DIDACTICS OF MATHEMATICS

8(12)

The Publishing House
of Wrocław University of Economics
Wrocław 2011

Referee
Henryk Zawadzki
(University of Economics in Katowice)

Copy-editing
Dorota Pitulec

Proof-reading
Barbara Łopusiewicz

Typesetting
Elżbieta Szlachcic

Cover design
Robert Mazurczyk

Front cover painting: W. Tank, Sower
(private collection)

This publication is available at: www. journal.ue.wroc.pl and www.ibuk.pl.
Abstracts of published papers are available in the international database
The Central European Journal of Social Sciences and Humanities
http://cejsh.icm.edu.pl

Information on submitting and reviewing paper is available
on the Publishing House's website www.wydawnictwo.ue.wroc.pl

ISSN 1733-7941

The original version: printed
Printing: Printing House TOTEM
Print run: 200 copies

TABLE OF CONTENTS

D I D A C T I C S O F M A T H E M A T I C S

No. 8 (12) 2011

ROZKŁAD STATYSTYKI *T*-STUDENTA PRZY DANEJ WARIANCJI Z PRÓBY O ROZKŁADZIE NORMALNYM

Albert Gardoń

Abstract. It is not so easy to lecture on higher mathematics for economy students. Advanced notions must be often presented for people without an appropriate theoretical background, which forces the teacher to simplify. Unfortunately, the praxis shows that the frontier between a simplification and a factual error is often very subtle and it happens this frontier is sometimes crossed. Such a situation occurs just in the case of the problem, which will be described in this paper. It is a known fact that the so called Student's T statistics from a normal distributed sample is t-Student distributed, without any doubt. But in handbooks for economy students several authors try to use this statistics for exercises with mathematical tables of the t-Student distribution, ordering a calculation of the probability that the sample average will belong to the given interval, in the case when the theoretical variance is unknown but the sample variance has been calculated. Unfortunately, such a situation has nothing to do with the t-Student distribution and this error is systematically copied in successive handbooks.

Keywords: Student's T statistics, t-Student distribution, normal distribution.

1. Wstęp

Niech standardowo będzie dana przestrzeń probabilistyczna (Ω, F, P), gdzie Ω jest tzw. przestrzenią zdarzeń, czyli zbiorem wszystkich zdarzeń elementarnych ω, $\boldsymbol{F}$ jest tzw. σ-ciałem zbiorów mierzalnych, czyli zbiorem wszystkich zdarzeń losowych A, a więc takich podzbiorów $A \subset \Omega$, dla których możliwe jest określenie ich miary, natomiast P jest miarą probabilistyczną określoną na tej przestrzeni, zwaną krótko prawdopodobieństwem. Zakłada się, że czytelnik zna i rozumie formalne definicje powyższych i innych pojęć probabilistycznych (jak np. zmienna losowa, rozkład, gęstość), dlatego nie będą tu one szczegółowiej objaśniane, można się z nimi zapoznać dokładniej np. w monografii (Billingsley, 2009).

Albert Gardoń
Department of Statistics, Wrocław University of Economics, Komandorska Street 118/120, 53-345 Wrocław, Poland.
E-mail: albert.gardon@ue.wroc.pl

Dalej rozważana będzie n-elementowa próba prosta $\mathbb{X}$ z rozkładu normalnego. Przed dokonaniem obserwacji wektor wartości dla tej próby nie jest znany i formalnie jest traktowany jak wektor lub ciąg niezależnych zmiennych losowych o jednakowym rozkładzie normalnym, co można zapisać w następujący sposób:

$$\mathbb{X} = (X_i)_{i=1}^{n} \ iid \ \sim N(\mu, \sigma) .$$

Wykorzystując normalność próby, można określić rozkłady poniższych popularnych statystyk próbkowych, oznaczanych tu przez Z, T i H:

$$Z = \frac{\overline{X} - \mu}{\sigma} \sqrt{n} \sim N(0,1) , \tag{1}$$

$$H = \frac{nS^2}{\sigma^2} \quad \sim \chi^2(n-1) = \Gamma\left(\frac{n-1}{2}, \frac{1}{2}\right), \tag{2}$$

$$T = \frac{\overline{X} - \mu}{S} \sqrt{n-1} \quad \sim t(n-1) , \tag{3}$$

gdzie $\overline{X}$ i S^2 są odpowiednio średnią arytmetyczną i wariancją (w sensie największej wiarygodności) z próby:

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i ,$$

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

Jak to zostało zapisane, statystyka Z, przy znanych parametrach μ i σ, ma standardowy rozkład normalny, co wynika z podstawowych własności rozkładu normalnego: suma zmiennych o rozkładzie normalnym ma dalej rozkład normalny, podobnie centrowanie i normalizowanie nie powoduje zmiany typu rozkładu w tym przypadku. Druga ze statystyk, H, przy znanym parametrze σ, ma rozkład χ^2 o $(n-1)$ stopniach swobody (zob. (Pawłowski, 1980)) lub ogólniej rozkład Γ z parametrami $\frac{n-1}{2}$ i $\frac{1}{2}$. Ostatnia statystyka, nazywana statystyką T Studenta, ma, przy znanym parametrze μ, rozkład t-Studenta o $(n-1)$ stopniach swobody, co wynika niemal bezpośrednio z jednej z definicji tego rozkładu (zob. również

(Pawłowski, 1980)). Mówi ona, że rozkład t-Studenta o danej liczbie stopni swobody ma zmienna losowa będąca ilorazem dwóch niezależnych zmiennych losowych, pomnożonym przez pierwiastek kwadratowy z liczby stopni swobody; licznik wspomnianego ilorazu musi być zmienną losową o standardowym rozkładzie normalnym, a jego mianownik zmienną losową o rozkładzie χ o danej liczbie stopni swobody[1].

2. Problem

Dotychczasowe stwierdzenia nie budzą wątpliwości. Problem pojawia się jednak, gdy należy określić rozkład statystyki T przy znanym parametrze μ, nieznanym parametrze σ, ale danej wariancji z próby $S^2 = s^2$. W wielu podręcznikach do statystyki dla uczelni ekonomicznych (zob. np. (Kassyk-Rokicka, 2001); (Ostasiewicz, Rusnak, Siedlecka, 2004); (Wawrzynek, 2007); (Witkowski, 2010)) rozkład t-Studenta jest wykorzystywany w zadaniach do obliczania prawdopodobieństwa, że średnia arytmetyczna $\overline{X}$ z niezbyt licznej ($n < 30$) próby prostej o rozkładzie normalnym, ze znaną wartością oczekiwaną μ i nieznaną wariancją, znajdzie się w określonym przedziale, dla ustalenia uwagi niech to będzie półprosta $(-\infty, u]$. Prezentowane tam „rozwiązania” wyglądają mniej więcej tak:

$$P(\overline{X} \le u) = P\left(\frac{\overline{X}-\mu}{S}\sqrt{n-1} \le \frac{u-\mu}{S}\sqrt{n-1}\right) \overset{sic!}{=}$$

$$\overset{sic!}{=} P\left(T \le \frac{u-\mu}{s}\sqrt{n-1}\right) = F_T\left(\frac{u-\mu}{s}\sqrt{n-1}\right),$$

gdzie F_T oznacza dystrybuantę statystyki T, czyli dystrybuantę rozkładu t-Studenta o (n–1) stopniach swobody. Pierwsze przejście polega na zamianie zmiennej losowej $\overline{X}$ na T w wyniku wykonania tego samego przekształcenia na obu stronach nierówności definiującej zdarzenie losowe. Błąd popełniany jest w drugim kroku, a niewłaściwe przejście zostało zaznaczone nad odpowiednim znakiem równości, która w tym przypadku nie zachodzi! Przyczyną takiego stanu jest niekonsekwentne potraktowanie

[1] Jeśli zmienna losowa ma rozkład χ^2, to jej pierwiastek kwadratowy ma rozkład χ o tej samej liczbie stopni swobody.

informacji o znanej dyspersji z próby. Z lewej strony warunku traktowana jest ona jako zmienna losowa $S(\omega)$, współtworząca zmienną losową $T(\omega)$, natomiast z prawej – jako liczba s. Niestety, jest to niedozwolone! Nie można utożsamiać funkcji z jedną z jej wartości! I niczego nie zmienia tu fakt, że przy normalnej próbie prostej ma miejsce niezależność: $\overline{X} \perp\!\!\!\perp S$ (zob. (Pawłowski, 1980)). Zresztą, w powyższym „rozwiązaniu" nawet nie da się zauważyć momentu, gdzie się z tej niezależności korzysta, co zresztą jest kolejnym niedociągnięciem. A przecież nawet na podstawowych kursach rachunku prawdopodobieństwa uczy się studentów, że wszelkie informacje dane przed obliczaniem zadanego prawdopodobieństwa należy umieszczać w warunku.

3. Rozwiązanie

Aby rozwiązać powstały problem, należy wrócić do podstaw. Niech

$$A = \left\{ \omega \in \Omega : T(\omega) \le \frac{u-\mu}{S(\omega)} \sqrt{n-1} \right\},$$

$$B = \{\omega \in \Omega : S(\omega) = s\}. \tag{4}$$

Powstaje pytanie, jaki jest rozkład statystyki $\overline{X}$, gdy znana jest wariancja z próby $S^2 = s^2$ lub równoważnie – jaki jest rozkład statystyki T, pod warunkiem że $S = s$. Informacja, że znana jest dyspersja z próby i wynosi s, oznacza, że należy ograniczyć rozważania dotyczące prawdopodobieństwa P zdarzenia A w przestrzeni Ω tylko do tych zdarzeń elementarnych ω, dla których $S(\omega) = s$, czyli do zbioru $A \cap B$. Innymi słowy, należy rozważyć rozkład zmiennej losowej T, uwzględniając nie całą dziedzinę tej funkcji, czyli Ω, lecz jedynie zbiór B. W pewnym sensie można powiedzieć, że przestrzeń Ω w wyniku tej informacji zostaje ograniczona do zbioru B, σ-ciało $\boldsymbol{F}$ do jego przekroju ze zdarzeniem B, a miara P zastąpiona prawdopodobieństwem warunkowym $P(\cdot \mid B)$. W pewnym sensie, gdyż $P(B) = 0$ (ze względu na ciągłość rozkładu statystyki S), więc nie można bezpośrednio zastosować klasycznego wzoru na prawdopodobieństwo warunkowe. Niemniej jednak w nowoczesnej teorii prawdopodobieństwa, opartej na teorii miary, warunkowanie względem zdarzeń prawie niemożliwych nie stanowi już zapory nie do przejścia (zob.

(Billingsley, 2009)). Zatem powinno to być obliczane w następujący sposób:

$$P(\overline{X} \le u \mid S = s) \overset{\overline{X} \perp\!\!\perp S}{=} P(\overline{X} \le u) = P\left(Z \le \frac{u-\mu}{\sigma}\sqrt{n} \right) = \Phi\left(\frac{u-\mu}{\sigma}\sqrt{n} \right),$$

gdzie Z dane jest równaniem (1), a Φ to dystrybuanta standardowego rozkładu normalnego. Pierwsze przejście jest wynikiem niezależności $\overline{X} \perp\!\!\perp S$, w kolejnych dokonuje się już tylko standaryzacji zmiennej losowej o rozkładzie normalnym. Równoważnie, odnosząc wyniki do statystyki T Studenta, można zapisać:

$$T = \sqrt{\frac{(n-1)\sigma^2}{nS^2}} \cdot Z \quad \overset{Z \perp\!\!\perp S}{\Rightarrow} \quad T|_{S=s} \sim N\left(0, \sqrt{\frac{(n-1)\sigma^2}{ns^2}} \right), \tag{5}$$

ponieważ drugi czynnik, Z, ma standardowy rozkład normalny, a pierwszy, który jest od niego niezależny, przy podanym warunku staje się stałą skalującą jedynie rozkład normalny. Podsumowując: warunkowy rozkład statystyki T Studenta przy danym $S = s$ jest normalny ze średnią 0 i wariancją $\frac{(n-1)\sigma^2}{ns^2}$.

4. Gęstość warunkowa statystyki *T* Studenta

Powyższy wynik można uzasadnić jeszcze bardziej formalnie, wyznaczając warunkową gęstość zmiennej losowej $T|_{S=s}$. W tym celu wystarczy, korzystając z niezależności $Z \perp\!\!\perp H$, gdzie zmienne te dane są odpowiednio wzorami (1) i (2), przekształcić wektor losowy (Z, H) w następujący sposób (zob. (Feller 2007)):

$$(Z,H) \xrightarrow{\xi} (T,W) = \left(\frac{\overline{X}-\mu}{S}\sqrt{n-1}\,,\, S^2 \right). \tag{6}$$

Wtedy, jeśli przez f_{ZH} oznaczy się gęstość łączną tego wektora, co można zapisać $(Z,H) \sim f_{ZH}$, to:

$$(T,W) = \xi(Z,H) \sim f_{TW}(t,w) = f_{ZH}\big(z(t,w), h(t,w)\big) \left| \det J_{\xi^{-1}}(t,w) \right|, \tag{7}$$

gdzie f_{TW} jest gęstością łączną wektora (T,W), a J oznacza jakobian przekształcenia, który w tym przypadku dany jest wzorem:

$$J_{\xi}(z,h)=\begin{bmatrix}\frac{\partial t(z,h)}{\partial z} & \frac{\partial t(z,h)}{\partial h}\\ \frac{\partial w(z,h)}{\partial z} & \frac{\partial w(z,h)}{\partial h}\end{bmatrix} \Rightarrow J_{\xi^{-1}}(t,w)=\begin{bmatrix}\frac{\partial z(t,w)}{\partial t} & \frac{\partial z(t,w)}{\partial w}\\ \frac{\partial h(t,w)}{\partial t} & \frac{\partial h(t,w)}{\partial w}\end{bmatrix}. \quad (8)$$

Ostatecznie wystarczy wyznaczyć gęstość warunkową $f_{T|W=s^2}$ za pomocą formuły:

$$T\,|_{S^2=s^2}=T\,|_{W=s^2}\sim f_{T|W=s^2}(t)=\frac{f_{TW}(t,s^2)}{f_W(s^2)}, \quad (9)$$

gdzie

$$f_W(w)=\int_{\Re} f_{TW}(t,w)dt$$

jest gęstością brzegową zmiennej losowej W.

Tyle ogólnej teorii dotyczącej przekształceń wektorów losowych. W tym konkretnym przypadku zmienne losowe Z i H, dane równaniami (1) i (2), mają następujące gęstości:

$$f_Z(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}},$$

$$f_H(h)=\frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\Gamma(\frac{n-1}{2})}h^{\frac{n-1}{2}-1}e^{-\frac{h}{2}}\mathbb{I}_{\Re_+}(h),$$

gdzie Γ to tzw. funkcja specjalna gamma Eulera:

$$\Gamma(u)=\int_0^{+\infty} y^{u-1}e^{-y}dy,$$

natomiast $\mathbb{I}$ to tzw. funkcja charakterystyczna (indykator) zbioru:

$$\mathbb{I}_A(u)=\begin{cases}1 & : \ u\in A\\ 0 & : \ u\notin A\end{cases}.$$

Z niezależności tych zmiennych losowych wynika, że gęstość łączna wektora (Z,H) jest iloczynem gęstości brzegowych, czyli:

$$f_{ZH}(z,h) = f_Z(z)\,f_H(h) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}} \cdot \frac{\left(\frac{1}{2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} h^{\frac{n-1}{2}-1} e^{-\frac{h}{2}} \mathsf{I}_{\Re_+}(h) =$$

$$= \sqrt{\frac{h^{n-3}}{2^n \pi}}\, \frac{e^{-\frac{z^2+h}{2}}}{\Gamma\left(\frac{n-1}{2}\right)} \mathsf{I}_{\Re_+}(h)\,.$$

Wyrażając „stare” zmienne (Z,H) za pomocą „nowych” (T,W), przy użyciu zależności (6), otrzymuje się przekształcenie ξ^{-1}:

$$Z = T\sqrt{\frac{nW}{(n-1)\sigma^2}}\,, \qquad H = \frac{nW}{\sigma^2}\,,$$

a następnie, wykorzystując (8), wyznacznik jego jakobianu:

$$\det J_{\xi^{-1}}(t,w) = \det\begin{bmatrix} \sqrt{\frac{nw}{(n-1)\sigma^2}} & \frac{\partial z(t,w)}{\partial w} \\ 0 & \frac{n}{\sigma^2} \end{bmatrix} = \sqrt{\frac{n^3 w}{(n-1)\sigma^6}}\,.$$

Pozwala to na obliczenie gęstości łącznej wektora (T,W) zdefiniowanej równaniem (7):

$$f_{TW}(t,w) = f_{ZH}\left(t\sqrt{\frac{nw}{(n-1)\sigma^2}}\,, \frac{nw}{\sigma^2}\right)\left|\sqrt{\frac{n^3 w}{(n-1)\sigma^6}}\right| =$$

$$= \sqrt{\frac{(nw)^{n-2}}{2^n \pi (n-1)\sigma^{2(n-2)}}}\; \frac{e^{-\frac{nw}{2(n-1)\sigma^2}(t^2+n-1)}}{\Gamma\left(\frac{n-1}{2}\right)} \mathsf{I}_{\Re_+}(w)$$

Dalej, całkując f_{TW} po pierwszej zmiennej, otrzymuje się gęstość brzegową zmiennej losowej W:

$$f_W(w) = \int_{\Re} \sqrt{\frac{(nw)^{n-2}}{2^n \pi (n-1) \sigma^{2(n-2)}}} \frac{e^{-\frac{nw}{2\sigma^2}}}{\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{t^2}{2\frac{(n-1)\sigma^2}{nw}}} \mathbf{I}_{\Re_+}(w)\, dt =$$

$$= \sqrt{\frac{(nw)^{n-2}}{2^n \pi (n-1) \sigma^{2(n-2)}}} \frac{e^{-\frac{nw}{2\sigma^2}}}{\Gamma\left(\frac{n-1}{2}\right)} \underbrace{\left(\int_{\Re} e^{-\frac{t^2}{2\frac{(n-1)\sigma^2}{nw}}} dt \right)}_{=\sqrt{\frac{2\pi(n-1)\sigma^2}{nw}}} \mathbf{I}_{\Re_+}(w) =$$

$$= \sqrt{\frac{(nw)^{n-3}}{2^{n-1} \sigma^{2(n-3)}}} \frac{e^{-\frac{nw}{2\sigma^2}}}{\Gamma\left(\frac{n-1}{2}\right)} \mathbf{I}_{\Re_+}(w) .$$

Ostatnie przejście bierze się z faktu, że całka umieszczona w nawiasie jest, z dokładnością do stałej $\sqrt{\frac{nw}{2\pi(n-1)\sigma^2}}$, całką po całej prostej rzeczywistej z gęstości rozkładu $\sim N\left(0, \sqrt{\frac{(n-1)\sigma^2}{nw}}\right)$. Ostatecznie zastosowanie wzoru (9) na gęstość warunkową daje następujący wynik:

$$f_{T|W=s^2}(t) = \frac{\sqrt{\frac{(ns^2)^{n-2}}{2^n \pi (n-1) \sigma^{2(n-2)}}} \frac{e^{-\frac{ns^2}{2(n-1)\sigma^2}(t^2+n-1)}}{\Gamma\left(\frac{n-1}{2}\right)} \overbrace{\mathbf{I}_{\Re_+}(s^2)}^{=1}}{\sqrt{\frac{(ns^2)^{n-3}}{2^{n-1} \sigma^{2(n-3)}}} \frac{e^{-\frac{ns^2}{2\sigma^2}}}{\Gamma\left(\frac{n-1}{2}\right)} \underbrace{\mathbf{I}_{\Re_+}(s^2)}_{=1}} =$$

$$= \sqrt{\frac{ns^2}{2\pi(n-1)\sigma^2}}\, e^{-\frac{ns^2 t^2}{2(n-1)\sigma^2}} = \frac{1}{\sqrt{2\pi \frac{(n-1)\sigma^2}{ns^2}}} e^{-\frac{t^2}{2\frac{(n-1)\sigma^2}{ns^2}}} .$$

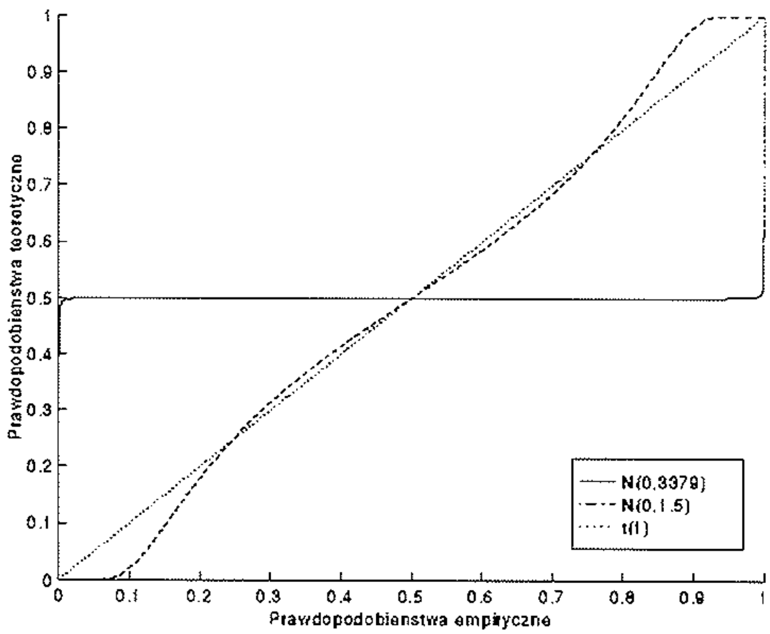
W powyższej formule indykator $\mathsf{I}_{\Re_+}(s^2)=1$, ponieważ zaobserwowana wartość wariancji z próby $s^2 \in \Re_+$. Jak łatwo zauważyć, otrzymana gęstość warunkowa to gęstość rozkładu normalnego o średniej 0 i wariancji $\frac{(n-1)\sigma^2}{ns^2}$, czyli:

$$T\,|_{S^2=s^2} \sim N\left(0, \sqrt{\frac{(n-1)\sigma^2}{ns^2}}\right),$$

co potwierdza wynik (5) podany w poprzedniej sekcji.

5. Symulacje komputerowe

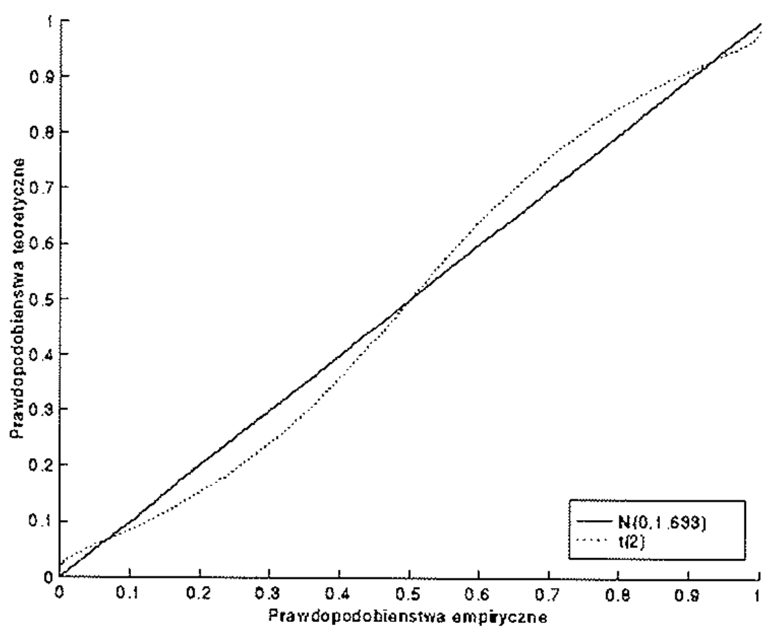
Rozważany w tym artykule problem pojawia się w książkach, które traktują rachunek prawdopodobieństwa jako teoretyczne narzędzie do praktycznego stosowania, np. w naukach ekonomicznych. Niestety, wydaje się, że tzw. praktycy nieufnie podchodzą do teoretycznych uzasadnień, szczególnie gdy te w jakiś sposób nie są zgodne z ich intuicją. Na szczęście obecnie nauka dysponuje bardzo szybkimi maszynami obliczeniowymi, które pozwalają na ,,praktyczną" weryfikację wielu teorii probabilistycznych za pomocą tzw. metod Monte Carlo, czyli przybliżania teoretycznych prawdopodobieństw częstościami z bardzo dużych prób tworzonych za pomocą generatorów liczb pseudolosowych. Dlatego teraz zostaną przedstawione wyniki eksperymentów komputerowych dotyczących rozkładu statystyki T Studenta, przeprowadzonych z wykorzystaniem aplikacji MATLAB, w której generowane będą liczby pseudolosowe, wykonywane obliczenia i tworzone rysunki, oraz aplikacji Statistica, która posłuży do wykonania testów normalności.



Rys. 1. PP-plot $K = 10^8$ realizacji zmiennej losowej T dla $n = 2$

Źródło: opracowanie własne.

Każdorazowo, dla n = 2, 3, 4 oraz $K = 10^8$, generowana była $(K \times n)$-elementowa macierz realizacji niezależnych zmiennych losowych o rozkładzie normalnym ze średnią 1 i dyspersją 2, czyli $[X_{ij}]_{K \times n}\, iid \sim N(1, 2)$. Na podstawie poszczególnych wierszy tej macierzy obliczone zostały realizacje statystyki T Studenta (3). Pierwszy eksperyment ma charakter kontrolny. Chodzi o krótką weryfikację działania generatora liczb pseudolosowych programu MATLAB oraz testów normalności programu Statistica. Sprawdzony zostanie niebudzący wątpliwości fakt, że bezwarunkowy rozkład statystyki T Studenta jest rozkładem t-Studenta odpowiednio o $(n - 1) = 1, 2, 3$ stopniach swobody. Dla n = 2 wykres prawdopodobieństwo-prawdopodobieństwo (pierwsza współrzędna to prawdopodobieństwo empiryczne, druga – odpowiednie prawdopodobieństwo teoretyczne), czyli tzw. PP-plot, można zobaczyć na rys. 1. Rozkład empiryczny został na nim porównany, obok rozkładu $\sim t(1)$ (linia kropkowana), również z rozkładem $\sim N(0{,}3379)$ (linia ciągła) oraz $\sim N\left(0, \frac{3}{2}\right)$ (linia przerywana). Drugie parametry w testowych rozkładach normalnych zostały wybrane na podstawie danych. W pierwszym przypadku σ została wyestymowana metodą największej wiarygodności, a bardzo wysoką wartość tego estymatora łatwo wytłumaczyć faktem, że dla 1 stopnia swobody zmienna losowa o rozkładzie t-Studenta nie ma określonej dyspersji (odpowiednia całka nie jest zbieżna). W drugim przypadku dyspersja została dobrana tak, by wizualnie gęstość odpowiedniego rozkładu normalnego możliwie najlepiej pokrywała się z histogramem gęstości empirycznej. Nie zmieniło to jednak oczywistego faktu, że do danych najlepiej dopasowany był w tym przypadku rozkład $\sim t(1)$, którego PP-plot ułożył się niemal idealnie na przekątnej (czyli prawdopodobieństwa empiryczne niemal idealnie pokrywały się z teoretycznymi). Ponadto wszystkie trzy testy normalności obliczone w programie Statistica (Kołmogorowa-Smirnowa, Lilienforsa i χ^2) dały we wszystkich trzech przypadkach (n =2,3,4) empiryczne poziomy istotności (p-value) dużo poniżej $\frac{1}{100}$, czyli sugerowały odrzucenie hipotezy o rozkładzie normalnym, a więc wszystko zgodnie z teorią.



Rys. 2. PP-plot K_s = 25947 realizacji zmiennej losowej $T|_{S=1}$ dla n = 3

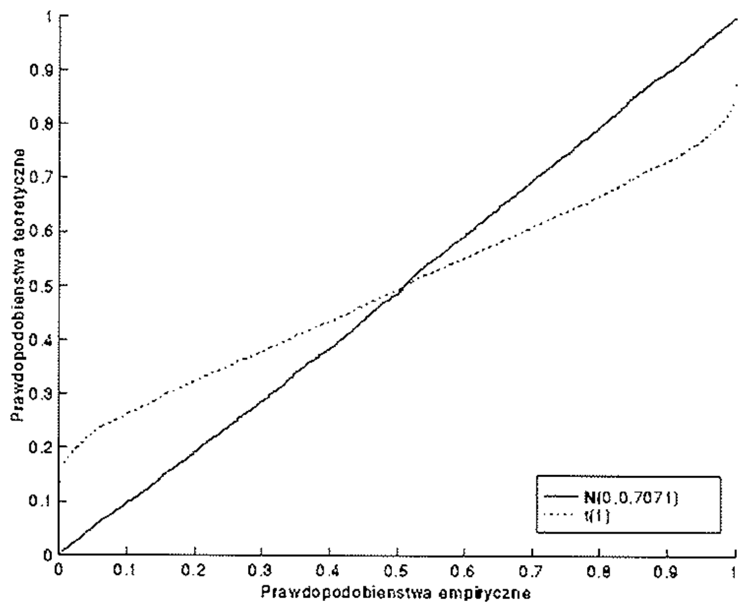
Źródło: opracowanie własne.

Teraz kolej na właściwe testy. Dla wygenerowanych już realizacji statystyk T Studenta, tym razem pod warunkiem zadanej wartości statystyki S, zostaną porównane rozkłady empiryczne z dwoma rozkładami teoretycznymi: niewłaściwie sugerowanym we wspomnianych podręcznikach rozkładem t-Studenta i teoretycznie uzasadnionym w niniejszym artykule rozkładem normalnym (5). Aby wyznaczyć ten empiryczny rozkład warunkowy, należało ustalić wartość statystyki S. Problem polega jednak na tym, że ma ona rozkład ciągły, a więc odpowiednie zdarzenie B, zdefiniowane równością (4), będzie prawie niemożliwe. Aby go rozwiązać, należy postąpić podobnie jak przy uogólnianiu wzoru na prawdopodobieństwo warunkowe dla prawie niemożliwych warunków, czyli potraktować to zdarzenie jako granicę $\lim_{d \to 0^+} \{S \in [s-d, s+d]\}$. Wtedy dla wystarczająco małego d otrzyma się pewien podzbiór realizacji statystyki (3) wywołany zdarzeniami elementarnymi w przybliżeniu zawartymi w B. Ostatecznie w eksperymentach zostało przyjęte $d = 5 \cdot 10^{-4}$ oraz $s = 1$. To ostatnie ze względu na bliskość modalnej rozkładu statystyki S, co zwiększa szanse na wygenerowanie realizacji S bliskich 1. Wyniki symulacji zostały zawarte w poniższej tabeli.

n	K_s	$D(T \mid S=1)$	K-S	Lil.	χ^2
2	21973	$\sqrt{2}$	n.i.	n.i.	0,44
3	25947	$\frac{2}{3}\sqrt{6}$	n.i.	n.i.	0,35
4	24065	$\sqrt{3}$	n.i.	n.i.	0,50

Pierwsza kolumna zawiera liczbę n zmiennych losowych iid $\sim N(1,2)$, składających się na statystykę T Studenta, druga – liczbę K_S realizacji statystyki T, dla których $S \in [1-d, 1+d]$ (wśród wszystkich $K = 10^8$ realizacji), czyli tych, dla których był wyznaczany empiryczny rozkład $T|_{S=1}$. Trzecia kolumna zawiera warunkową dyspersję $T|_{S=1}$ obliczoną na podstawie (5), a ostatnie trzy kolumny – wartości empirycznych poziomów istotności testów normalności: Kołmogorowa-Smirnowa, Lilienforsa i χ^2. Ponadto PP-plot dla odpowiednich rozkładów normalnego i t-Studenta został przedstawiony na rys. 2. Jak widać, rozkład empiryczny $T|_{S=1}$

pokrywa się niemal idealnie z rozkładem normalnym, co zresztą potwierdzają wysokie wartości p-value we wszystkich testach normalności[2]. Na koniec przeprowadzony został jeszcze analogiczny eksperyment przy warunku $S = 2$, a powstały w jego wyniku PP-plot przy $n = 2$, dla $K_S = 5189$ realizacji, można zobaczyć na rys. 3. Również i w tym przypadku p-value były bardzo wysokie, np. w teście χ^2 dla danych, których dotyczy wspomniany PP-plot, $p = 0{,}71$.



Rys. 3. PP-plot $K_s = 5189$ realizacji zmiennej losowej $T\,|_{S=2}$ dla $n = 2$

Źródło: opracowanie własne.

6. Podsumowanie

Autor ma nadzieję, że przedstawione powyżej analityczne i empiryczne uzasadnienia definitywnie przekonały wszystkich wątpiących, że warunkowy rozkład statystyki T Studenta przy danej wariancji z próby nie jest rozkładem t-Studenta, lecz rozkładem normalnym (5). Do rozwiązania

[2] Program Statistica podaje p-value „n.i.”, gdy jest ono tak duże, że nie pozwala na odrzucenie hipotezy zerowej przy żadnym sensownym poziomie istotności.

pozostał już tylko jeden problem. Skoro właściwy rozkład statystyki (3) zależy od nieznanej wartości σ, to jak sobie poradzić z obliczaniem prawdopodobieństw, np. z omawianym na początku artykułu zadaniem:

$$P(\overline{X} < u \mid S = s) = \Phi\left(\frac{u-\mu}{\sigma}\sqrt{n}\right)?$$

Jeśli taki problem pojawi się w praktyce, to najsensowniejszym rozwiązaniem wydaje się być wyestymowanie parametru σ na podstawie próby za pomocą jednego ze znanych estymatorów dyspersji lub, po spierwiastkowaniu, wariancji, co oczywiście nie zmieni faktu, że rozkład $T|_{S=s}$ pozostanie normalny. Często jednak ten problem jest wywoływany sztucznie przez autorów podręczników do statystyki dla ekonomistów, w celu przećwiczenia wiadomości dotyczących rozkładu t-Studenta, np. umiejętności korzystania z tablic tego rozkładu. W takich wypadkach najlepiej po prostu zrezygnować z tego typu ćwiczeń i zastąpić je zadaniami o innej treści, choćby takimi, jak np.: „Jakie jest prawdopodobieństwo, że zmienna losowa o rozkładzie $\sim \mathrm{t}(n)$ nie przekroczy wartości u?”. Może mniej wysublimowane, ale przynajmniej poprawne...

Literatura

Billingsley P. (2009). *Prawdopodobieństwo i miara*. Wydawnictwo Naukowe PWN. Warszawa.

Feller W. (2007). *Wstęp do rachunku prawdopodobieństwa*. Wydawnictwo Naukowe PWN. Warszawa.

Kassyk-Rokicka H. (red.) (2001). *Statystyka. Zbiór zadań*. PWE. Warszawa.

Ostasiewicz S., Rusnak Z., Siedlecka U. (2004). *Statystyka. Elementy teorii i zadania*. Wydawnictwo Akademii Ekonomicznej we Wrocławiu. Wrocław.

Pawłowski Z. (1980). *Statystyka matematyczna*. PWN. Warszawa.

Wawrzynek J. (2007). *Metody opisu i wnioskowania statystycznego*. Wydawnictwo Akademii Ekonomicznej we Wrocławiu. Wrocław.

Witkowski M. (red.) (2010). *Statystyka matematyczna w zarządzaniu*. Wydawnictwo Uniwersytetu Ekonomicznego w Poznaniu. Poznań.