Testul Fisher F cu valori semnificative mici în eşantioane de volum redus*

STUDIU DE CAZ

Drd. Alina BARBU
Academia de Studii Economice, Bucureşti

Abstract

Testarea diferenţelor între grupuri se realizează îndeosebi cu ajutorului testului Fisher F. În practică, acesta returnează uneori valori mici, dar semnificative. Situaţia a fost rareori discutată din punct de vedere teoretic, în general fiind pusă pe seama unor probleme în specificarea modelului sau în analiză.

Se prezintă un caz real în care au fost întâlnite valori F semnificative mici şi modul de tratare a acestora.

Cuvinte cheie: testul Fisher F; valori F semnificative mici; aditivitate; omiterea factorilor semnificativi; distribuţie normală; eşantioane de volum redus.

***

Testarea diferenţelor între grupuri: aspecte teoretice

Verificarea (testarea) ipotezelor statistice, ansamblu de metode ale statisticii inductive, permit, plecând de la date experimentale sau de la observaţii, precizarea formei legilor de repartiţie a variabilelor aleatoare referitoare la populaţiile considerate sau validitatea anumitor supoziţii referitoare la valorile parametrilor acestor legi. Testul statistic este criteriul pentru verificarea ipotezei statistice, constînd în calculul unei statistici şi stabilirea unei reguli fixate în prealabil de acceptare sau de respingere a ipotezei nule, H_0, cu o anumită probabilitate de a lua o decizie inexactă cînd H_0 este confruntată cu H_1. (Trebici, 1985)

* Articol elaborat ca parte a proiectului "Doctorat şi doctoranzi în triunghiul educaţie-cercetare-inovare (DOC-ECI) ", proiect cofinanţat din Fondul Social European prin Programul Operaţional Sectorial Dezvoltarea Resurselor Umane 2007-2013 şi coordonat de Academia de Studii Economice Bucureşti.

Pentru verificarea existenței diferențelor între două grupuri se pot folosi diferite teste statistice, în funcție de tipul variabilelor şi de tipul eşantioanelor analizate.

Teste statistice pentru verificarea diferențelor dintre grupuri

Tabel 1

	Eşantioane dependente	Eşantioane independente
Scala nominală	▪ Testul McNemar	▪ Fisher ▪ X^2
Scala ordinală	▪ Testul Wilcoxon	▪ Mann- Whitney
Scala de interval (raport)	▪ Testul t pentru medii	▪ Testul Z sau t pentru medii ▪ Regresie ▪ Testul Fisher F

Dintre acestea, testele t, Z, Fisher, X^2 sunt teste parametrice – validarea concluziilor luând în calcul premise fundamentale legate de forma distribuției variabilei. Performanța testelor parametrice în ipoteza nerespectării normalității distribuției a fost verificată de-a lungul timpului cu ajutorul Metodelor de simulare Monte-Carlo: fiind generat un număr mare de seturi de date, se verifică erorile şi abaterile în cazul în care datele nu sunt distribuite normal. S-a observat că testele parametrice nu au rezultate atât de slabe cum s-a crezut inițial, ceea ce nu înseamnă că ipoteza de normalitate a distribuției ar trebui ignorată.

În cazul aplicării testului Fisher – F în modele liniare, majoritatea abordărilor teoretice prezintă situația în care se înregistrează valori mari, semnificative, deoarece acestea permit respingerea ipotezei nule (respingerea ipotezei de echivalență a celor două eşantioane). Puțini autori au menționat situația în care valorile înregistrate sunt mici semnificative, mult mai mici decât ar trebui să fie dacă nu s-ar înregistra nicio legătură între variabile.

Abordări în literatura de specialitate: „Small *F*-ratios: Red Flags in the Linear Model”

Singurul articol care dezbate pe larg problema valorilor semnificative mici ale Testului F este „Small *F*-ratios: Red Flags in the Linear Model” (Meek, Ozgur şi Dunning, 2007). Autorii remarcă faptul că singurele referințe la valori mici ale testului F au apărut în propriile lucrări: Meek şi Turner (1983) prezintă un model bifactorial analizat ca unul unifactorial, concluzia fiind că valoarea scăzută a testului F indica omiterea unui factor semnificativ, iar Meek, Ozgur şi Dunning (2005) prezintă rezultate parțiale ale unei discuții mai largi pe tema valorilor F mici semnificative.

Meek, Ozgur şi Dunning (2007) susțin că sunt mai multe cauze ale apariției valorilor F mici în testarea de ipoteze bazată pe analiza varianței şi că orice astfel de situație trebuie investigată în detaliu. Autorii detaliază trei cazuri în care ele pot apărea:

Interacțiune în blocuri randomizate, exemplificată prin construirea unui experiment încrucişat fără replicare (*two-factor design without replication*) între 3 colegii şi 5 programe de studii. Cinci studenți sunt aleşi aleator din fiecare grupă şi este calculat scorul mediu pe care ei îl obțin la Graduate Management Admission Test (GMAT). Programul de studii înregistrează F=0,14, p=0,962 (1-p=0,038), valoare semnificativă neobişnuit de mică. Interacțiunea între rând (colegiu) şi coloană (program de studii) este verificată cu ajutorul testului Tukey pentru (non-)aditivitate, însă nu este observat un efect semnificativ de aditivitate (interacțiune).

Experimentul este redefinit cu repetare (*two-factor design with replication*): se aleg aleator 10 studenți de la fiecare colegiu şi se alocă aleator câte doi fiecărui program de studii. Este posibilă evaluarea independentă a interacțiunii colegiu-program de studii. Noul tabel ANOVA prezintă F=3,09, p=0,048 pentru programul de studii şi F=3,15, p=0,026 pentru efectul de interacțiune, ambele valori semnificative pentru α=0,05. Autorii concluzionează că valoarea F scăzută obținută în prima variantă indica o problemă cu modul în care experimentul a fost construit, dar că efectul interacțiunii nu este în mod necesar responsabil pentru valoarea F scăzută.

Omiterea factorilor semnificativi, exemplificată prin analiza numărului de zile petrecute de o femeie în spital după naştere, urmărind 4 spitale, fiecare cu 9 observații. Analiza unifactorială (variabila de influență: spitalul) returnează F=0,08, p=0,971 (1-p=0,029), valoare neobişnuit de mică. Datele sunt reprezentate grafic şi se observă gruparea acestora, în cadrul fiecărui spital, în două categorii; testul Hartlett *F* max permite acceptarea ipotezei variantelor egale între spitale, însă distribuțiile variabilelor nu par a respecta forma normală. Gruparea datelor în clustere distincte sugerează omiterea unui factor semnificativ.

Este introdus un factor suplimentar: tipul de naştere (naturală, cezariană sau asistată medical) şi analiza este reluată. Sunt evidențiate diferențe între tipurile de naştere (F=560,67, p=0,0) şi, la un prag de semnificație α=0,10, între spitale (F=2,83, p=0,06). În acest caz, omiterea factorului semnificativ a dus la apariția valorilor F mici, semnificative.

Neliniaritatea sau specificarea incorectă a modelului, exemplificată prin analiza vânzărilor de casete video (VHS) între 1995-2004 printr-un model liniar. Valoarea F obținută este 0,00, p=0,984 (1-p=0,016), iar reprezentarea grafică a datelor exprimă clar o distribuție neliniară a valorilor. Folosind un model (o ecuație) de gradul II, autorii obțin F=41,15, p=0,0, demonstrând că modelul inițial liniar nu era potrivit pentru datele disponibile.

Meek, Ozgur şi Dunning încheie cu observația că **valoarea testului F nu ar trebui niciodată să se apropie de 0; o valoare F apropiată de 0 şi semnificativă indică, cel mai probabil, o problemă în construcția modelului sau în analiza datelor** şi ar trebui investigată în aceeaşi măsură ca o valoare mare semnificativă. În afara cauzelor prezentate pe larg în articol, ei enumeră şi: nerespectarea ipotezelor legate de distribuție, multicoliniaritate în cazul regresiei şi/sau falsificarea datelor.

***

Investigarea apariției valorilor F mici a pornit de la două studii de caz reale din domeniul cercetării medicale: în două studii de piață, medicii au fost solicitați să evalueze performanța reprezentanților medicali ai unor companii farmaceutice pe o serie de atribute, folosind note de la 1 la 10. Fiecare eşantion a fost împărțit în două sub-eşantioane, în funcție de specialitatea medicală a respondenților, iar testul F a fost folosit pentru evaluarea diferențelor semnificative între specialități. Sunt redate evaluările medii; din motive de confidențialitate, nu s-au redat numele specialităților şi atributelor.

Rezultatele testului F pentru atributele A, B, C şi specialitățile 1 şi 2

Tabelul 2

medie	**Atributul A**	**Atributul B**	**Atributul C**
Specialitate 1 (N=56)	9,50	9,52	9,66
Specialitate 2 (N=33)	9,52	9,68	9,67
Total eşantion* (N=89)	9,51	9,57	9,66
Testul *F*	*F=0,01, p=0,086*	*F=1,32, p=0,253*	***F=0, p=0,037***

* Sunt răspunsuri complete; eşantionul cuprinzând şi non-răspunsurile numără 150 de respondenți. S-a ales eliminarea non-răspunsurilor din analiză pentru a spori acuratețea, susținut de faptul că eşantionul redus cuprinde > 50 de respondenți.

Studiu de caz 2: rezultatele testului F pentru atributele D, E şi specialităţile 3 şi 4

Tabelul 3

medie	**Atributul D**	**Atributul E**
Specialitate 3 (N=30)	8,79	8,74
Specialitate 4 (N=21)	8,75	8,75
Total eşantion* (N=51)	8,77	8,74
Testul *F*	*F=0,01, p=0,077*	***F=0, p=0,02***

Sunt evidenţiazte două aspecte:

- Eşantioanele sunt mici (89, respectiv 51 de evaluări), fiind o caracteristică a cercetării prin sondaj în domeniul medical comparativ cu domeniul „*consumer*” (totalitatea consumatorilor unui produs sau serviciu).

- Evaluarea pe o scală de la 1 la 10 nu este recomandată de teoria cercetării de piaţă, deoarece nu diferenţiază destul de bine între mai multe atribute; în exemplele anterioare, rezultă că evaluările medii sunt foarte apropiate.

Valori foarte scăzute ale testelor F se înregistrează pe atributele A, C, D şi E, însă doar pentru atributele C şi E există semnificaţie la pragul 0,05, indicată de valoarea $p \leq 0{,}05$. Situaţia este neobişnuită şi se încadrează, potenţial, în categoria „*red flag*” identificată de Meek, Ozgur şi Dunning (2007). Înainte de a considera că aceste valari sunt întâmplătoare, s-a verificat în ce măsură ele pot fi justificate de soluţiile identificate de autori.

Omiterea factorilor semnificativi din model

Nu există o metodă universal valabilă de a detecta factori semnificativi omişi, deoarece în realitate relaţiile dintre variabile sunt complexe şi nu întotdeauna uşor de observat. În cele două studii prezentate, dispunem de o singură altă variabilă socio-demografică, localitatea, cu trei valori: Bucureşti, oraşe mari, oraşe medii. În continuare s-a testat dacă aceasta influenţează semnificativ evaluarea medicilor.

Pentru un prag de semnificaţie $\alpha = 0{,}05$, testul F indică:

- O legătură semnificativă între localitate şi evaluările pe atributul A (F=3,87, p=0,024).
- Absenţa unei legături semnificative între localitate şi evaluările pe atributul C (F=2,45, p=0,09), însă relaţia devine semnificativă

* 31 de răspunsuri valide; din cauza volumului foarte mic al eşantionului, valorile lipsă au fost înlocuite prin mediere globală pentru a diminua pierderea de informaţie. Sunt cunoscute implicaţiile ratei mari a non-răspunsurilor asupra validităţii concluziilor. Un eşantion N=31 ar fi prea redus pentru orice fel de test.

dacă pragul de semnificație este coborât la 0,10. Eşantionul este mic şi distribuit neuniform: N=45 în Bucureşti, N=87 în oraşe mari, N=18 în oraşe medii (deoarece nu a fost construit pentru a fi reprezentativ la nivel regional). Nu se recomandă coborârea pragului de semnificație.

- Absența unei legături semnificative între localitate şi evaluările pe atributul D (F=0,44, p=0,345).
- Absența unei legături semnificative între localitate şi evaluările pe atributul E (F=0,3, p=0,25).

Încadrarea valorilor F în limite „normale”, ar putea însemna că localitatea este un factor explicativ mai bun decât specialitatea. **Absența legăturilor semnificative generale între localitate şi evaluări nu permite acceptarea ipotezei omiterii factorilor semnificativi** ca explicație a valorilor F semnificative mici.

Interacțiunea (aditivitatea)

Aditivitatea reprezintă proprietatea variabilelor independente de a interacționa semnificativ pentru a influența o variabilă dependentă. Pe lângă influența individuală a variabilelor independente se adaugă influența interacțiunii. Există o serie de teste care verifică aditivitatea, printre care: Tukey, Mandel, Johnson-Graybill, *locally best invariant* (LBI) şi Tussel. Šimečková, Šimeček şi Rasch (2008) compară aceste cinci teste pentru a observa riscul de tip I înregistrat efectiv pe baza simulărilor pe calculator.

Testul Tukey (1949) este una dintre cele mai cunoscute opțiuni pentru a verifica aditivitatea; se urmăreşte un tip specific de aditivitate, de tipul celei în care interacțiunea este proporțională cu efectul rândului şi coloanei. Šimečková, Šimeček şi Rasch argumentează că este foarte potrivit pentru detectarea acestui tip de interacțiune, mai puțin de 4% dintre simulări înregistrând valori α>0,05. Statistica testului este un raport între media pătratelor efectelor interacțiunii şi erori:

$$S_T = \frac{MS_{interactiune}}{MS_{eroare}}$$

Statistica testului Tukey este F-distribuită cu *1; (a-1)(b-1)* grade de libertate, unde a = numărul de rânduri şi b = numărul de coloane din model.

Modelul inițial este unifactorial, ipoteza interacțiunii nu este aplicabilă. În schimb, s-a testat dacă un model încrucişat specialitate-localitate poate fi încadrat în această situație. Nu se recomandă testarea aditivității pentru cazul 2, deoarece:

- Distribuția răspunsurilor pe rânduri (localitate) şi coloane (specialitate) generează foarte puține valori în fiecare grupă – 4 din

6 grupe conţin cel mult 6 răspunsuri.
- La punctul anterior s-a observat că nu există legătură semnificativă între localitate şi evaluările pe atributele D şi E.

Pentru cazul 1, s-a obţinut următoarea distribuţie a răspunsurilor:

număr de respondenţi	**Specialitate 1**	**Specialitate 2**	**Total**
Bucureşti	16	8	24
Oraşe mari	37	18	55
Oraşe medii	3	7	10
Total	**56**	**33**	**89**

medie	**Specialitate 1**	**Specialitate 2**	**Total**
Bucureşti	9,81	9,75	**9,79**
Oraşe mari	9,62	9,78	**9,67**
Oraşe medii	9,33	9,29	**9,30**
Total	**9,66**	**9,67**	**9,66**

Prin analiza ANOVA se susţine faptul că efectul localităţii este la limita semnificaţiei pentru un prag α=0,05 (F=17,03, p=0,055, F_{critic}=19), iar specialitatea înregistrează o valoare F mică, nesemnificativă (F=0,04, p=0,84).

Testul Tukey: T=-3207,14, valoare inferioară în modul celei critice $F_{0,01;1,1}$=4052,18 – dar nu şi $F_{0,05;1,1}$=161,45. Ţinând cont de dimensiunea sub-eşantioanelor, rezultă că, pe baza testului Tukey, susţinut de analiza ANOVA, **nu există suficiente dovezi pentru a accepta existenţa interacţiunilor între variabile**.

Folosirea unui model incorect

Ţinând cont de evaluările medii înregistrate, este puţin probabil să existe o legătură neliniară între specialitate şi evaluare. Poate fi o problemă însă aplicarea testului (parametric) F dacă distribuţiile evaluărilor nu respectă ipoteza de normalitate – în special în eşantioane mici, deoarece pentru eşantioanele mari, Teorema Limită Centrală garantează faptul că media variabilei va urma o distribuţie normală chiar dacă variabila din eşantion nu urmăreşte distribuţia normală.

Normalitatea distribuţiei este de obicei verificată cu ajutorul testelor:

- Kolmogorov-Smirnov, care nu mai necesită prezentare;

- Shapiro-Wilk W, tot mai popular datorită puterii bune comparativ cu alte teste de normalitate, în special în eşantioane medii şi mici (Shapiro şi Wilk, 1965; Shapiro şi Wilk, 1968). În particular, puterea testului W este mai mare decît a testului Kolmogorov-Smirnov deoarece detectează inclusiv deviații cauzate de asimetrie (*skewness*) şi aplatizare (*kurtosis*). Relația de calcul a testului:

$$W = \frac{\left(\sum_{i=1}^{n} c_i X_i\right)^2}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

Coeficienții c_i şi valorile critice ale statisticii W sunt constante, tabelate inițial de Shapiro şi Wilk până la n=50. În prezentarea originală, testul W nu a fost extins pentru eşantioane mai mari de n=50 din cauza calculului anevoios al coeficienților, dar extensia Shapiro-Francia oferă coeficienți pentru $50 \le n \le 99$. Nici în prezent testul nu este implementat de toate pachetele statistice, fiind dificil de calculat*. Burdenski (2000) recomandă testul Kolmogorov-Smirnov pentru eşantioane mai mari de 25 de unități şi testul Shapiro-Wilk pentru eşantioane mai mici de 25 de unități. Testul Kolmogorov-Smirnov ar trebui să fie aplicabil în ambele studii de caz.

Testul Kolmogorov-Smirnov indică deviații de la distribuția normală pentru atributele A şi C (89 răspunsuri complete), dar normalitate pentru atributele D şi E (31 răspunsuri complete). Deşi eşantionul este mai mare de 25 de unități, valorile F mici indică o potențială problemă a aplicării testelor parametrice pentru ultimele două atribute. Se verifică normalitatea distribuției atributelor D şi E cu ajutorul testului Shapiro-Wilk.

Verificarea normalității distribuțiilor pentru atributele D şi E folosind testul Shapiro-Wilk

Tabelul 4

	Atributul D	Atributul E
$W_{calculat}$	0,834	0,830
W_{critic} pentru α=0,05	0,902	0,902
Decizie	*Nu este distribuție normală*	*Nu este distribuție normală*

Pentru toate atributele care înregistrează valori F mici se constată abateri de la distribuția normală. În ciuda recomandării lui Burdenski, testul de normalitate Kolmogorov-Smirnov nu se dovedeşte adecvat în cazul eşantionului 2 (N=31 răspunsuri complete), fiind preferabil testul Shapiro-Wilk.

* O versiune online a testului Shapiro-Wilk este disponibilă la adresa: http://dittami.gmxhome.de/shapiro/. Este suficientă copierea datelor , iar rezultatul, inclusiv valorile critice şi decizia de acceptare/ respingere, este calculat automat. Conform testului original, eşantioanele acceptate sunt cuprinse între 5 şi 50 de unități.

Teorema Limită Centrală nu poate fi aplicată pentru aceste eşantioane, întrucât ele sunt de volum mediu şi mic. Este recomandată folosirea testelor neparametrice pentru evaluarea atributelor cu probleme: A, C, D şi E:

Testul Mann-Whitney U cel mai uzual test neparametric pentru variabile ordinale (s-a folosit o scală „inferioară"), calculat pe baza sumei rangurilor înregistrate de fiecare sub-eşantion:

$U = \max\left(n_1 n_2 + \frac{n_i\left(n_i + 1\right)}{2} - R_i \right)$, unde suma rangurilor este $R_i = \sum_{j=1}^{n_i} r_{ji}$ şi i=1,2.

Ipoteza nulă se respinge dacă valoarea calculată $U' > Ucritic$ $U' > U_{critic}$. Sunt sintetizate rezultatele aplicării testului Mann-Whitney pentru atributele A, C, D şi E. (tabelul următor)

Testul median este o altă opţiune neparametrică, fiind calculat pe baza numărului de observaţii mai scăzute şi mai ridicate decât mediana din fiecare grup:

$$X^2 = \sum_{i=1}^{2} \frac{\left(n_{1i} - n_{1i}^{'}\right)^2 + \left(n_{2i} - n_{2i}^{'}\right)^2}{n_{1i}} \approx X_1^2$$

Testul median este mai slab decât Mann-Whitney, dar este mai potrivit dacă evaluările conţin multe valori extreme (evaluări cu notă 10).

Testarea diferenţelor cu ajutorul testului Mann-Whitney

Tabelul 5

	Atributul A	Atributul C	Atributul D	Atributul E
N	*89*	*89*	*31*	*31*
U'	*911*	*890*	*103.5*	***124.5***
(U_{critic})	*(<963)*	*(<990)*	*(<128.5)*	***(>107)***

Pentru **atributele A, B, C, D nu se poate spune că există diferenţe semnificative** între cele două specialităţi medicale.

Pentru atributul E, testul F nu a evidenţiat existenţa unei diferenţe între evaluări, însă testul Mann-Whitney arată diferenţe semnificative. Testul median nu este aplicabil deoarece, în urma eliminării valorilor egale cu mediana, rămân mai puţin de 20 de valori volum minim de date cu care poate opera testul. Valorile efective (8,74 şi 8,75) nu susţin existenţa diferenţelor. **Pentru atributul E rezultatele statistice sunt neconcludente** – în principal, din cauza eşantionului mic - de aceea nu se poate admite existenţa unor diferenţe semnificative.

***

Examinarea minuțioasă a celor două studii de caz nu este justificată din cauza evaluărilor foarte apropiate şi a dimensiunii reduse a eşantioanelor. Analiza se bazează pe studii de caz reale, situații care nu sunt acoperite de teoria statistică. Studiile provin din domeniul medical, unde este mai comună abordarea unor eşantioane de volum redus deoarece populația de proveniență este, la rândul său, mai redusă. Clarificarea acestor situații în teorie conduce la completarea unei lacune în verificarea ipotezelor statistice (potențiala problemă a testului F în eşantioane medii şi mici) şi la o mai bună abordare a apariției valorilor F semnificative mici în practică.

Singura lucrare teoretică ce dezbate pe larg problema valorilor F semnificative mici este a lui Meek, Ozgur şi Dunning (2007). S-a atras atenția că valorile F mici pot ascunde probleme ale modelului dacă sunt semnificative şi ar trebui verificate cauzele posibile înainte de a le considera valori întâmplătoare. Sugestiile identificate de autori sunt adaptate în studiul de caz prezentat pentru a identifica şi trata cauza valorilor F mici. Rezultă că principala cauză a valorilor F mici a fost aplicarea testului F pe date ce nu urmăresc distribuția normală, provenite din eşantioane medii-mici. S-a observat şi că testul Shapiro-Wilk este mai potrivit decât Kolmogorov-Smirnov pentru a testa normalitatea în eşantioane de volum redus (N<50, pentru siguranță).

Problema valorilor F mici nu este lămurită în totalitate. Se poate suspecta faptul că însăşi dimensiunea redusă a eşantionului poate afecta testele statistice: în cazul prezentat, N=51 şi N=89, iar Meek, Ozgur şi Dunning au prezentat cazuri cu 10-50 de observații. În eşantioanele medii şi mici nu se poate aplica Teorema Limită Centrală, iar aceasta poate fi cauza principală a valorilor F mici. În eşantioane de volum redus ar trebui aplicate testele neparametrice (ex. Mann-Whitney) ca regulă generală.

Concluzii

Se recomandă pentru cercetările viitoare în acest domeniu orientarea asupra a trei probleme esențiale:

- Identificarea frecvenței de apariție a valorilor F mici semnificative în eşantioane mai mari, de exemplu N>100-150;
- Dacă valorile F mici semnificative apar şi în acest caz, interesează evaluarea normalității distribuției şi compararea testelor Kolmogorov-Smirnov şi Shapiro-Wilk;
- În cazul în care variabilele nu urmăresc distribuția normală, este indicată evaluarea aplicabilității Teoremei Limită Centrală când sunt deja înregistrate valori F mici.

În cazul în care se constată apariţia valorilor F mici doar în cazul eşantioanelor de volum redus, putem considera aplicarea exclusivă a testelor neparametrice; trebuie însă identificat şi volumul-limită al eşantionului pentru care valorile F mici încetează să apară.

Bibliografie

- **Burdenski, T.** (2000), ***Evaluating Univariate, Bivariate and Multivariate Normality using Graphical and Statistical Procedures***. Multiple Linear Regression Viewpoints, nr 26
- **Meek, G., Ozgur, C., Dunning, K.** (2005). ***Some implications of significantly small F-ratios***. Proceedings of the 2005 Annual National Meeting of the Decision Sciences Institute
- **Meek, G., Ozgur, C., Dunning, K.** (2007). ***Small F-ratios: Red Flags in the Linear Model***. Journal of Data Science, nr. 5,
- **Meek, G., Turner, S.** (1983). ***Statistical Analysis for Business Decisions***. Boston: Houghton Mifflin.
- **Shapiro, S.S., Wilk, M.B.** (1965). ***An analysis of variance test for normality (complete samples).*** Biometrika, nr. 52
- **Shapiro, S.S., Wilk, M.B., Chen, H.J.** (1968). ***A comparative study of various tests for normality***. Journal of the American Statistical Association, nr. 63
- **Šimečková, M., Šimeček, P., Rasch, D. (**2008) ***Tests of Additivity in two-way ANOVA Models with Single Subclass Numbers***, Statistical Papers of the Union of Czech Mathematicians and Physicists – JČMF
- **Trebici, V.** (coord) (1985), ***Mică Enciclopedie de Statistică***, Bucureşti: Editura Ştiinţifică şi Enciclopedică
- **Tukey, J.W.** (1949). ***One Degree of Freedom for Non-Additivity***. Biometrics, nr. 5