# Adaptacyjne wersje gładkiego testu zgodności z rozkładem logistycznym

Alicja Janic

Instytut Matematyki i Informatyki, Politechnika Wrocławska

5 grudnia 2009

## Problem testowania

Niech $\underline{X} = (X_1, ..., X_n)$ będzie próbą z ciagłego rozkładu o gęstości $g(x)$. Testujemy

$$H_0 : g(x) \in \{f(x;\beta), \beta \in \mathcal{B}\}, \quad \text{gdzie} \quad f(x;\beta) = \frac{1}{\beta_2} f\left(\frac{x-\beta_1}{\beta_2}\right) \quad \text{i}$$

$$f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}.$$

## Klasa gładkich alternatyw

$$g_k(x;\eta) = c_k(\theta)\exp\left\{\sum_{j=1}^{k}\theta_j\phi_j(F(x;\beta))\right\}f(x;\beta), \qquad (1)$$

gdzie $\theta \in R^k$, $\eta = (\theta, \beta)$, $c_k(\theta)$- stała normująca, $\phi_0 \equiv 1, \phi_1, \phi_2, ...$ ortonormalny układ wielomianów Legendre'a na $[0, 1]$,
$F$- dystrybuanta rozkładu logistycznego.

W rodzinie (1) hipoteza $\mathcal{H}_0$ jest równoważna hipotezie

$$\mathcal{H}_0^*(k) : \eta = \eta_0 = (0, \beta).$$

- Janic i Ledwina (2009) *Journal of Statistical Theory and Practice*.

## Wprowadzamy oznaczenia:

- $$\ell(x;\eta) = \log g_k(x;\eta);$$
- $$\dot{\ell}_\eta(x;\eta) = \frac{\partial}{\partial \eta}\ell(x;\eta) = \Big(\dot{\ell}_\theta(x;\eta), \dot{\ell}_\beta(x;\eta)\Big).$$

## Dla rodziny wykładniczej (1) uzyskujemy:

- $$\dot{\ell}_\theta(x;\eta_0) = \dot{\ell}_\theta(x;\beta) = \Big(\phi_1(F(x;\beta)), ..., \phi_k(F(x;\beta))\Big)$$

$$\text{oraz} \quad \dot{\ell}_\beta(x;\eta_0) = \dot{\ell}_\beta(x;\beta) = \frac{\partial}{\partial \beta}\log f(x;\beta).$$

- Dla rozkładu logistycznego:

$$\dot{\ell}_{\beta_1}(x;\beta) = \frac{1}{\sqrt{3}}\dot{\ell}_{\theta_1}(x;\beta) = \frac{1}{\sqrt{3}}\phi_1(F(x;\beta)).$$

- Zdefiniujmy efektywny wektor wynikowy:

$$\ell^*(x;\beta) = \dot{\ell}_\theta(x;\beta) - \dot{\ell}_\beta(x;\beta){I_{\beta\beta}}^{-1}I_{\beta\theta}.$$

- Wprowadzamy teraz

$$I_* = E_{\eta_0}[\ell^*(X;\beta)]^T[\ell^*(X;\beta)]$$

i otrzymujemy

- $$I_*^{-1} = I + I_{\theta\beta}[I_{\beta\beta} - I_{\beta\theta}I_{\theta\beta}]^{-1}I_{\beta\theta}.$$

- Dla rozkładu logistycznego:

$$I_{\beta\beta} - I_{\beta\theta} I_{\theta\beta} = \frac{1}{\beta_2^2} \begin{pmatrix} 0 & 0 \\ 0 & \frac{3+\pi^2}{9} - \sum_{j=1}^{k} m_j^2 \end{pmatrix}, \quad \text{gdzie}$$

$$m_j = \begin{cases} 0, & j \text{ parzyste}, \\ \frac{2\sqrt{2j+1}}{(j-1)(j+2)}, & j \text{ nieparzyste}. \end{cases}$$

- Wyrzucamy z rodziny wykładniczej (1) wielomian $\phi_1$ i k-wymiarowa rodzina ma postać:

$$g_k^*(x;\eta) = c_k(\theta) \exp\left\{ \sum_{j=2}^{k+1} \theta_j \phi_j(F(x;\beta)) \right\} f(x;\beta). \qquad (2)$$

## Efektywna statystyka wynikowa

Niech $\tilde{\beta}$ jest niezmienniczym estymatorem $\beta$. Efektywna statystyka wynikowa dla $\mathcal{H}_0^*(k)$ w rodzinie (2) dana jest wzorem

$$W_k(\tilde{\beta}) = \Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell^*(X_i;\tilde{\beta})\Big] I_*^{-1}\Big[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\ell^*(X_i;\tilde{\beta})\Big]^T \qquad (3)$$

$$= \sum_{j=1}^{k}\{\sqrt{n}U_j(\tilde{\beta})\}^2, \quad \text{gdzie} \quad U_j(\beta) = \Big[\frac{1}{n}\sum_{i=1}^{n}\ell^*(X_i;\beta)\Big](I_*^{-1})^{1/2}.$$

Efektywne statystyki wynikowe zostały wprowadzone przez Neymana (1954, 1959) a testy oparte na tych statystykach nazwane testami $C(\alpha)$ (Neyman, 1959).

## Zastosowane reguły wyboru:

- Janic i Ledwina (2009)

$$S1 = S1(\tilde{\beta}) = \min\Big\{k,\ k = 1, ..., d(n) : W_k(\tilde{\beta}) - k\log n \geq W_j(\tilde{\beta}) - j\log n,\ j = 1, ..., d(n)\Big\}. \tag{4}$$

W przypadky $\tilde{\beta} = \hat{\beta}$ reguła $S1$ została przebadana przez Kallenberga i Ledwine (1997). Reguła $S1$ z estymatorami największej wiarygodności została zaproponowana również przez Aerts, Claeskens i Hart (2000).

- Inglot i Ledwina (2006)
  Definiujemy nową karę

$$\Pi(j,n) = j\log n \cdot \mathbf{1}_{A_n'} + 2j \cdot \mathbf{1}_{A_n}, \quad \text{gdzie}$$

$$A_n = \{\max_{1\le j\le d(n)} |\sqrt{n}U_j(\tilde{\beta})| > \sqrt{c\log n}\} \tag{5}$$

  i nową regułę

$$T1 = T1(\tilde{\beta}) = \min\Big\{k,\ k = 1,...,d(n) : W_k(\tilde{\beta}) - \Pi(k,n) \ge W_j(\tilde{\beta}) - \Pi(j,n),\ j = 1,...,d(n)\Big\}. \tag{6}$$

## Jak wyznaczamy $c$? (Inglot i Janic, 2009)

Barierę $c$ wybieramy tak, aby $P_0(A_n) = \delta$. Przy $H_0$

$$(\sqrt{n}U_1(\tilde{\beta}), ..., \sqrt{n}U_{d(n)}(\tilde{\beta})) \approx (Z_1, ...Z_{d(n)}).$$

Zatem, wykorzystyjąc powyższą aproksymację, wyznaczamy $c$ z równania

$$\Phi(\sqrt{c \log n}) = 1 - \frac{\delta}{2d(n)}.$$

Przyjmijmy $\delta = 0.025$, w symulacjach $d(n) = 12$. Zatem dla $n = 50$ dostajemy $c = 2.394$, a dla $n = 100$ dostajemy $c = 2.033$.

## Estymatory:

Stosujemy estymatory uzyskane metodą momentów

$$\tilde{\beta}[m] = \left(\bar{X}, \frac{\sqrt{3}}{\pi} S\right).$$

## Statystyki:

- $WS1 = W_{S1(\tilde{\beta}[m])}(\tilde{\beta}[m])$;
- $WT1 = W_{T1(\tilde{\beta}[m])}(\tilde{\beta}[m])$.

## Twierdzenie 1

Janic i Ledwina (2009)
Przy pewnych założeniach o szybkości zbieżności $d(n)$

$$P_0(S1 > 1) \to 0$$

i w konsekwencji

$$W_{S1} \xrightarrow{\mathcal{D}} \chi^2_1 \text{ przy } H_0.$$

## Twierdzenie 2

Przy pewnych założeniach o szybkości zbieżności $d(n)$ i dla $c \geq 2$

$$P_0(A_n) \to 0 \text{ dla } n \to \infty.$$

Zatem

$$P_0(T1 = S1) \to 1 \text{ gdy } n \to \infty$$

i w konsekwencji

$$P_0(T1 > 1) \to 0 \text{ oraz } W_{T1} \xrightarrow{\mathcal{D}} \chi^2_1 \text{ przy } H_0.$$

**Tabela 1.** Wartości krytyczne $WS1$ i $WT1$, $d(n) = 12$, $\alpha = 0.05$, 30 000 MC.

| | $WS1$ | $WT1$ |
|---|---|---|
| $n = 50$ | 5.8831 | 7.3262 $(c = 2.4)$ |
| $n = 100$ | 5.3429 | 7.0440 $(c = 2.0)$ |
| $n = 500$ | 4.2810 | 4.5561 $(c = 2.0)$ |

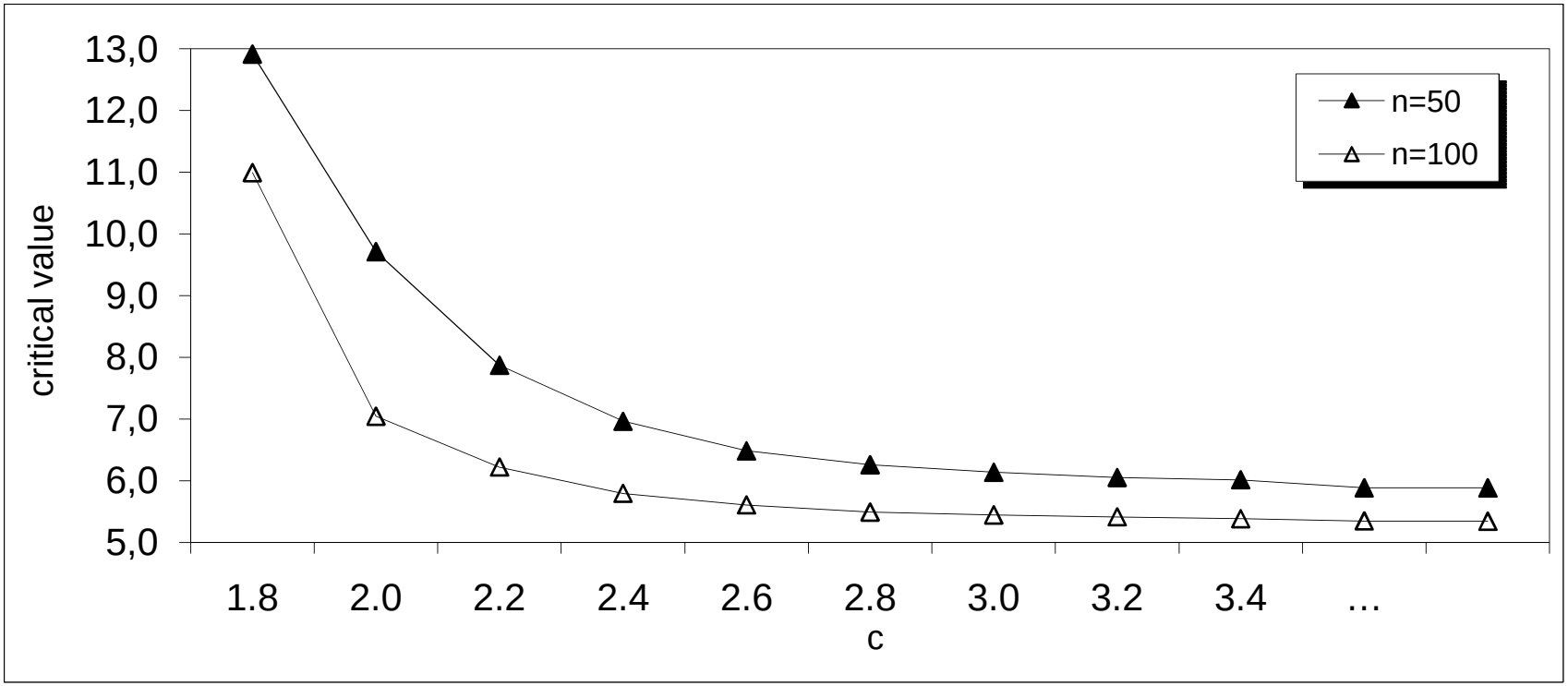
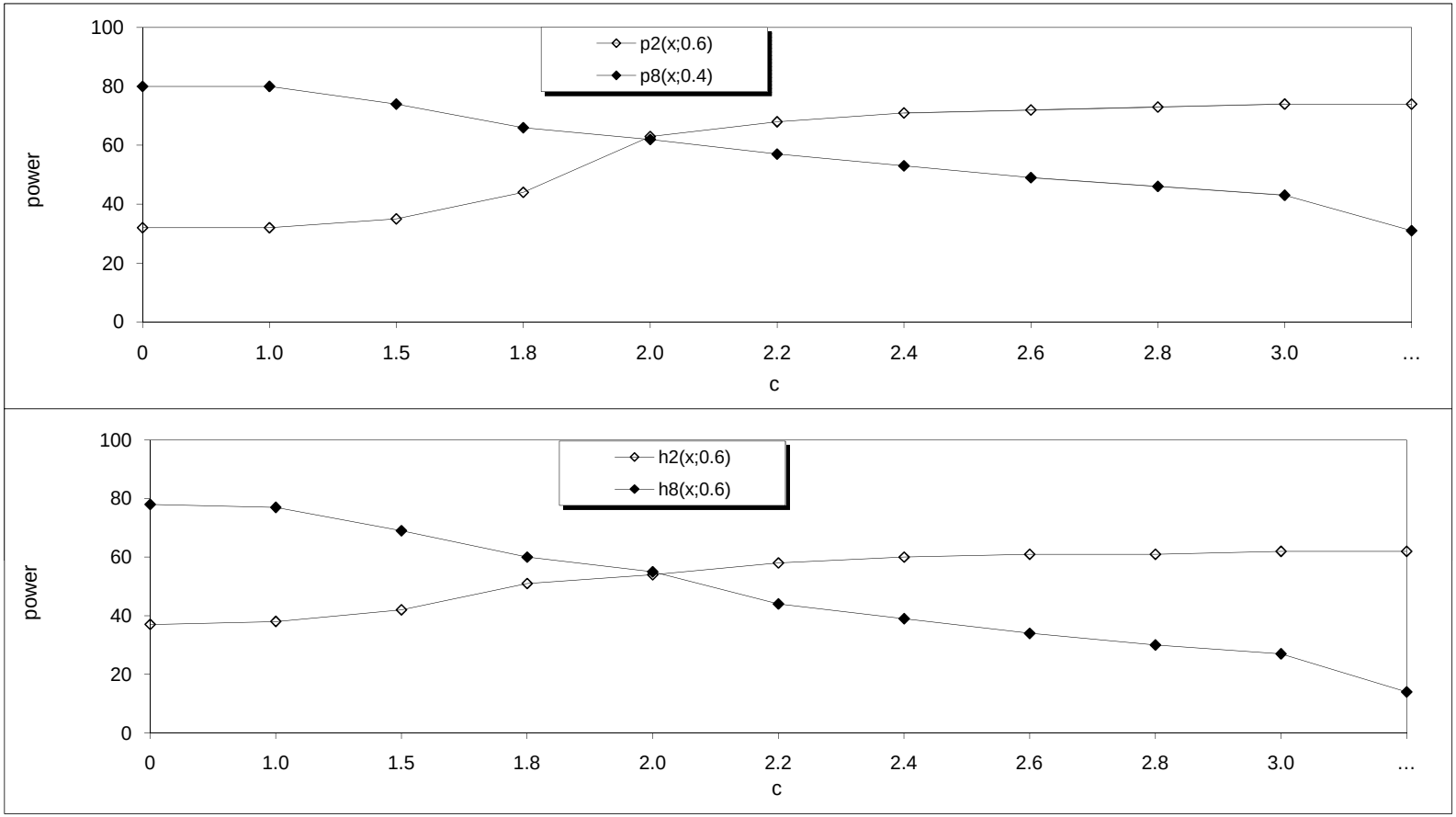

13,0
12,0
11,0
10,0
9,0
8,0
7,0
6,0
5,0
critical value
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
…
c
n=50
n=100

**Tabela 2.** Opis przykładowych alternatyw użytych w symulacjach.

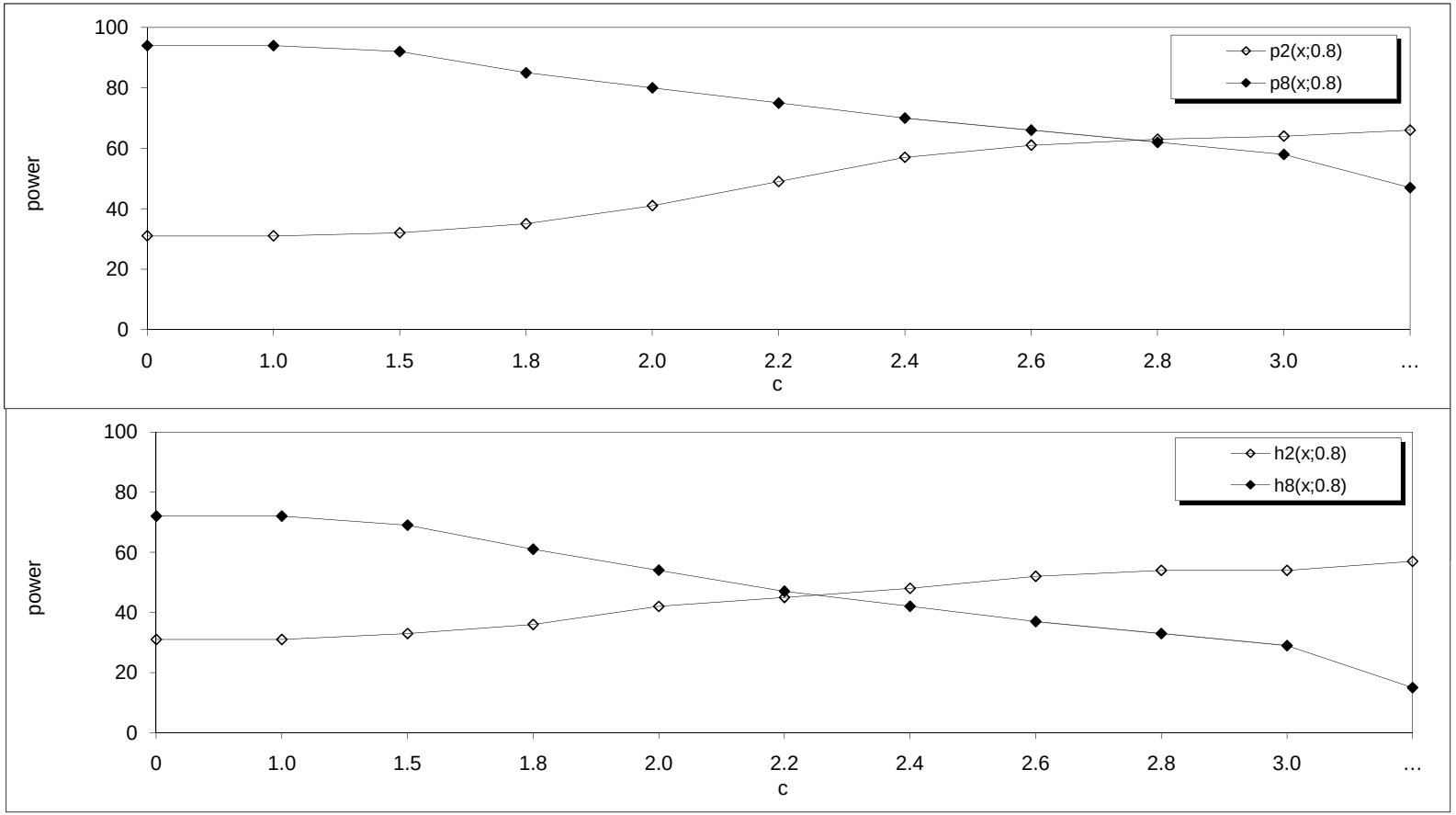
| Symbol | Gęstość/Definicja |
|---|---|
| Beta$(p,q)$ | $x^{p-1}(1-x)^{g-1}\{B(p,q)\}^{-1},\ x\in(0,1)$ |
| LC$(p,m)$ | $p\varphi(x-m)+(1-p)\varphi(x),\ x\in R$ |
| LG$(p,q)$ | $q^{-p}\{\Gamma(p)\}^{-1}\exp\{px-q^{-1}\exp(x)\},\ x\in R$ |
| LN$(g,d)$ | $X=\exp\{d^{-1}(Z-g)\}$ |
| SB$(g,d)$ | $X=\exp\{d^{-1}(Z-g)\}[1+\exp\{d^{-1}(Z-g)\}]^{-1}$ |
| SC$(p,d)$ | $d^{-1}p\varphi\left(d^{-1}x\right)+(1-p)\varphi(x),\ x\in R$ |
| SU$(g,d)$ | $X=\sinh\{d^{-1}(Z-g)\}$ |
| TU$(\lambda_1,\lambda_2)$ | $X=[U^{\lambda_1}-1]/\lambda_1-[(1-U)^{\lambda_2}-1]/\lambda_2$ |

$p_j(x;\rho)=g_{12}(x;\eta)$, gdzie $\eta=(\theta,\beta_1,\beta_2)=(\rho e_j,0,1)$;
$h_j(x,\rho)=f(x)(1+\rho\cos(\pi jF(x)))$;
$e_1,...,e_{12}$ jest standardową bazą w przestrzeni Euklidesowej $R^{12}$.

# Rysunek 2. $n = 100, d(n) = 12, \alpha = 0.05$, 10000 MC.



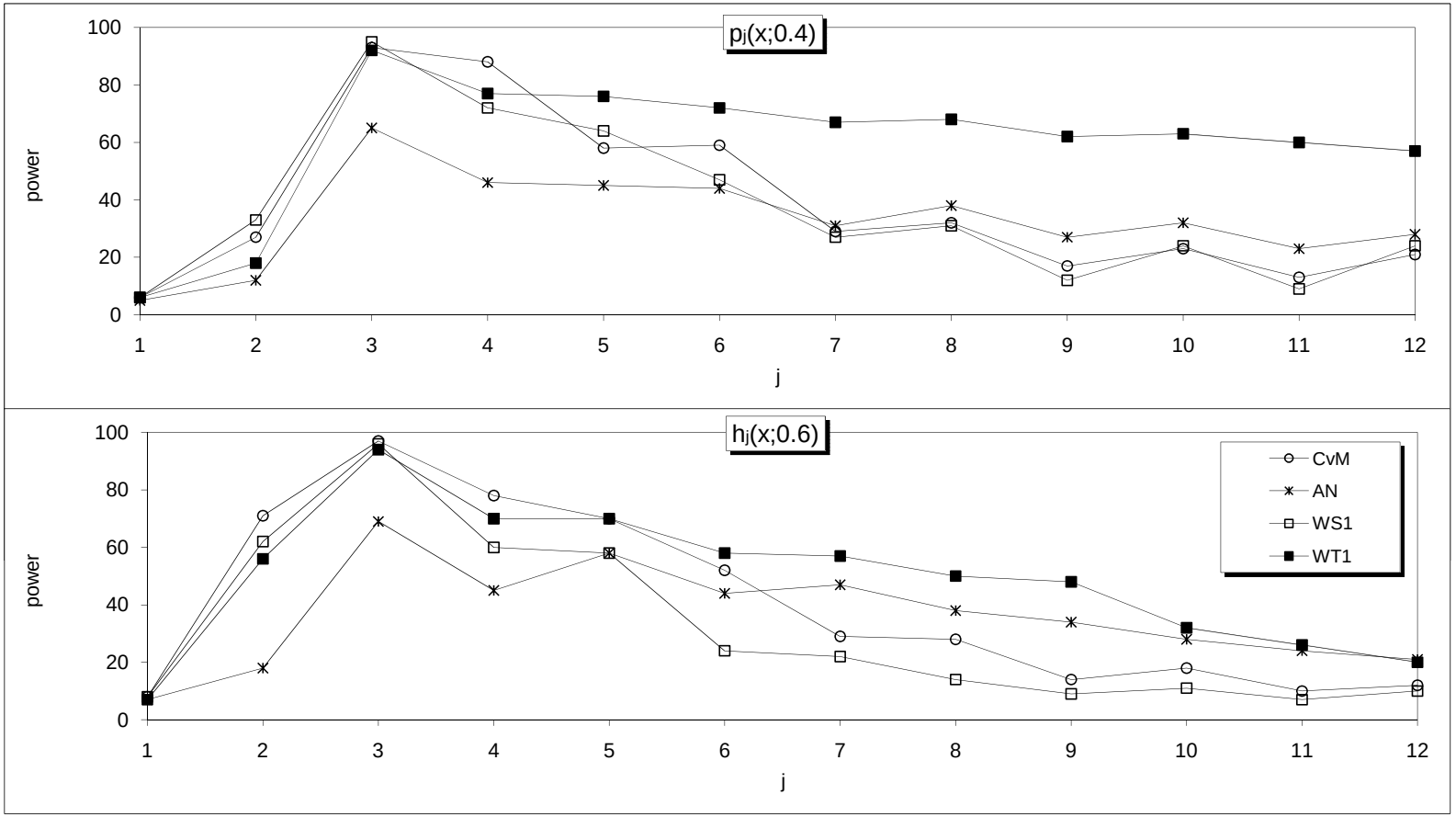
# Rysunek 3. $n = 50, d(n) = 12, \alpha = 0.05$, 10000 MC.



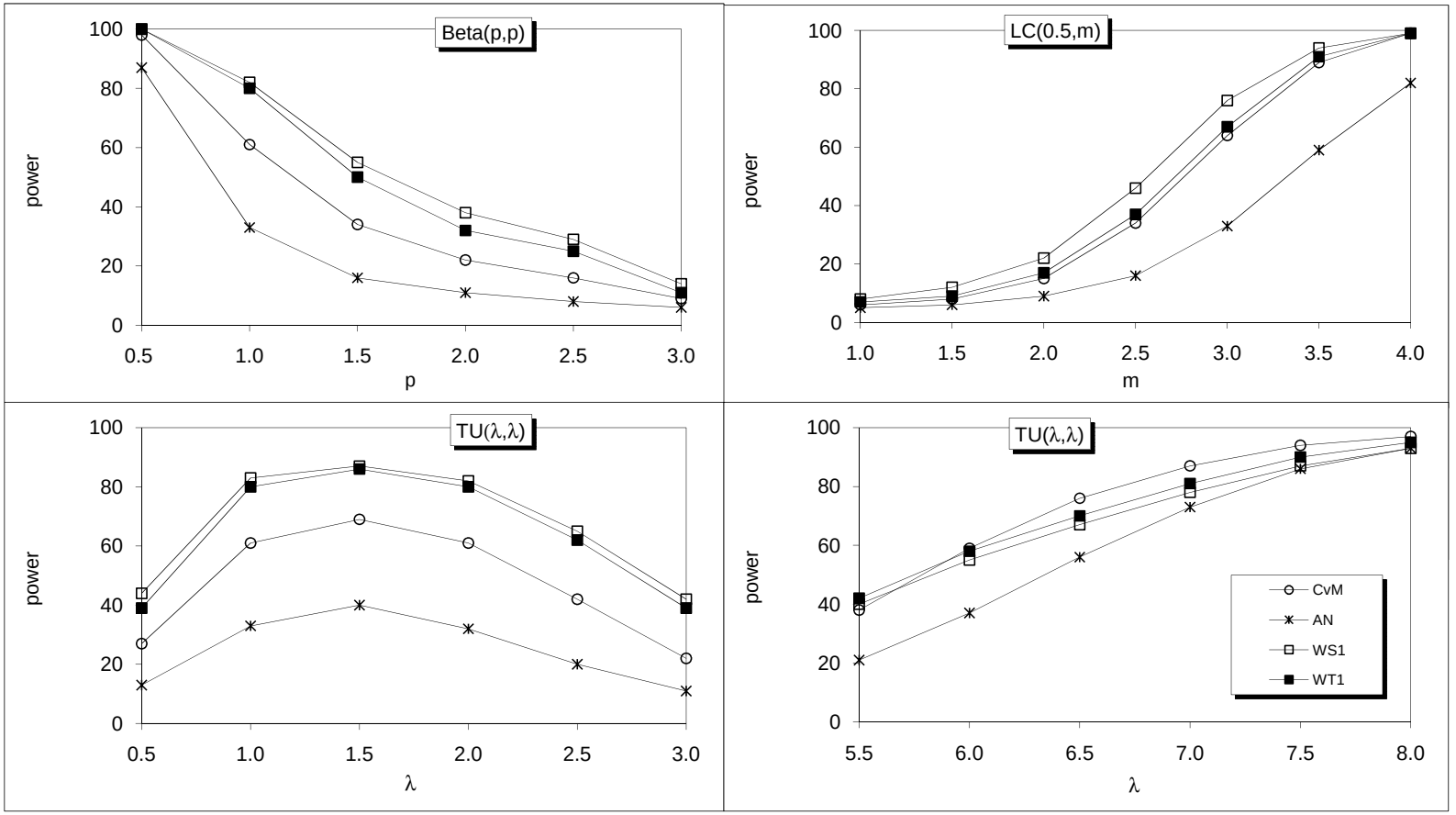
**Porównujemy empiryczne moce testów opartych na statystykach:**

- $CvM$ (statystyka - Craméra - von Misesa) - test typu EDF (Stephens, 1992);
- $AN$ - test typu $\chi^2$ (Aguirre i Nikulin, 1994);
- $WS1$ (Janic i Ledwina, 2009);
- $WT1$.

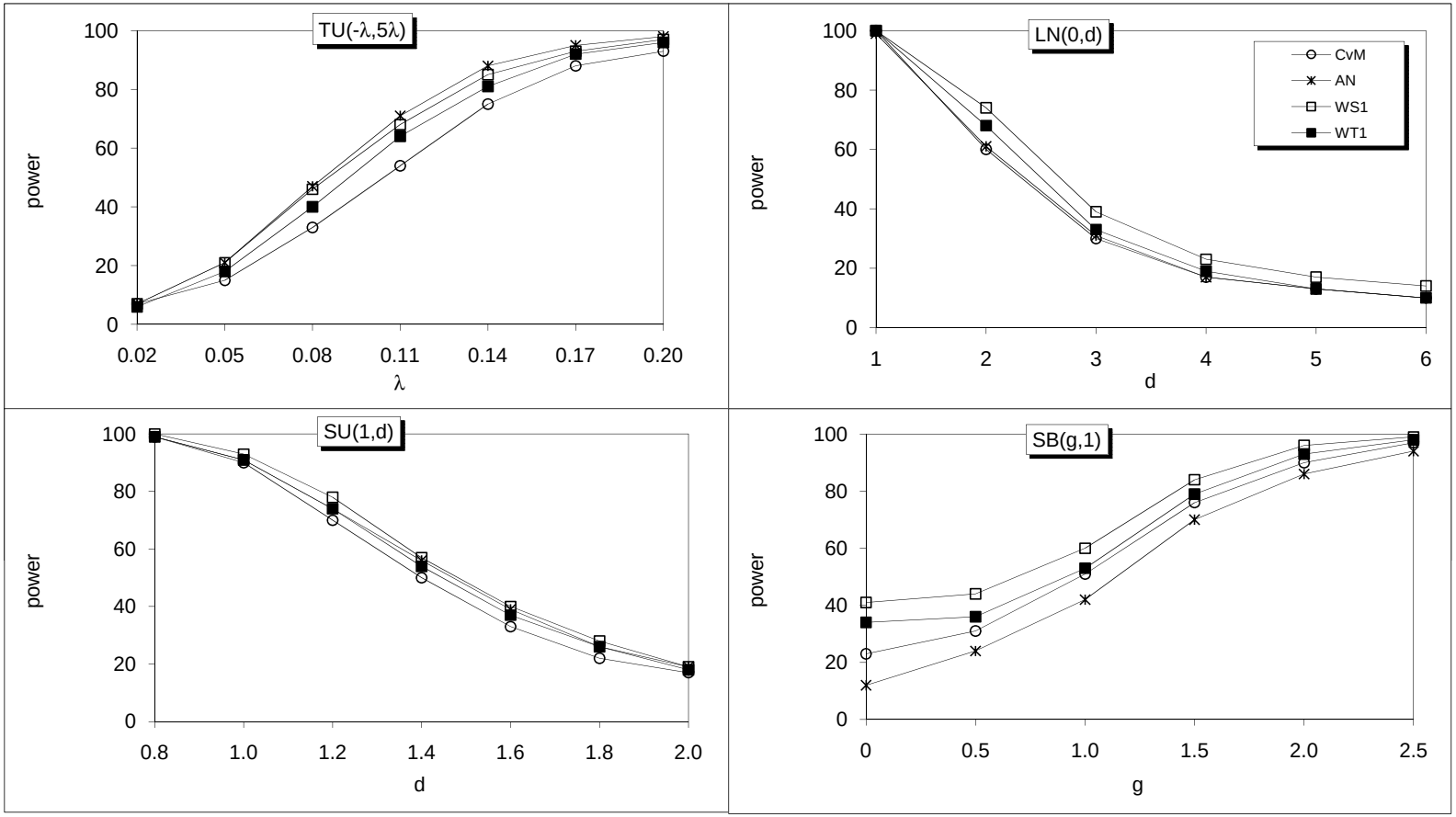
# Rysunek 4. $n = 100, d(n) = 12, \alpha = 0.05$, 10000 MC.



# Rysunek 5. Alternatywy symetryczne. $n = 50, d(n) = 12$.



# Rysunek 6. Alternatywy skośne. $n = 50, d(n) = 12$.



Aerts, M., Claeskens, G., Hart, J.D., 2000. Testing lack of fit in multiple regression. *Biometrika* 87, 405–424.

Aguirre, N., Nikulin, M., 1994. Chi-squared goodness-of-fit test for the family of logistic distribution. *Kybernetika* 30, 214–222.

D'Agostino, R.B., Stephens, M.A., 1986. *Goodness–of–Fit Techniques.* Dekker, New York.

Inglot, T., Kallenberg, W.C.M., Ledwina, T., 1997. Data driven smooth tests for composite hypotheses. *Ann. Statist.* 25, 1222–1250.

Inglot, T., Ledwina, T., 2006. Towards data driven selection of a penalty function for data driven Neyman tests. *Linear Algebra and its Appl.* 417, 124–133.

Inglot, T., Janic, A., 2009. How powerful are data driven score tests for uniformity. *Applicationes Mathematicae* 36, 375–395.

Jakubiec, A., 1998. Power simulations of goodness-of-fit tests for logistic distribution. *Diploma thesis (in Polish)*, Wrocław University of Technology.

Janic, A., Ledwina, T., 2009. Data-driven smooth tests for a location-scale family revisited. *J. Statist. Theory and Practice* 3, 645-663.

Kallenberg, W.C.M., Ledwina, T., 1997. Data driven smooth tests for composite hypotheses: Comparison of powers. *J. Statist. Comput. Simul.* 59, 101–121.

LaRiccia, V.L., 1991. Smooth goodness-of-fit tests: a quantile function approach. *J. Amer. Statist. Assoc.* 86, 427–431.

Neyman, J., 1954. Sur une famille de tests asymptotiques des hypothèses statistiques composées. *Trabajos de Estadistica* 5, 161-168.

Neyman, J., 1959. Optimal asymptotic tests of composite statistical hypotheses. In *Probability and Statistics*, (ed. U. Grenander), Harald Cramér Volume, 212–234, Wiley, New York.

Rayner, J.C.W., Best, D.J., 1989. *Smooth Tests of Goodness of Fit*. Oxford University Press, New York.

Stephens, M.A., 1992. Tests of fit for logistic distribution based on EDF. Biometrika 66, 591–595.

Thomas, D.R., Pierce, D.A., 1979. Neyman's smooth goodness–of–fit test when the hypothesis is composite. *J. Amer. Statist. Assoc.* 74, 441–445.