, AJ, BQ, ND, EP, AC, DY: varietăți de cartofi (UD: Up to Date; KK: K of K; KP: Kerr's Pink; TP: Tinwald Perfection; ID: Iron Duke; GS: Great Scott; AJ: Ajax; BQ: British Queen; ND: Nithsdale; EP: Epicure; AC: Arran Comrade; DY: Duke of York)
- ÷ DS, DC, DB, US, UC, UB: tratamente (D* - cu fertilizant natural; U* - fără; S - sol fertilizat cu sulfat; C - sol fertilizat cu cloruri; B - sol fertilizat cu baze)

Tabelul 2. Valorile produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ calculate cu relația (21)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.61	26.30	22.68	22.51	21.71	21.33	20.86	20.69	17.36	15.28	13.57	11.40
DC	27.21	25.92	22.35	22.18	21.40	21.02	20.55	20.39	17.11	15.06	13.37	11.24
DB	25.56	24.35	21.00	20.84	20.10	19.75	19.31	19.15	16.07	14.15	12.56	10.56
US	21.04	20.04	17.28	17.15	16.55	16.25	15.90	15.76	13.23	11.65	10.34	8.69
UC	21.24	20.23	17.44	17.31	16.70	16.41	16.04	15.91	13.35	11.76	10.44	8.77
UB	6.14	5.85	5.05	5.01	4.83	4.75	4.64	4.60	3.86	3.40	3.02	2.54

Tabelul 3. Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (18)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.07	26.42	22.64	21.85	21.85	21.94	20.94	20.63	17.93	15.48	13.54	11.61
DC	26.66	26.02	22.29	21.52	21.52	21.60	20.62	20.32	17.66	15.24	13.33	11.43
DB	24.91	24.32	20.83	20.11	20.11	20.19	19.27	18.99	16.50	14.25	12.46	10.69
US	20.64	20.15	17.26	16.66	16.66	16.73	15.96	15.73	13.67	11.80	10.32	8.85
UC	20.58	20.09	17.21	16.61	16.61	16.68	15.92	15.69	13.63	11.77	10.29	8.83
UB	6.29	6.14	5.26	5.08	5.08	5.10	4.86	4.79	4.17	3.60	3.14	2.70

Tabelul 4. Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (19)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.57	26.08	23.04	22.61	21.48	21.61	21.13	20.69	17.66	15.23	13.79	11.56
DC	27.38	25.9	22.88	22.45	21.34	21.46	20.99	20.55	17.54	15.13	13.69	11.48
DB	25.84	24.44	21.59	21.19	20.14	20.26	19.8	19.4	16.56	14.28	12.92	10.83
US	21.23	20.08	17.74	17.4	16.54	16.64	16.27	15.93	13.6	11.73	10.62	8.9
UC	21.47	20.31	17.94	17.61	16.73	16.83	16.46	16.12	13.76	11.86	10.74	9
UB	7.02	6.64	5.87	5.76	5.47	5.51	5.38	5.27	4.5	3.88	3.51	2.94

Tabelul 5. Valorile optimizate ale produselor $(a_ib_j)_{1\le i\le 6;1\le j\le 12}$ folosind relațiile (20)

TV	UD	KK	KP	TP	ID	GS	AJ	BQ	ND	EP	AC	DY
DS	27.64	26.19	22.85	22.60	21.59	21.44	20.98	20.71	17.49	15.24	13.67	11.47
DC	27.35	25.91	22.61	22.36	21.36	21.22	20.76	20.50	17.30	15.08	13.52	11.35
DB	25.74	24.40	21.28	21.05	20.11	19.97	19.55	19.29	16.29	14.20	12.73	10.68
US	21.17	20.06	17.50	17.31	16.53	16.42	16.07	15.87	13.39	11.68	10.47	8.78
UC	21.40	20.28	17.69	17.50	16.71	16.60	16.25	16.04	13.54	11.80	10.58	8.88
UB	6.57	6.23	5.43	5.37	5.13	5.10	4.99	4.93	4.16	3.63	3.25	2.73

După cum se observă în Tabelul 6, fiecare dintre metodele definite de relațiile (18)-(20) îmbunătățește valoarea sumei obiectiv în raport cu expresia definită de formula aproximativă (21) şi reprezintă corecții ale acesteia. Astfel, relația (18) îmbunătățește soluția propusă de relația (21) în ipoteza erorii experimentale uniform distribuite între clase (eroarea experimentală absolută), relația (19) îmbunătățește soluția propusă de relația (21) în ipoteza erorii experimentale proporționale cu magnitudinea fenomenului observat (eroarea experimentală relativă) în timp ce relația (20) îmbunătățește soluția propusă de relația (21) minimizând statistica Pearson-Fisher X^2 (a cărei expresie este o Pearsoniană de tipul III [17]).

Tabelul 6. Valori comparative pentru eroarea experimentală întâmplătoare

	S^2				X^2				CV^2			
Cat	eq(21)	eq(18)	eq(20)	eq(19)	eq(21)	eq(18)	eq(20)	eq(19)	eq(21)	eq(18)	eq(20)	eq(19)
DS	23.4	18.76	24.12	57.97	1.10	0.937	1.127	2.308	0.056	0.0515	0.0573	0.0971
DC	59.7	48.48	59.86	104.95	3.08	2.497	3.052	4.847	0.164	0.133	0.1611	0.2365
DB	69.8	66.77	71.47	95.21	3.78	3.596	3.796	4.803	0.221	0.2078	0.2167	0.2633
US	41.6	49.03	41.66	35.34	2.72	3.19	2.709	2.358	0.186	0.2158	0.183	0.1635
UC	57.6	59.01	56.53	82.16	3.46	3.66	3.339	4.367	0.218	0.2375	0.2065	0.2444
UB	37.5	40.1	37.13	28.26	7.89	8.295	7.659	5.956	1.751	1.8018	1.6696	1.3512
UD	30.3	26.3	28.2	78.9	2.66	2.35	2.15	3.58	0.335	0.293	0.235	0.232
KK	15.3	13.5	15.8	18.7	0.76	0.64	0.73	0.88	0.045	0.033	0.035	0.044
KP	63	62.7	64	67.5	3.11	3.15	3.13	3.19	0.155	0.162	0.159	0.155
TP	34.3	31.4	33.3	76.5	2.79	2.69	2.37	3.67	0.357	0.340	0.256	0.242
ID	3.4	3.9	4	4.5	0.21	0.27	0.28	0.26	0.017	0.028	0.029	0.021
GS	26.2	25.6	26.9	28.6	2.29	2.45	2.52	2.42	0.319	0.349	0.352	0.327
AJ	45	47	45.3	43.4	2.56	2.71	2.60	2.44	0.152	0.168	0.164	0.148
BQ	21.5	20.4	21	31.8	1.93	1.71	1.67	2.19	0.253	0.205	0.182	0.193
ND	18.3	17.9	19.1	20.5	2.13	2.29	2.35	2.27	0.393	0.424	0.427	0.403
EP	2.9	3.2	3.3	3.8	0.53	0.64	0.66	0.62	0.133	0.158	0.163	0.142
AC	18.2	18.8	18.7	19.3	1.76	1.87	1.84	1.83	0.209	0.232	0.233	0.221
DY	11.1	11.5	11.2	10.6	1.31	1.40	1.39	1.27	0.228	0.255	0.258	0.227
Σ	289.5	**282.2**	290.8	404.1	22.04	22.17	**21.69**	24.62	2.596	2.647	2.493	**2.355**

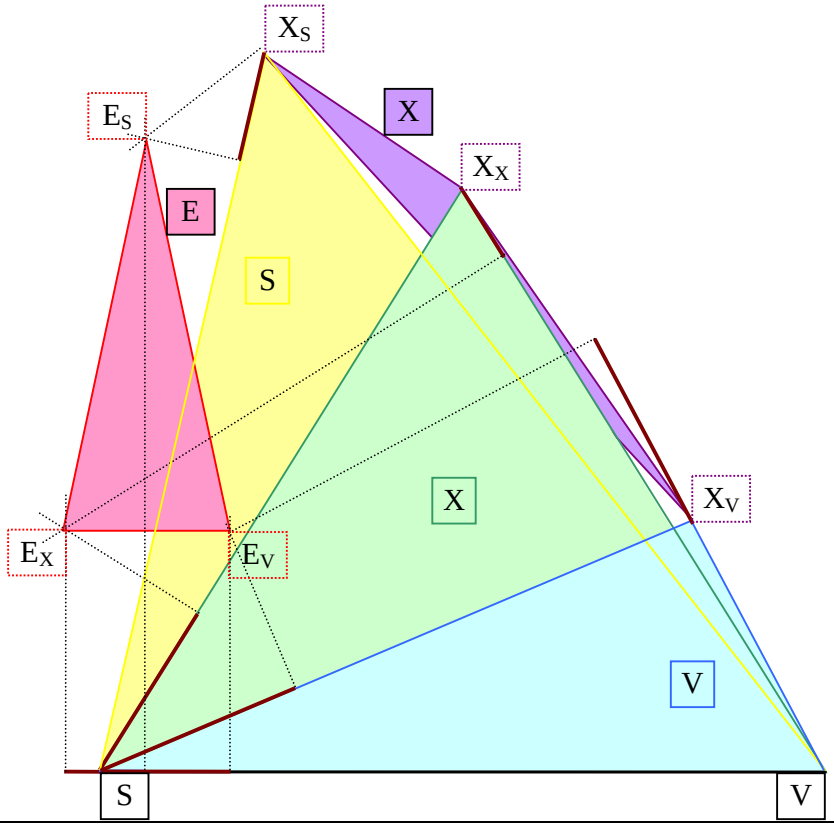
Valorile obținute în Tabelul 6 pentru eroarea experimentală în cele 3 forme ale sale (pătratică absolută S^2, pătratică relativă CV^2, şi Pearsoniană X^2) pentru cele 4 cazuri (frecvență teoretică estimată din contingență - eq. 21, frecvență teoretică estimată din minimizarea erorii pătratice absolute - eq. 18, frecvență teoretică estimată din minimizarea erorii pătratice relative - eq. 19, frecvență teoretică estimată din minimizarea statisticii Pearson-Fisher - eq. 20 sunt valori obținute într-un design de experiment în care există exact doi factori independenți (tip tratament şi tip sol sau factor A şi factor B) ceea ce permite o reprezentare în plan a distanțelor Euclidiene între rezultate.

În Figura 2 au fost reprezentate distanțele Euclidiene între erorile experimentale estimate de fiecare formulă (18)-(20) folosind triunghiuri Snyder [25] (diagrame frecvent folosite în cromatografie pentru a reprezenta 3 sau mai mulți parametrii ce depind de doi factori).

Figura 2 a fost realizată impunând reprezentarea la aceeaşi scară a ariei de eroare în raport cu cei doi factori (prin fixarea distanței dintre coordonata erorii experimentale în ipoteza S^2 = min. şi coordonata erorii experimentale în ipoteza CV^2 = min.) când coordonata în ipoteza X^2=min. s-a obținut prin maximizarea ariei de eroare (maximizarea ariilor triunghiurilor S, V şi X). Coordonatele contingenței s-au obținut astfel încât proiecțiile

[25] SNYDER Lloyd R. 1974. Classification of the solvent properties of common liquids. Journal of Chromatography A 92(2):223-230.

contingenței pe laturile triunghiurilor să împartă laturile în rapoartele observate între diferențele din Tabelul 6.



Legendă:
S - coordonata erorii experimentale în ipoteza S^2=min. (ec. 9);
V - coordonata erorii experimentale în ipoteza CV^2=min. (ec. 10);
X_S - coordonata erorii experimentale pătratice absolute (S^2) în ipoteza X^2=min. (ec.11);
X_V - coordonata erorii experimentale pătratice relative (CV^2) în ipoteza X^2=min. (ec.11);
X_X - coordonata statisticii X^2 în ipoteza X^2=min. (ec.11);
E_S - coordonata erorii experimentale pătratice absolute (S^2) în ipoteza contingenței (ec.4);
E_V - coordonata erorii experimentale pătratice relative (CV^2) în ipoteza contingenței (ec.4);
E_X - coordonata statisticii Pearson-Fisher (X^2) în ipoteza contingenței (ec.4);
S - triunghiul erorilor pătratice absolute (S^2); V - triunghiul erorilor pătratice relative (CV^2);
X - triunghiul statisticilor X^2; E - triunghiul de contingență; X - triunghiul de variație a statisticii X^2;

Figura 2. Distanțe Euclidiene între estimările erorilor experimentale

Construcția din Figura 1 permite aprecieri calitative cu privire la modelul de contingență definit de ec. (21) şi la statistica Pearson-Fisher în raport cu natura erorii experimentale. Astfel, se observă (în Figura 2) că singura intersecție între aria de contingență şi ariile de eroare se realizează cu eroarea pătratică absolută, deci contingența definită de ecuația (21) asigură agrementul între observație şi model numai pentru acest tip de erori din cele 3 cuprinse în studiu. De asemenea, singura intersecție a triunghiului de variație a statisticii X^2 este cu triunghiul statisticii X^2 ceea pe de o parte recomandă folosirea optimizării definite de ec. (14) [24] sau de ec. (16) [18] şi pe de altă parte demonstrează de ce testul Chi

Square este mai expus [26] decât alte teste cum ar fi Kolmogorov-Smirnov ([27,28]) şi Anderson-Darling ([29,30]) la erori de tip I respingând ipoteza nulă că variabila linie nu este în relație cu variabila coloană (asocierea este întâmplătoare) chiar când de fapt ipoteza este adevărată.

Se poate reprezenta poziția relativă a soluției propuse de relația (21) în raport cu valorile optimale propuse de relațiile (18)-(20). Pentru aceasta datele din Tabelul 6 au fost transformate cum arată Tabelul 7.

Tabelul 7. Transformarea valorilor reziduale din Tabelul 7 în valori relative la minim

Valori absolute	S^2	X^2	CV^2
E	289.5	22.04	2.596
S^2=min.	**282.2**	22.17	2.647
X^2=min.	290.8	**21.69**	2.493
CV^2=min.	404.1	24.62	**2.355**
Valori relative	S^2	X^2	CV^2
E	1.026	1.016	1.102
S^2=min.	**1**	1.022	1.124
X^2=min.	1.030	**1**	1.059
CV^2=min.	1.432	1.135	**1**

În Figura 3 s-a reprezentat în coordonatele definite de valorile pentru S^2, CV^2 şi X^2 valorile relative ale erorii (excesul de eroare) pentru rezultatele obținute prin estimarea simplă (E, relația 21), minimizarea erorii pătratice absolute (S^2=min., relația 18), minimizarea erorii pătratice relative (CV^2=min., relația 19) şi minimizarea statisticii X^2 (X^2=min., relația 20).

Rezultatul reprezentării din Figura 3 este consistent cu rezultatul proiecțiilor în plan din Figura 2. Figura 3 evidențiază că soluția propusă de (21) este foarte aproape de soluția propusă de (18) şi (20) fiind intermediară acestora şi este foarte departe de soluția propusă de (19).

[26] STEELE Mike, CHASELING Janet, HURST Cameron. 2005. Simulated Power of the Discrete Cramer-von Mises Goodness-of-Fit Tests. International Congress on Modelling and Simulation. Advances and Applications for management and decision making. MODSIM 2005:1300-1304.
[27] KOLMOGOROV Andrey. 1941. Confidence Limits for an Unknown Distribution Function. The Annals of Mathematical Statistics 12(4):461-463.
[28] SMIRNOV Nikolay V. 1948. Table for estimating the goodness of fit of empirical distributions. The Annals of Mathematical Statistics 19(2):279-281.
[29] ANDERSON Theodore W, DARLING Donald A. 1952. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics 23(2):193-212.
[30] SCHOLZ Fritz W, STEPHENS Michael A. 1987. K-sample Anderson-Darling Tests. Journal of the American Statistical Association 82(399):918-924.

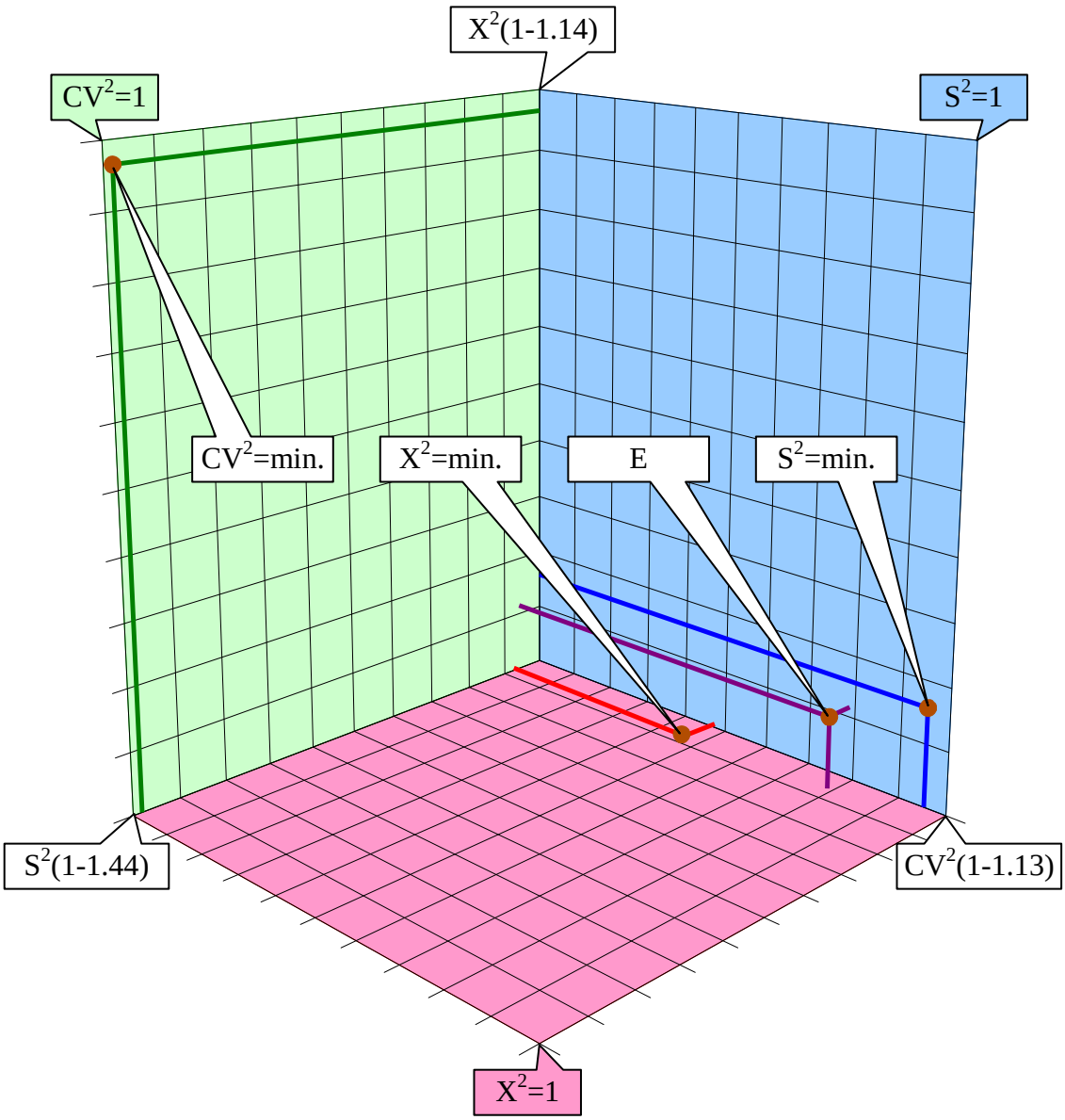


Figura 3. Poziția estimării empirice (21) în spațiul erorilor minime relative (18)-(19)-(20)

Probleme frecvent întâlnite în aplicarea testului χ^2 ca măsură a independenței

Nici aplicarea testului χ^2 pentru verificarea independenței nu este scutită de dificultăți în practică [31]. Astfel, FISHER a propus ca alternativă la testul χ^2 [32] testul care astăzi îi poartă numele (Fisher Exact Test, [33]), care se bazează pe calculul probabilităților marginale.

Pentru o tabelă de contingență 2X2, se cunoaşte că există exact un singur grad de libertate. Tabelul de mai jos (Tabelul 8) ilustrează această situație, în care impunerile sunt date de sumele observațiilor.

[31] FISHER Ronald A. 1934. Statistical Methods for Research Workers. Oliver and Boyd, Edinburgh.
[32] FISHER Ronald A. 1935. The Logic of Inductive Inference. Journal of the Royal Statistical Society 98:39-54.
[33] AGRESTI Alan. 1992. A Survey of Exact Inference for Contingency Tables. Statistical Science 7(1):131-177.

Tabelul 8. O tabelă de contingență 2X2 are un sigur grad de libertate (x)

X^2	Clasa A	Clasa Ω_1\A	Total Ω_1
Clasa B	x	n_1-x	n_1
Clasa Ω_2\B	n_2-x	n_3-n_1+x	n_2+n_3-n_1
Total Ω_2	n_2	n_3	n_2+n_3

Probabilitatea de a observa configurația din Tabel este dată de distribuția multinomială (relația 24), în timp ce valoarea statisticii Chi Square (X^2) este dată de relația (25):

$$p_{MN}(x;n_1,n_2,n_3)=\frac{n_1!\cdot n_2!\cdot n_3!\cdot(n_2+n_3-n_1)!}{x!\cdot(n_1-x)!\cdot(n_2-x)!\cdot(n_3-n_1+x)!\cdot(n_2+n_3)!} \tag{24}$$

$$X^2(x;n_1,n_2,n_3)=\frac{(xn_2+xn_3-n_1n_2)^2(n_2+n_3)}{n_1n_2n_3(n_2+n_3-n_1)} \tag{25}$$

Intervalul pe care observabila x poate lua valori este [0..min(n_1,n_2)].

Pentru exemplificarea problematicii s-au folosit datele din [32] (n_1 = 13, n_2 = 12, n_3 = 18) când intervalul de variație al lui x este [0..12] în timp ce valoarea observată a fost 10. Valoarea statisticii X^2 (relația 25) a fost reprezentată în Figura 4.

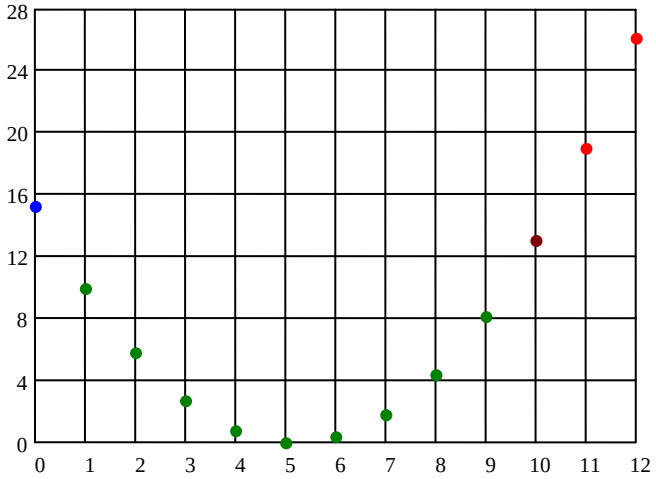


Figura 4. Valoarea statisticii X^2 în funcție de observabila independentă x

Așa cum se evidențiază în Figura 4, spațiul observațiilor posibile cu privire la valoarea statisticii X^2 în funcție de observabila independentă x este discret. Valoarea observată (x=10) este situată într-o vecinătate a unei margini (x=12) având două observații mai defavorabile decât ea (cu o valoare X^2 mai mare) în aceeași vecinătate (x=11 şi x=10) şi o observație mai defavorabilă în vecinătatea opusă (x=0).

Din acest moment există două abordări, corespunzător cu obiectivul comparației din tabela de contingență. Dacă obiectivul observației este evidențierea probabilității ca să se observe depărtări mai mari de la omogenitate decât depărtarea observată, atunci probabilitatea asociată observației se obține din cumularea probabilităților în x=0, x=10, x=11 şi x=12. Dacă obiectivul observației este evidențierea probabilității ca să se observe depărtări mai mari de la omogenitate în sensul depărtării observate, atunci probabilitatea asociată observației se obține

din cumularea probabilităților în x=10, x=11 şi x=12.

În Figura 5 a fost reprezentată probabilitatea observației (relația 24).

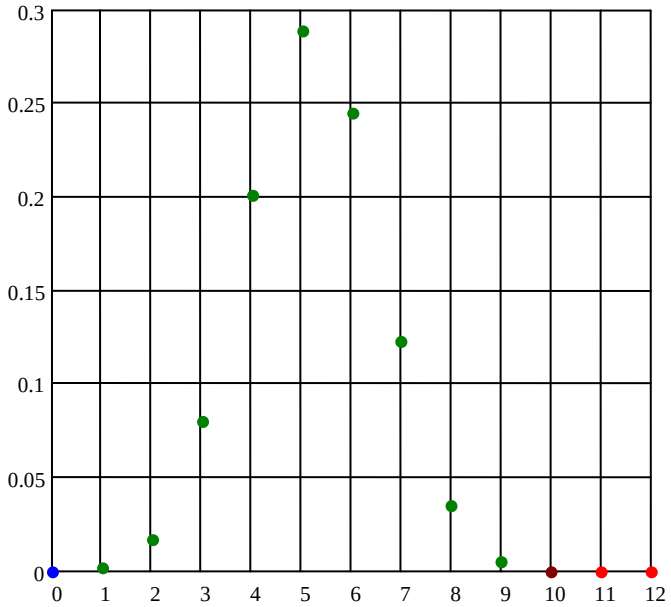


Figura 5. Valoarea statisticii probabilității observației în funcție de observabilă

Tabelul 9 prezintă pentru comparație valorile a trei probabilități: din distribuția χ^2 (p_{X2}), a probabilității de observare a unei depărtări de la omogenitate mai mari în sensul celei observate (p_{O2}) şi respectiv a unei depărtări mai mari în orice sens (p_{D2}). În această construcție probabilitatea din distribuția χ^2 (p_{X2}) este un estimator al unei depărtări mai mari în orice sens (p_{D2}).

Tabelul 9. Probabilități de observare

Probabilitate	Expresie de calcul	Valoare
p_{X2}	$\chi^2_{CDF}(X^2=13.03,df=1)$	$3.063\cdot10^{-4}$
$p_{O2}\,(x^2\geq X^2)$	$p_{MN}(10,13,12,18)+p_{MN}(11,13,12,18)+p_{MN}(12,13,12,18)$	$4.625\cdot10^{-4}$
$p_{O2}\,(x^2>X^2)$	$p_{MN}(11,13,12,18)+p_{MN}(12,13,12,18)$	$1.548\cdot10^{-5}$
$p_{D2}\,(x^2\geq X^2)$	$p_{O2}(x^2\geq X^2)+p_{MN}(0,13,12,18)$	$5.367\cdot10^{-4}$
$p_{D2}\,(x^2>X^2)$	$p_{O2}(x^2>X^2)+p_{MN}(0,13,12,18)$	$8.702\cdot10^{-5}$

Tabelul 9 arată cum testul χ^2 este în eroare atunci când valorile din tabelul de contingență se abat de la condițiile impuse asupra frecvențelor observate (cel mult 20% dintre celulele contingenței să conțină valori mai mici decât 5). Tabelul 9 mai arată cum în aceste cazuri testul Chi Square este expus la erori de tipul I (acordând o probabilitate mai mică decât cea reală evenimentului de a se produce observația observată, se află în riscul de a accepta ipoteza contrară chiar dacă ea nu este adevărată, ceea ce este totuna cu a respinge ipoteza nulă chiar dacă ea este adevărată).

Pentru a corecta semnificația statistică pentru tabele de contingență (sau frecvență) cu

puține observații, Frank YATES a propus [34] o corecție la continuitate în care în expresia ecuației statisticii (relațiile (10), (11) şi (12)) din modulul diferenței între frecvența observată şi frecvența estimată în ipoteza independenței estimare se scade 0.5 simbolizând mijlocul intervalului de frecvență în timp ce MANTEL şi HAENSZEL au propus [35] ponderarea (împărțirea) statisticii X^2 cu df/(df-1), unde df este numărul de grade de libertate ale asocierii. Nici una dintre aceste ajustări însă nu este o alternativă decât la χ^2, testul Fisher Exact reprezentând testul de aur (Golden Test) pentru valoarea adevărată a probabilității de apariție a evenimentului observat.

3.4. Regresii Liniare Multiple

O ecuație de regresie liniară multiplă este o ecuație de forma:

$$b_0 + b_1X_1 + ... + b_nX_n = \hat{Y} \sim Y \quad (1)$$

sau

$$b_1X_1 + ... + b_nX_n = \hat{Y} \sim Y \quad (2)$$

unde Y este un şir de observații experimentale supuse erorii experimentale întâmplătoare iar $\{X_1, ..., X_n\}$ reprezintă o mulțime de descriptori $\{X_i\}_{1\le i\le n}$ asupra cărora se formulează ipoteza că o asociere liniară a acestora explică observațiile experimentale efectuate, iar şirul $(b_i)_{i\le n}$ reprezintă parametrii modelului (şi în acelaşi timp coeficienții ecuației).

Următoarele caracteristici definesc ecuațiile (1) şi (2):

÷ numărul de variabile independente: n = |X|;

÷ numărul de observații experimentale: m = |Y| = |X_1| = ... = |X_n|;

÷ numărul de parametrii ai modelului: |b| = n+1 pentru (1) şi |b| = n pentru (2).

În obținerea parametrilor ecuației de regresie (1) sau (2) se asumă următoarele ipoteze:

÷ valorile variabilei Y sunt normal distribuite; eroarea de măsură a lui Y este întâmplătoare şi de asemenea distribuită normal;

÷ variabilele $X_1, ..., X_n$ au valori distribuite normal şi nu sunt afectate de erori.

Obținerea parametrilor unei ecuații de regresie $(b_i)_{i\le n}$ din observații este întotdeauna însoțită de un risc de a fi în eroare, iar în ipoteza că există relația liniară definită de (1) sau (2) folosind distribuția Student t se poate aprecia semnificația statistică şi intervalul de încredere al acestora.

Pentru ca ecuația (1) sau (2) să admită soluție unică este necesar (nu însă şi suficient) ca $n \le m-1$. Pentru ca parametrii ecuației de regresie $(b_i)_{0\le i\le n}$ să aibă şi semnificație statistică

[34] YATES Frank. 1934. Contingency table involving small numbers and the χ^2 test. Journal of the Royal Statistical Society (Supplement) 1: 217-235.
[35] MANTEL Nathan, HAENSZEL William. 1959. Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4):719-748.

este necesar (nu însă şi suficient) ca $n \leq m-6$.

În cazul absenței semnificației statistice pentru coeficientul b_0, ecuația (1) se poate restrânge la ecuația (2).

Absența semnificației statistice pentru un coeficient b_i al unei variabile X_i ($1 \leq i \leq n$) în ecuația de regresie (1) asociată cu absența semnificației statistice a acestuia şi în ecuația de regresie (2) impune respingerea ipotezei legăturii liniare între observabila Y şi variabila X_i.

În aceste ipoteze problema determinării coeficienților (b_i) ale ecuației se rezolvă prin minimizarea sumei erorilor observat vs. cunoscut:

$$\Sigma_{1 \leq i \leq m}(\hat{Y}_i - Y_i)^2 \rightarrow \min. \qquad (3)$$

Rezolvarea ecuației de minimizare presupune rezolvarea unui sistem de ecuații liniar şi omogen ale cărei necunoscute sunt coeficienții (b_i).

Rezolvarea ecuației de regresie (1) prin minimizarea pătratelor erorilor (LSE - least squares error) dată de relația (3) implică:

÷ exprimarea matriceală a sistemului de ecuații liniare şi omogene date de (3):

$$\begin{array}{c} \begin{matrix} 0 & 1 & \dots & n \end{matrix} \\ a = \begin{pmatrix} 1 & M(X_1) & \dots & M(X_n) \\ M(X_1) & M(X_1X_1) & \dots & M(X_1X_n) \\ \dots & \dots & \dots & \dots \\ M(X_n) & M(X_nX_1) & \dots & M(X_nX_n) \end{pmatrix} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix} ; \\ \begin{matrix} 0 \end{matrix} \qquad \begin{matrix} 0 & 1 & \dots & n \end{matrix} \\ b = \begin{pmatrix} M(Y) \\ M(X_1Y) \\ \dots \\ M(X_nY) \end{pmatrix} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix} ; \; c = \begin{pmatrix} 1/m & 0 & \dots & 0 \\ 0 & 1/m & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1/m \end{pmatrix} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix} \end{array} \qquad (4)$$

÷ construcția matricei extinse a sistemului:

$$\begin{matrix} -1 & 0 & 1 & \dots & n & n+1 & n+2 & \dots & 2n+1 \end{matrix}$$
$$\begin{pmatrix} M(Y) & 1 & M(X_1) & \dots & M(X_n) & 1/m & 0 & \dots & 0 \\ M(X_1Y) & M(X_1) & M(X_1X_1) & \dots & M(X_1X_n) & 0 & 1/m & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ M(X_nY) & M(X_n) & M(X_nX_1) & \dots & M(X_nX_n) & 0 & 0 & \dots & 1/m \end{pmatrix} \begin{matrix} 0 \\ 1 \\ \dots \\ n \end{matrix}$$

÷ transformarea matricei extinse a sistemului folosind algoritmul Gauss-Jordan (prin operații elementare efectuate asupra liniilor matricei) având ca obiectiv (şi până când) se obține matricea unitară în spațiul matricei a şi când se obțin coeficienții $(b_i)_{0 \leq i \leq n}$ şi erorile standard ale acestora $(s(b_i))_{0 \leq i \leq n}$:

$$\begin{array}{c} \begin{matrix} -1 & 0 & 1 & ... & n & n+1 & n+2 & ... & 2n+1 \end{matrix} \\ \begin{pmatrix} b_0 & 1 & 0 & ... & 0 & s(b_0) & 0 & ... & 0 \\ b_1 & 0 & 1 & ... & 0 & 0 & s(b_1) & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... & ... \\ b_n & 0 & 0 & ... & 1 & 0 & 0 & ... & s(b_n) \end{pmatrix} \end{array} \begin{matrix} 0 \\ 1 \\ ... \\ n \end{matrix} \qquad (5)$$

Rezolvarea ecuației de regresie (2) prin minimizarea pătratelor erorilor (LSE - least squares error) dată de relația (3) implică:

÷ exprimarea matriceală a sistemului de ecuații liniare şi omogene date de (3):

$$b=\begin{matrix} 0 \\ \begin{pmatrix} M(X_1Y) \\ ... \\ M(X_nY) \end{pmatrix} \end{matrix} \begin{matrix} 1 \\ ... \\ n \end{matrix}; \; a=\begin{matrix} 1 & ... & n \\ \begin{pmatrix} M(X_1X_1) & ... & M(X_1X_n) \\ ... & ... & ... \\ M(X_nX_1) & ... & M(X_nX_n) \end{pmatrix} \end{matrix} \begin{matrix} 1 \\ ... \\ n \end{matrix}; \; c=\begin{matrix} 1 & ... & n \\ \begin{pmatrix} 1/m & ... & 0 \\ ... & ... & ... \\ 0 & ... & 1/m \end{pmatrix} \end{matrix} \begin{matrix} 1 \\ ... \\ n \end{matrix} \qquad (6)$$

÷ construcția matricei extinse a sistemului:

$$\begin{matrix} 0 & 1 & ... & n & n+1 & ... & 2n \\ \begin{pmatrix} M(X_1Y) & M(X_1X_1) & ... & M(X_1X_n) & 1/m & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ M(X_nY) & M(X_nX_1) & ... & M(X_nX_n) & 0 & ... & 1/m \end{pmatrix} \end{matrix} \begin{matrix} 1 \\ ... \\ n \end{matrix}$$

÷ transformarea matricei extinse a sistemului folosind algoritmul Gauss-Jordan (prin operații elementare efectuate asupra liniilor matricei) având ca obiectiv (şi până când) se obține matricea unitară în spațiul matricei a şi când se obțin coeficienții $(b_i)_{0\le i\le n}$ şi erorile standard ale acestora $(s(b_i))_{0\le i\le n}$:

$$\begin{array}{c} \begin{matrix} 0 & 1 & ... & n & n+1 & ... & 2n \end{matrix} \\ \begin{pmatrix} b_1 & 1 & ... & 0 & s(b_1) & ... & 0 \\ ... & ... & ... & ... & ... & ... & ... \\ b_n & 0 & ... & 1 & 0 & ... & s(b_n) \end{pmatrix} \end{array} \begin{matrix} 1 \\ ... \\ n \end{matrix} \qquad (7)$$

Coeficientul de corelație oferă o măsură a legăturii liniare între cele două variabile (Y şi $\hat{Y}$) şi se calculează pe baza formulei (unde M este valoarea medie):

$$r(Y,\hat{Y}) = \frac{cov(Y,\hat{Y})}{s(Y)\cdot s(\hat{Y})} = \frac{M(Y\hat{Y}) - M(Y)\cdot M(\hat{Y})}{\sqrt{M(Y^2) - M^2(Y)}\sqrt{M(\hat{Y}^2) - M^2(\hat{Y})}} \qquad (8)$$

Semnificația statistică a legăturii liniare caracterizate de corelația dată de relația (8) este obținută din statistica Fisher F (unde |b| este numărul de coeficienți folosiți în estimare), iar probabilitatea asociată respingerii modelului liniar din funcția cumulativă de probabilitate (CDF) a distribuției Fisher:

$$F(r) = \frac{r^2}{1-r^2}\cdot\frac{m-|b|}{n}; \; p_F = F_{CDF}\left(F(r), n, m-|b|\right) \qquad (9)$$

În ipoteza că sistemul de ecuații admite o soluție unică pentru ecuația de regresie, ipotezele asumate permit şi obținerea semnificațiilor statistice ale parametrilor $t(b_i)$ şi a

probabilităților asociate valorilor semnificativ statistic nenule ale acestora folosind distribuția Student t (unde $s(b_i)$ este dat de (5) pentru (1) și de (7) pentru (2)):

$$t(b_i) = \frac{b_i}{s(b_i)} \sqrt{\frac{m-|b|}{\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2}};\ p_t = t_{CDF}\left(t(b_i), m-|b|\right) \tag{10}$$