

■ ■ ■ ■ ■ PRZEGLĄD ■ ■ ■ ■ ■ ZAGADNIEŃ NAUKOWYCH

Robert TWARDOSZ

Zakład Klimatologii UJ

Adam WALANUS

Wydział Geologii, Geofizyki i Ochrony Środowiska AGH — Kraków

PRAKTYCZNE TESTOWANIE ISTOTNOŚCI RÓŻNICY WARTOŚCI ŚREDNICH W SERIACH OPADOWYCH

A PRACTICAL METHOD FOR TESTING OF THE SIGNIFICANCE OF DIFFERENCE BETWEEN AVERAGE VALUES IN PRECIPITATION SERIES

Opady atmosferyczne są jednym z najbardziej zmiennych w czasie i przestrzeni elementów klimatu. Zakres wahań sum opadów, bez względu na rozdzielczość czasową i strefę klimatyczną, jest bardzo duży (Prokop, 2006). Na przykład, w Polsce opady dobowe dochodzą do 300 mm, większość sum jednak osiąga wartość do 5 mm (Olechnowicz-Bobrowska, 1970; Twardosz, 2007). W związku z tym ze statystycznego punktu widzenia opady cechują się silną asymetrią rozkładu prawdopodobieństwa opisującego ich wielkość. Dodatnia asymetria (skośność, współczynnik γ_1) rozkładu wynika z oczywistej cechy opadów, iż nie przyjmują one wartości ujemnych. Stopień tej asymetrii zwiększa się wraz ze skracaniem okresu sumowania opadów, wtedy, gdy prawdopodobieństwo braku opadu w przyjętej jednostce czasu jest wyraźnie większe od zera. W przypadku opadów dobowych obserwujemy silną asymetrię, natomiast już w miesięcznych prawdopodobieństwo braku opadu jest znikome, co więcej, suma opadu jest sumą wielu niezależnych zdarzeń opadowych, co – zgodnie z centralnym twierdzeniem granicznym rachunku prawdopodobieństwa – powoduje, że rozkład prawdopodobieństwa sum opadu będzie zbliżony do rozkładu normalnego. Tak więc w sumach rocznych asymetria

jest najczęściej niewielka i opady mają rozkład zbliżony do normalnego (Kozuchowski, 1985; Twardosz, 2007). Oznacza to m.in., że prawdopodobieństwo pojawienia się sumy rocznej mniejszej lub większej od średniej jest takie samo. Zmienność opadów dobowych jest niezwykle duża. Zdarza się, że w niektórych miesiącach maksymalne dobowe sumy opadów przekraczają nawet średnie sumy miesięczne. W badaniach klimatologicznych często bada się średnie sumy opadów w porach roku lub typach cyrkulacji (Ustrnul, Czekierda, 2000; Twardosz, 2005; Świątek, 2009), ale nie zawsze dokonuje się oceny istotności różnicy średnich sum. Stosowanie standardowego testu t Studenta wymaga spełnienia założenia o normalności rozkładów obu serii danych opadowych.

W pracy zaproponowane są dwa odmienne sposoby rozwiązania problemu testowania istotności różnic średnich opadów w przypadku istotnej asymetrii rozkładów. Pierwszy z nich polega na transformacji danych za pomocą stosowanego w statystyce przekształcenia Boxa-Coxa. Drugie podejście polega na zastosowaniu testów istotności niezależnych od rozkładu. Zaproponowano trzy takie testy: test serii Walda-Wolfowitza, test sum rang U Manna-Whitney'a i test zgodności Kołmogorowa-Smirnowa. Wszystkie testy są zilustrowane przykładami zrealizowanymi za pomocą standardowych formuł arkusza MS Excel. Formuły te są praktycznie niezależne od wersji programu, jedynie wygląd zamieszczonych „zrzutów z ekranu” może być w nowszych wersjach nieco inny, co nie ma oczywiście żadnego merytorycznego znaczenia.

Przykłady obliczeń pokazano na podstawie sekularnej serii obserwacji opadu w Krakowie oraz typów cyrkulacji w Polsce Południowej T. Niedźwiedzia (1981, 2010). Wykorzystano dobowe sumy opadów oraz informacje o postaci i typie opadów atmosferycznych z lat 1863-2008. Na ich podstawie wydzielono sumy opadów ciekłych i śnieżnych oraz mieszanych i burzowych w podziale na typy cyrkulacji, przy jakich te opady występowały.

Współczynnik skośności jako charakterystyka rozkładu sum opadów

Wysokość (suma) opadu jest wielkością z natury niesymetryczną, co wynika z faktu, że nie przyjmuje wartości ujemnych, a nie ma ograniczenia po stronie wartości dodatnich. Miarą asymetrii (skośności) rozkładu zmiennej losowej jest współczynnik skośności, zdefiniowany jako moment centralny rzędu trzeciego, unormowany przez moment centralny rzędu drugiego (wariancję). Współczynnik skośności, oznaczany w statystyce często przez γ_1 , jest wielkością niemianowaną, niezależną od skali zmiennej losowej. Przyjmuje on wartości dodatnie i ujemne. Rozkłady opadów cechują się dodatnim współczynnikiem skośności, czyli mają asymetrię prawostronną, a więc w serii pomiarowej jest wiele wartości małych

i niewiele bardzo dużych, które jednak „przeciągają na swoją stronę” średnią. Na przykład, seria zawierająca 20 jedynek i jedną dwójkę ma $\gamma_1 = 4,6$, natomiast przy pięciu dwójkach i jednej jedynce jest to już tylko $\gamma_1 = 1,6$. Oczywiście przy równej liczbie jedynek i dwójek $\gamma_1 = 0$. Rozkład normalny, jak każdy rozkład symetryczny, ma skośność zerową, skośność rozkładu wykładniczego natomiast wynosi 2.

Estymator współczynnika skośności jest łatwy do obliczenia w zasadzie nawet za pomocą kalkulatora. W MS Excel do jego obliczania służy funkcja SKOŚNOŚĆ. Mając obliczony współczynnik skośności dla danej serii pomiarowej, trzeba sobie postawić pytanie, czy jest on istotnie różny od zera. Odpowiedzi łatwo udzielić, znając następujący wzór na odchylenie standardowe estymatora współczynnika skośności: $\sigma = (6/n)^{1/2}$, gdzie n jest liczebnością serii. Wzoru tego nie ma w Excelu, wprowadzenie go jako formuły jest jednak bardzo łatwe. Na przykład, we wspomnianym już przykładzie, gdzie 20 razy „opad” wynosił 1, a 5 razy 2, $\gamma_1 = 1,6 \pm 0,49$. Estymator współczynnika skośności ma rozkład normalny przy dużych n , jednak również przy $n=25$ można posłużyć się w przybliżeniu rozkładem normalnym, który w przypadku odchyłki od zera (1,6) ponad trzykrotnie większej od niepewności (0,49) każe odrzucić hipotezę o braku skośności ($\gamma_1=0$), czyli o symetrii rozkładu.

Testowanie równości średnich w przypadku rozkładu normalnego

Jak już wspomniano we wstępie, sumy opadów z dłuższych okresów podlegają rozkładowi normalnemu, przynajmniej w zadowalającym przybliżeniu. W takiej sytuacji standardowym testem równości średnich jest test oparty na statystyce t Studenta. Warto przypomnieć, że test ten służy do porównania średnich w dwóch grupach wyników niekoniecznie będących szeregami czasowymi.

Na rysunku 1 widać, jak łatwo jest zastosować ten test w arkuszu MS Excel. Służy do tego funkcja TEST.T, której wynikiem jest p -wartość, która, jeżeli jest mniejsza od standardowej wartości poziomu istotności $\alpha = 0,05$, to oznacza to podstawę do odrzucenia hipotezy o równości wartości oczekiwanych w analizowanych seriach pomiarowych. Komentarza wymagają jedynie dwa parametry funkcji TEST.T, które w przykładzie na rys. 1 obydwa mają wartość 2 (jest to przypadek najczęstszy). Pierwszy z tych parametrów, nazywany błędnie w polskojęzycznym Excelu parametrem „ślad” (ang. *Tails!* – ogony), informuje program, czy jesteśmy zainteresowani testem jedno czy dwustronnym. Jest to ważne zagadnienie, zmieniające p -wartość dwukrotnie. Test jednostronny stosuje się wtedy, gdy wiadomo, że jedna z serii może mieć średnią wyłącznie większą (mniejszą) od drugiej serii. Taka niezależna wiedza skutkuje większą mocą testu. W przypadku pytania, czy średnie są w ogóle różne czy równe, wartością parametru powinno być 2. Drugi

parametr określa typ testu, przy czym dostępne są 3 typy: 1 – sparowany, 2 – z jednakowym odchyleniem standardowym w obu próbach, 3 – z różnymi odchyleniami (brak wiedzy, czy są równe). Pierwszy typ, częsty w badaniach medycznych, w klimatologii ma mniejsze zastosowanie. Trzeba tu mieć serię par wartości, co do których zachodzi przypuszczenie, że systematycznie pierwsza wartość jest mniejsza (większa). Zagadnienie równości odchyłeń standardowych (wariancji) jest natomiast bardzo aktualne w badaniach klimatycznych. Posiadanie wiedzy o równości wariancji (2) poprawia moc testu. Jeżeli jednak zachodzi przypuszczenie, że średnie w dwóch seriach opadowych są takie same, to można też domniemywać, że rozkłady opadów są takie same, a więc tym bardziej wariancje są takie same. Wydaje się więc, że przypadek ten będzie miał częste zastosowanie. Typ 3. testu stosuje się, gdy nic nie wiadomo o wariancjach obu serii, a w szczególności, jeżeli test F (dostępny w Excelu) kazał odrzucić hipotezę o równości wariancji.

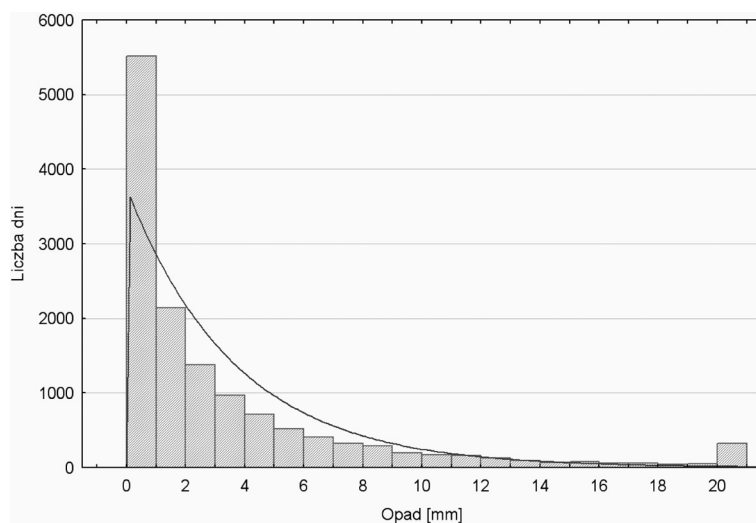
C1		=TEST.T(A1:A24;B1:B20;2;2)				
	A	B	C	D	E	F
1	25	44	0.201142			
2	38	42				
3	42	49				
4	17	39				
5	38	30				

Rys. 1. Fragment arkusza MS Excel z testem t Studenta równości średnich. Wynik to p -wartość, widoczna wartość 0,2 jest większa od 0,05, co oznacza nieistotność różnicy średnich. Widoczne w formule adresy komórek A24 i B20 są związane z długościami badanych serii

Fig. 1. Part of MS Excel spreadsheet with the Student t test applied to the series contained in the columns A and B. The p -value is given as final result, since the obtained value 0.2 is greater than the standard significance level 0.05, the series can't be treated as different

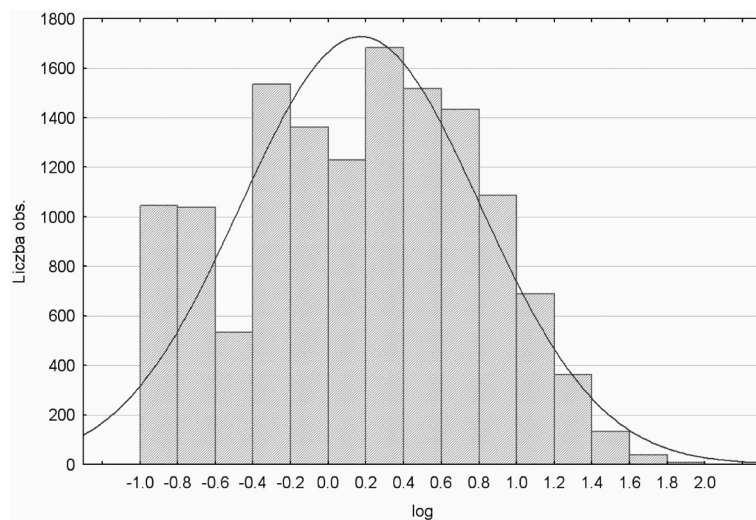
Rozkład wykładniczy

Dobowe opady charakteryzują się wyraźną dodatnią skośnością. Można jednak powiedzieć więcej, podać kształt rozkładu, jakiemu podlegają. Otóż z dobrym przybliżeniem jest to rozkład wykładniczy (rys. 2). W takiej sytuacji można się spodziewać, że zlogarytmowanie liczbowych wartości opadu (milimetrów logarytmować nie można) da zmienną losową o rozkładzie normalnym (rys. 3).



Rys. 2. Histogram dobowych sum opadów ciekłych w Krakowie (1863-2008). Do histogramu dopasowano rozkład wykładniczy. Słupkę >20 mm zawiera wszystkie wyższe opady

Fig. 2. Histogram of the daily liquid precipitation in Krakow (1863-2008). The exponential distribution is fitted to the histogram. The last rectangle contain all values >20 mm



Rys. 3. Histogram opadów (jak na rys. 2) po zlogarytmowaniu ($\log_{10}$, wartość 0 to 1 mm opadu). Do histogramu dopasowano rozkład normalny. Zwraca uwagę dobra zgodność z modelem prawej strony rozkładu

Fig. 3. Histogram of the data as in Fig. 2, however, after log transformation ($\log_{10}$, so 0 means 1 mm). The Gaussian curve is fitted, with good result, especially if the right side is considered

Zgodność danych prezentowanych na rys. 2 z modelem wykładniczym zaburza zbyt duża liczebność najniższych opadów (słupek histogramu 0-1 mm). Widoczne jest to też na rys. 3, gdzie dwa pierwsze słupki są wyraźnie za wysokie względem dopasowanej krzywej Gaussa. Warto zauważyć, że proponowany często do opisu opadu model rozkładu gamma (Wilks, 2006) nie będzie lepszy w tej sytuacji, gdyż właśnie rozkład wykładniczy, będący jego skrajnym przypadkiem, daje największe prawdopodobieństwo zerowego (bardzo niskiego) opadu.

Przekształcenie Boxa-Coxa

Tytułowe przekształcenie, ogólniejsze od zwykłego logarytmowania zmiennej, pozwala w wielu przypadkach efektywnie przetransformować zmienną o skośnym rozkładzie w zmienną normalną. To potęgowe przekształcenie ma względnie prostą postać:

$$y = (x^\lambda - 1) / \lambda$$

gdzie λ jest dobieranym parametrem przekształcenia, w zakresie mniej więcej od -0,5 do 0,5.

Wartość $\lambda = 0$ jest tu oczywiście niedopuszczalna, jednak dla zachowania ciągłości przekształcenie definiuje się je wtedy jako logarytmiczne:

$$y = \log(x).$$

W każdym przypadku wartości transformowane muszą być dodatnie; $x > 0$.

Na rysunku 4 widoczna jest formuła przekształcenia Boxa-Coxa wpisana do arkusza Excela. Parametr λ , z którego korzysta się w formule, jest wpisany do komórki B1. W formule przeliczającej oryginalne wartości zawarte w kolumnie A należy zwrócić uwagę na znak \$ zapewniający poprawne odwoływanie się do parametru λ . W przypadku przekształcenia logarytmicznego formuła jest prostsza i nie wymaga parametru: =LN(A2). Dla kwestii skośności nie ma znaczenia, jakiego logarytmu się używa.

Wartość λ można dobrać eksperymentalnie. Nieco bardziej złożona formuła:

$$=JEŻELI(B\$1=0;LOG(A2);(POTĘGA(\$A2;B\$1)-1)/B\$1)$$

„obsługuje” też przypadek $\lambda = 0$. W komórce wpisywać należy różne wartości z przedziału od -0,5 do 0,5 i obserwować współczynnik skośności danych przekształconych (na rys. 4 formuła =SKOŚNOŚĆ(B2:B15) jest umieszczona w wierszu 17). Zwykle dość szybko osiąga się zadowalająco niską wartość γ_1 . Na rysunku 4 $\gamma_1 = -0,006$ osiągnięto przy $\lambda = 0,15$.

Wracając do zagadnienia porównywania średnich, trzeba podkreślić, że do obu serii danych trzeba zastosować to samo przekształcenie Boxa-Coxa, przy tym

samym λ . Tak więc trzeba znaleźć takie λ , które daje zadowalająco niskie γ_1 dla obu serii. Oczywiście nie należy w celu poszukiwania γ_1 łączyć serii. Byłoby to zaprzeczeniem idei badania zgodności średnich, gdyż wymagałoby założenia, że są one równe.

λ = (POTĘGA(A2:B\$1)-1)/B\$1			
	A	B	
1		0.15	
2	7.7	2.38	
3	11.9	2.999	
4	11.3	2.923	
5	45.3	5.145	
6	7.4	2.336	
7	8.6	2.547	
8	19.1	3.713	
9	14.4	3.282	
10	8.9	2.581	
11	7.9	2.419	
12	1.2	0.185	
13	8.4	2.5	
14	2.1	0.785	
15	6.4	2.134	
16			
17	2.675	-0.006	

Rys. 4. Przekształcenie Boxa-Coxa w arkuszu MS Excel. W kolumnie A znajdują się przykładowe dobowe dane opadowe, w kolumnie B przekształcone. Wartość λ wpisuje się do komórki B1, a w komórce B17 obserwuje się uzyskany współczynnik skośności (formuła: =SKOŚNOŚĆ(B2:B15))

Fig. 4. The Box-Cox transform in MS Excel spreadsheet. The raw data fill the column A, while in B there are transformed ones. The trial values of parameter λ are in cell B1, and the resulting skewness coefficient values can be observed in B17

W podobny sposób, jak to jest pokazane na rys. 4, w arkuszu trzeba umieścić drugą serię opadową i przekształcać ją za pomocą takiego samego przekształcenia, odwołującego się „po lambdę” do tej samej komórki B1 (którą można adresować jako \$B\$1, wtedy będzie można całą formułę swobodnie kopiować).

Po znalezieniu właściwej wartości λ przekształcenia Boxa-Coxa obie przekształcone serie poddaje się standardowemu testowi t . Test Studenta oczywiście można wpisać do arkusza przed przystąpieniem do dobierania λ , trzeba jednak pamiętać, że logiczna kolejność jest następująca: najpierw przekształcić zmienne tak, aby miały rozkład zbliżony do normalnego (γ_1 bliskie zero), a potem testować testem wymagającym normalności.

Na rysunku 5 widoczny jest przykład realizacji zadania porównania średnich opadowych przy 21 typach cyrkulacji atmosferycznej. Dane oryginalne umieszczone

są w kolumnach od A do U, w wierszach od 11. W kolumnach od X do AQ, w wierszach od 11 znajdują się dane przekształcone według formuły (zapis dla komórki W11):

$$=JEZELI(\$V\$2=0;LOG(A11);(POTĘGA(A11;\$V\$2)-1)/\$V\$2)$$

gdzie komórka V2 zawiera dobieraną wartość współczynnika transformacji Boxa-Coxa (wartość 0 jest dopuszczalna). W wierszu 2 znajdują się średnie obliczone w serii przekształconych danych, niżej odchylenia standardowe, dalej liczebności serii i odchylenia standardowe średnich (formuła w komórce W5: $=W3/PIERWIASTEK(W4)$). Wiersz 7 zawiera współczynniki skośności, wiersz 8 zawiera wartości z wiersza 7 pomnożone przez liczebności serii (wiersz 4), czyli ważne współczynniki skośności. Wiersz 9 zawiera niepewności współczynników skośności, formuła w komórce W9:

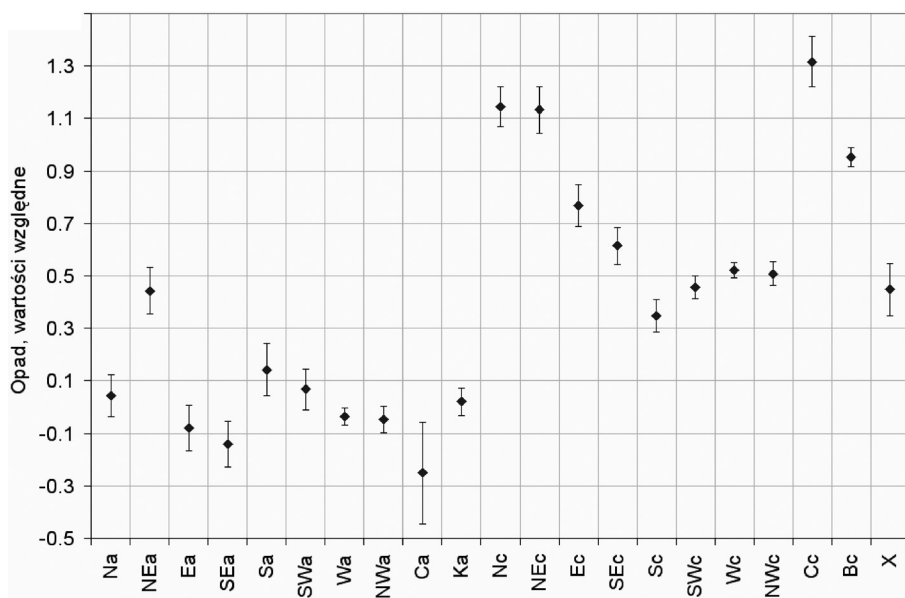
$$=PIERWIASTEK(6/W4)$$

Optymalizacja przekształcenia Boxa-Coxa polega na szukaniu wartości λ minimalizującej średnią ważonych współczynników skośności (V8). Trzeba podkreślić, że wartość 0,887, widoczna na rys. 5, wskazuje na bardzo małą średnią skośność, gdyż ważne wartości skośności, jak widać z rysunku, osiągają wartości rzędu 100. Na rysunku 6 widoczne są średnie opady po przekształceniu, które pozwala interpretować słupki błędów zgodnie z rozkładem normalnym. Warto przypomnieć, że w przypadku stykania się zakresów błędów 1σ (przypadek Ec i Sec na rys. 6) p -wartość testu równości średnich, przy podobnych wartościach obu σ , wynosi 0,16.

	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ
1	Cc	Bc	X		Na	NEa	Ea	SEa	Sa	SWa	Wa	NWa	Ca	Ka	Nc	NEc	Ec	SEc	Sc	SWc	Wc	NWc	Cc	Bc	X
2				0.07	0.04	0.44	-0.08	-0.1	0.14	0.07	-0.04	-0	-0.3	0.02	1.14	1.13	0.77	0.61	0.35	0.46	0.521	0.51	1.32	0.953	0.45
3					1.42	1.53	1.47	1.35	1.42	1.33	1.257	1.37	1.39	1.41	1.73	1.78	1.55	1.5	1.41	1.37	1.378	1.43	1.42	1.549	1.53
4					328	293	289	241	208	288	1390	700	51	698	515	404	379	432	537	953	2356	1027	214	1821	240
5					0.08	0.09	0.086	0.09	0.1	0.08	0.034	0.05	0.19	0.05	0.08	0.09	0.08	0.07	0.06	0.04	0.028	0.04	0.1	0.036	0.1
6																									
7					0.26	0.29	0.42	0.62	0.31	0.29	0.116	0.21	0.46	0.24	-0.1	-0.2	-0.1	-0.1	-0.1	-0.1	-0.17	0.07	-0.3	-0.2	0.05
8				0.887	86.1	83.8	121.3	151	65.4	83.3	161.9	150	23.3	169	-41	-100	-38	-33	-36	-71	-411	69.8	-72	-356	12.7
9					0.14	0.14	0.144	0.16	0.17	0.14	0.066	0.09	0.34	0.09	0.11	0.12	0.13	0.12	0.11	0.08	0.05	0.08	0.17	0.057	0.16
10																									
11	0.1	0.1	0.1		-2.13	-2.1	-2.13	-2.1	-2.1	-2.1	-2.13	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.13	-2.1	-2.1	-2.13	-2.1
12	0.1	0.1	0.1		-2.13	-2.1	-2.13	-2.1	-2.1	-2.1	-2.13	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.1	-2.13	-2.1	-2.1	-2.13	-2.1

Rys. 5. Fragment arkusza MS Excel zawierającego 21 serii opadowych o różnych długościach, od 51 do 2356, przekształconych w celu minimalizacji średniej ważonej skośności. Opis układu arkusza i użytych formuł znajduje się w tekście

Fig. 5. Part of MS Excel spreadsheet containing 21 precipitation series of length ranging from 51 to 2356. Data are transformed to obtain the minimal value of the weighted average skewness coefficient



Rys. 6. Wykres średnich wartości opadu (w umownej skali) w zależności od typu cyrkulacji atmosferycznej. Dane po przekształceniu Boxa-Coxa mają rozkłady bliskie normalnym, co pozwala interpretować zakresy odchyłeń standardowych średnich w naturalny sposób, jako 68% przedziały ufności.

Wykres ilustruje wartości z wierszy 2 i 5 arkusza widocznego na rys. 5

Fig. 6. Plot of average precipitation (relative scale) versus circulation type. Data are Box-Cox transformed and are almost normally distributed. As a result the typical interpretation can be attached to the 1σ error bars, connected with the normal distribution (68% confidence level). Plotted values are taken from the raw 2 and 5 of spreadsheet in Fig. 5

Test serii Walda-Wolfowitza

Przed przystąpieniem do omówienia tytułowego testu zgodności konieczna jest uwaga na temat pewnej typowej właściwości danych opadowych. Otóż ze względu na skończoną dokładność pomiaru częste są wartości 0,1 mm, 0,2 mm itd., co zilustrowano w tab. 1.

Nie wchodząc w zagadnienie rozkładu tych wartości, które – jak widać z tabeli – mogłyby być złożone, trzeba zwrócić uwagę, że wskazane zjawisko grupowania się wielu wyników pomiarów konkretnych wartościach jest krytyczne dla **testu serii** (ang. *Wald-Wolfowitz runs test*), który jest uniwersalnym rozwiązaniem zagadnienia porównywania dwóch zestawów liczb pod względem wartości. Problem ten ma charakter numeryczny, obliczeniowy. Otóż zakładając, że znacząca część wyników w dwóch zestawach ma identyczne wartości, widzimy, że test serii jest narażony na przypadkowe zgrupowanie wszystkich identycznych wartości w zaledwie

dwie serie (odpowiadające dwóm zestawom danych), co znacząco wpłynęłoby na wynik testu. Problem ten można łatwo rozwiązać, dodając do wartości pomiarowych małą wartość losową, rzędu np. 1% wartości oryginalnej. Ze względów numerycznych mogłoby to nawet być zaledwie 0,001%, a nawet mniej. W praktyce jednak pojawia się tu następne niebezpieczeństwo. Otóż zdarza się, że w wyniku przeliczeń w seriach danych występują licznie grupy „nierównych wartości”, np. 0,12. Nie można wykluczyć, że z jakichś względów jeden zestaw danych zawiera dużo wartości 0,1, a drugi „zaczyna się” od 0,12, podczas gdy oczywiście nie ma to takiego sensu, że najniższy opad był w drugim zestawie o 20% wyższy od najniższego opadu w zestawie pierwszym.

Tabela 1. Liczebność najniższych wartości sum dobowych opadów. Dane dotyczą opadu ciekłego, całkowita liczba dni z opadem wynosi 13695

Table 1. Number of repeats of a few lowest values of precipitation in the series of 13695 daily liquid precipitation

Opad [mm]	N
0,1	1049
0,2	1035
0,3	532
0,4	724
0,5	396
0,6	566
0,7	274
0,8	392
0,9	262

Test można wykonać w arkuszu MS Excel, co jest zilustrowane na rys. 7. Dwie serie opadowe umieszczone są w kolumnach A i B. W kolumnie C i E wykonano dodawanie małej liczby losowej, za pomocą formuły:

$$=ZAOKR(A1;1)*(0.995+0.01*LOS())$$

Funkcja zaokrąglająca do jednego miejsca po przecinku ma zabezpieczyć analizę przed przypadkiem wystąpienia drobnych różnic będących skutkiem przeliczeń. Funkcja LOS generuje liczbę losową o rozkładzie równomiernym z przedziału od 0 do 1. Przekształcenie liniowe wartości LOS zapewnia symetryczne dodanie lub odjęcie liczby nieprzekraczającej 1% oryginalnej wartości. Stosowanie funkcji LOS ma pewną specyfikę, o której trzeba tu pamiętać. Otóż jak każda inna funkcja przelicza się ona automatycznie przy każdej operacji w arkuszu, jednak w przypadku tej funkcji za każdym razem otrzymuje się inny wynik. Dlatego należy wykonać operację skopiowania szeregów i operację „wklej specjalnie” jako „war-

tości”. W kolumnach D i F są wpisane identyfikatory serii. Pary kolumn C i D oraz E i F zostały skopiowane do kolumn G i H, jedna pod drugą, po czym para kolumn G i H została posortowana. W komórce I2 znajduje się formuła:

$$=JE\acute{Z}ELI(H1<>H2;1;0)$$

przeciągnięta do końca podwójnej serii w kolumnie H. Suma liczb z kolumny I daje liczbę serii. Suma ta znajduje się w komórce K1. W komórkach K2 i K3 podane są liczebności serii, konieczne do wykonania testu statystycznego. Oczekiwana wartość liczby serii jest obliczona w komórce K4, odpowiednia formuła ma postać:

$$=2*K2*K3/(K2+K3)+1$$

W komórce K5 obliczona jest dyspersja (σ) liczby serii:

$$=PIERWIASTEK((K4-1)*(K4-2)/(K2+K3-1))$$

pozwalająca obliczyć liczbę z o rozkładzie unormowanym normalnym

$$=(K1-K4)/K5$$

która sama wiele mówi. Łatwo jednak obliczyć odpowiednie (jednostronne) prawdopodobieństwo:

$$=ROZKŁAD.NORMALNY.S(K6)$$

by zakończyć procedurę prostą odpowiedzią na pytanie o zgodność średnich w seriach A i B:

$$=K7>0,05$$

Trzeba pamiętać, że test serii w zasadzie jest testem zgodności rozkładów, a nie testem równości średnich. Można sobie np. wyobrazić nieco abstrakcyjną sytuację, gdy jeden zestaw danych składa się wyłącznie z wartości 3, a drugi z wartości 2 i 4 w równej liczbie. Średnie są wtedy takie same (równe 3), a serie będą tylko trzy (BBBBBAAAAAABBBBB), co nawet przy małej liczebności danych będzie skutkowało skrajnie niską wartością p , wskazującą na konieczność odrzucenia hipotezy o zgodności rozkładów, podczas gdy średnie są równe. Jednak w przypadku opadów rozkłady są zwykle podobne (co do postaci) i różnią się tylko głównym parametrem, czyli średnią.

KB		f _k = k/7 > 0.05									
	A	B	C	D	E	F	G	H	I	J	K
1	20.5	65.1	20.45113	A	65.15744	B	80.96731	A		k=	3983
2	20.0	59.2	19.99996	A	59.32073	B	65.15744	B	1	nA=	3536
3	18.1	53.5	18.04987	A	53.42975	B	62.89069	A	1	nB=	4393
4	16.5	46.1	16.56876	A	46.13695	B	59.32073	B	1	mi=	3919.2
5	15.2	44.2	15.26109	A	44.26572	B	55.47694	A	1	sigma=	44.0
6	13.8	39.9	13.82785	A	40.08144	B	53.42975	B	1	z=	1.45
7	13.7	38.8	13.76306	A	38.69729	B	49.47532	A	1	p=	0.926518
8	13.4	36.8	13.33415	A	36.80492	B	46.13695	B	1	śrA=śrB=	PRAWDA
9	13.2	35.3	13.21529	A	35.14397	B	45.20149	A	1		
10	13.1	35.2	13.09843	A	35.04243	B	44.26572	B	1		
11	12.6	33.4	12.57474	A	33.54102	B	44.00619	B	0		
12	12.5	32.8	12.51223	A	32.78176	B	40.54981	A	1		
4390		0.1			0.10024	B	1.20314	B	0		
4391		0.1			0.10046	B	1.20303	B	0		
4392		0.1			0.09953	B	1.20300	B	0		
4393		0.1			0.10000	B	1.20298	B	0		
4394							1.20294	A	1		

Rys. 7. Test serii wykonany na przykładzie opadu ciekłego w okresie wiosny (A) i jesieni (B).
Opis arkusza znajduje się w tekście

Fig. 7. The series test as comparing the spring (A) and the autumn (B) precipitation

Test sumy rang U Manna-Whitney'a

Test porównawczy dwóch serii pomiarowych oparty na statystyce *U* Manna-Whitney'a jest podobny do omówionego wcześniej testu serii. Jednak w tym przypadku statystyką testową jest nie liczba serii, która przy braku różnic w opadach nie może być zbyt mała, lecz suma tzw. rang, która nie powinna być zbyt mała ani zbyt duża.

Sposób obliczania statystyki testowej, a więc i sens rang ilustruje arkusz Excela przedstawiony na rys. 8. W kolumnach A i C umieszczone są dwie serie opadowe A i B. W kolumnach E i F serie te zostały wpisane jedna pod drugą, z zachowaniem znacznika serii (A i B). Kolumny G i H ilustrują wynik sortowania kolumn E i F. Kolumna I zawiera kolejne liczby naturalne, a więc rangi serii. Kolumna J zawiera rangi wyłącznie dla jednej z serii (obojętne której). Formuła obliczająca tę kolumnę widoczna jest na rysunku. Liczba *R* (komórka J27) to suma rang ($=\text{SUMA}(J1:J25)$). Liczba *U* otrzymana według formuły: $=J27-0,5*15*(15+1)$ jest tytułowym wynikiem testu. Wartość 15 w tej formule to liczebność serii, której rangi zsumowano (liczba *R*). Liczbę *U* można przekształcić liniowo na liczbę *Z*, która podlega rozkładowi normalnemu unormowanemu, formuła zawierająca wartość oczekiwaną *U* i jej dyspersję jest następująca:

$$=(K27-0,5*15*10)/\text{PIERWIASTEK}(15*10*(15+10+1)/12)$$

gdzie liczby 15 i 10 to liczebność serii, oczywiście różne w różnych przypadkach, a liczba 12 jest stałą (podobnie jak 0,5).

Test sumy rang Manna-Whitney'a porównujący opad ciekły w okresie wiosny i jesieni dał wynik $Z = 0,4$, co daje $p = 0,33$ w żadnym wypadku niewskazujące na różnicę w seriach.

J1	fx =JEŻELI(H1="A";1;0)											
	A	B	C	D	E	F	G	H	I	J	K	L
1	17.5 A	5.4 B	17.5 A	0.6 B	1	0						
2	2.9 A	11.2 B	2.9 A	1.8 B	2	0						
3	5.3 A	0.6 B	5.4 A	2.9 A	3	3						
4	106.2 A	15.5 B	106.2 A	4.1 A	4	4						
5	4.9 A	1.8 B	4.9 A	4.2 B	5	0						
6	5.0 A	36.1 B	5.0 A	4.9 A	6	6						
7	4.1 A	17.7 B	4.2 A	5.0 A	7	7						
8	9.2 A	4.2 B	9.2 A	5.3 A	8	8						
9	21.3 A	25.1 B	21.3 A	5.4 B	9	0						
10	14.4 A	18.9 B	14.4 A	9.2 A	10	10						
11	15.8 A		15.8 A	9.7 A	11	11						
12	14.3 A		14.3 A	11.2 B	12	0						
13	20.6 A		20.6 A	12.4 A	13	13						
14	12.4 A		12.4 A	14.3 A	14	14						
15	9.7 A		9.7 A	14.4 A	15	15						
16			5.4 B	15.5 B	16	0						
17			11.2 B	15.8 A	17	17						
18			0.6 B	17.5 A	18	18						
19			15.5 B	17.7 B	19	0						
20			1.8 B	18.9 B	20	0						
21			36.1 B	20.6 A	21	21						
22			17.7 B	21.3 A	22	22						
23			4.2 B	25.1 B	23	0						
24			25.1 B	36.1 B	24	0						
25			18.9 B	106.2 A	25	25						
26									R	U	Z	
27									194	74	-0.055	

Rys. 8. Arkusz MS Excel zawierający test Manna-Whitney'a porównujący dwie przykładowe serie opadowe. Dla jasności arkusza serie są krótkie, o liczebności odpowiednio 15 i 10. Wynikiem testu jest tu liczba Z o rozkładzie normalnym unormowanym, dla której łatwo otrzymać p -wartość. Formuły zawarte w arkuszu podane są w tekście

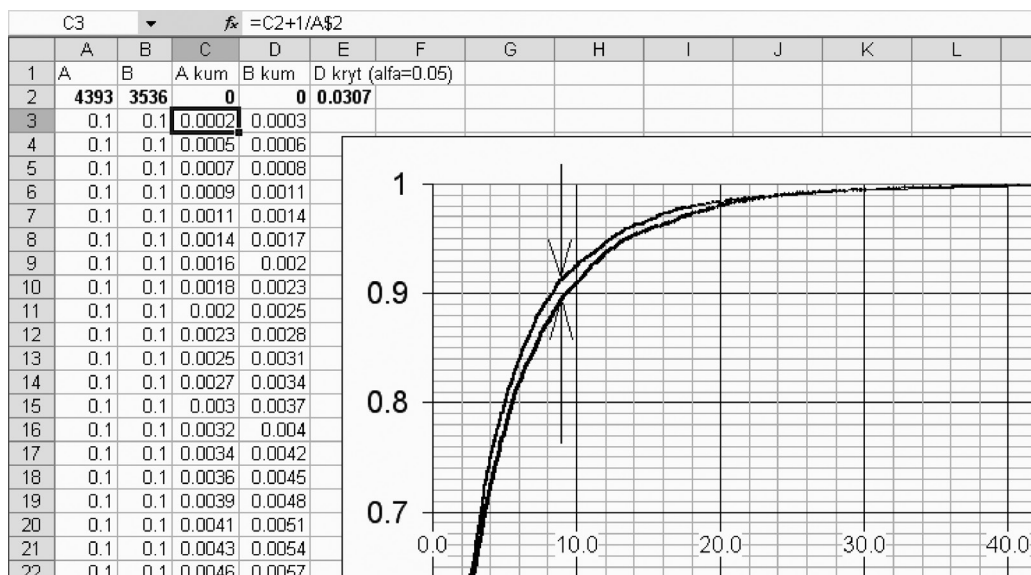
Fig. 8. MS Excel spreadsheet with the Mann-Whitney test as comparing two short exemplary data series. The standard normal Z value is given as the final calculation result

Test Kołmogorowa-Smirnowa

Średnia (właściwie wartość oczekiwana) to parametr rozkładu zmiennej losowej. Można powiedzieć, że jest to najważniejszy parametr. Tak więc do

porównywania średnich można oczywiście użyć testu zgodności dwóch rozkładów prawdopodobieństwa, przy czym chodzi tu o rozkłady empiryczne, otrzymane z danych. Test Kołmogorowa-Smirnowa jest bardzo intuicyjnym testem zgodności. Polega on na sprawdzeniu, czy maksymalna rozbieżność krzywych kumulacyjnych nie jest za duża.

Poniższy opis procedury zaimplementowanej w arkuszu MS przypomni pojęcie empirycznej krzywej kumulacyjnej (dystrybuanty empirycznej).



Rys. 9. Test Kołmogorowa-Smirnowa wykonany w arkuszu MS Excel za pomocą elementarnych formuł. Porównane są serie opadu ciekłego, wiosennej z jesienną. Brak różnic

Fig. 9. The Kolmogorov-Smirnov test in MS Excel spreadsheet. It is apparent that there is no significant difference between the winter and autumn precipitation series

Wkopiujmy dwie serie danych A i B do kolumny A i B Excela (rys. 9), zaczynając od trzeciego wiersza. Serie należy posortować rosnąco (każdą kolumnę osobno). W drugim wierszu wpisujemy liczebności tych serii (na rys. 9 są to wartości 4393 i 3536). Opis zamieszczony w pierwszym wierszu oczywiście nie jest konieczny, jest jednak zalecany. W komórkach C2 i D2 wpisujemy zera, które będą początkowymi wartościami krzywych kumulacyjnych. Następnie w komórce C3 należy wpisać formułę (widoczną na rysunku):

$$=C2+1/A\$2$$

w której znak dolara oznacza stałe adresowanie komórki zawierającej liczebność serii. Formuła ta, po skopiowaniu w dół tak daleko, ile jest danych w serii A,

obliczyć krzywą kumulacyjną. Ostatnią wartością powinno być 1. Analogicznie postępujemy w przypadku serii B, w komórce D3 wpisujemy:

$$=D2+1/B\$2$$

(można skopiować z C3, lecz ze skorygowaniem adresu liczebności serii).

Krzywe kumulacyjne należy wykreślić, używając wykresu XY (punktowego). Wykonuje się to zaznaczając dane z kolumn A i C (od wiersza 3) i wybierając odpowiedni wykres. Nieco trudniejsze jest dodanie do otrzymanego wykresu serii B. Klikając prawym klawiszem myszki na marginesie wykresu, należy wybrać opcję „Dane źródłowe”, następnie w otrzymanym oknie, na karcie „Serie” należy dodać nową serię i wybrać dla niej dane z kolumn B i D.

Otrzymany wykres dystrybuant empirycznych może mieć dużą wartość poznawczą. By zakończyć test, trzeba obliczyć wartość krytyczną statystyki $D_{n,k}$ Kołmogorowa-Smirnowa, co można wykonać w którejkolwiek komórce; na rys. 9 jest to komórka E2. Formuła jest następująca:

$$=1,36/PIERWIASTEK(A2*B2/(A2+B2))$$

Liczba 1,36 odpowiada poziomowi istotności $\alpha = 0,05$. Wzór ten zasadniczo ważny jest dla liczebności danych nie mniejszych od 35, jeśli jednak wynik testu nie będzie „na granicy”, to nie trzeba tego wymogu traktować krytycznie.

Odczyt wyniku testu polega na sprawdzeniu na wykresie, czy jest takie miejsce (wartość opadu), gdzie rozbieżność krzywych kumulacyjnych byłaby większa od wartości obliczonej w komórce G2 (na rys. jest to 0,0307).

Na rysunku 9 największa odległość dystrybuant występuje w przypadku opadu ok. 9 mm. Widać wyraźnie, że w żadnym wypadku nie jest ona jednak większa od 0,0307, gdyż wynosi ok. 0,02. Powiększając wykres, można statystykę D odczytać z wystarczającą dokładnością. Tak więc serie A i B, opady w jesieni i na wiosnę, nie różnią się średnimi, nie różnią się też rozkładami.

Wnioski

Na przykładzie sekularnej serii obserwacji opadu w Krakowie rozpatrzono zagadnienie testowania równości średnich sum opadów atmosferycznych. Opady, ze względu na wyraźną skośność rozkładu, nie podlegają standardowym procedurom wymagającym rozkładu normalnego.

Zaproponowano trzy podejścia testowania istotności różnicy średnich. Pierwsze polega na nieliniowym przekształceniu obserwacji doprowadzającym rozkład skośny do normalnego. Najczęściej będzie to logarytmowanie, a ogólniej przekształcenie Boxa-Coxa. Drugie podejście opiera się na dwóch nieco różnych testach, w których sortuje się „zmieszane” dwie serie pomiarowe, obserwując, czy „wymieszały

się” symetrycznie. W jednym teście liczy się serie, w drugim statystyką testową jest suma rang jednej z serii. Trzecie podejście to standardowy test zgodności rozkładów Kołmogorowa-Smirnowa. W tym wypadku produktem dodatkowym jest wykres dwóch krzywych kumulacyjnych (dystrybuant empirycznych). Statystyka testowa D pozwala obliczyć p -wartość testu, jednak ogląd całych krzywych zwykle wnosi dodatkową wiedzę, szczególnie w przypadku niezgodności rozkładów serii.

Podejście z przekształceniem Boxa-Coxa pozwala na stosowanie wielu standardowych procedur, w szczególności stosowanie wykresów ze słupkami błędów, których interpretacja może być wtedy zgodna z naturalnym założeniem, że zakres błędu jest 68% przedziałem ufności, jak to wynika z rozkładu normalnego.

Trzy, a właściwie cztery stosunkowo niezależne podejścia stwarzają pewien kłopot nadmiaru interpretacyjnego. Jeżeli przyjąć, że w zasadzie wynikiem testu jest odpowiedź TAK/NIE dotycząca zgodności średnich, to ostateczna decyzja jest jasna jedynie w przypadku TAK, TAK, TAK, TAK, albo NIE, NIE, NIE, NIE. Każda inna, nie tak jednoznaczna sytuacja pozwala na różne interpretacje.

Zbadanie mocy testów statystycznych nie rozwiąże tej możliwej niejednoznaczności, gdyż będzie wymagało poczynienia konkretnych założeń, co do rozkładów opadu, które zawsze są tylko przybliżeniem rzeczywistości. Nieuchronnie dochodzi się tutaj, po wykonaniu ścisłych obliczeń, do obszaru wymagającego zastosowania intuicji i doświadczenia.

W artykule przedstawiono jedynie sposób praktycznej realizacji testów z wykorzystaniem arkusza kalkulacyjnego. Nie zajmowano się precyzyjnymi sformułowaniami testów i wzorami „ukrytymi” w Excelu. Mając na uwadze konkretne zastosowanie, takie postępowanie wydaje się uzasadnione. Niemniej jednak, ze względu na ze wszech miar słuszną zasadę mówiącą, że nie da się skutecznie działać w niestandardowym środowisku (jakim jest nauka) bez wiedzy podstawowej, planowane jest uzupełnienie powyższego tekstu.

Będzie ono zawierało omówienie współczesnego paradygmatu wnioskowania statystycznego na przykładzie danych klimatycznych.

Literatura

- Kożuchowski K., 1985, *Zmienność opadów atmosferycznych w Polsce w stuleciu 1881-1980*. Acta Geogr. Lodziensia, 48, 1-158.
- Niedźwiedź T., 1981, *Sytuacje synoptyczne i ich wpływ na zróżnicowanie przestrzenne wybranych elementów klimatu w dorzeczu górnej Wisły*, Rozprawy Habilitacyjne UJ, 58, 1-165.
- Niedźwiedź T., 2010, *Katalog typów cyrkulacji atmosferycznej, mas powietrznych i frontów atmosferycznych w Polsce Południowej*. Zbiór komputerowy dostępny „on line” w Katedrze Klimatologii Uniwersytetu Śląskiego <http://klimat.wnoz.us.edu.pl/>.

- Olechnowicz-Bobrowska B., 1970, *Częstość dni z opadem w Polsce*. Prace Geograficzne PAN, 86.
- Paszyński J., Niedźwiedz T., 1999, *Klimat*, [W:] L. Starkel (red.), *Geografia Polski. Środowisko przyrodnicze*. Warszawa, Wyd. Nauk. PWN.
- Prokop P., 2006, *Maksymalne opady oraz czas ich trwania na świecie i w Polsce*. Prz. Geof., 51, 2, 147–160.
- Prokop P., Walanus A., 2003, *Trends and periodicity in the longest instrumental rainfall series for the area of most extreme rainfall in the world, Northeast India*. Geographia Polonica, 76, 2, 25–35.
- Świątek M., 2009, *Wpływ cyrkulacji atmosferycznej na zmienność opadów na polskim wybrzeżu Bałtyku*. Rozprawy i Studia, Uniwersytet Szczeciński.
- Twardosz R., 2005, *Dobowy przebieg opadów atmosferycznych w ujęciu synoptycznym i probabilistycznym na przykładzie Krakowa (1886-2002)*. Instytut Geografii i Gospodarki Przestrzennej UJ, Kraków.
- Twardosz R., 2007, *Opady atmosferyczne*. [W:] D. Matuszko (red.), *Klimat Krakowa w XX wieku*. Instytut Geografii i Gospodarki Przestrzennej UJ, 127–138.
- Ustrnul Z., Czekierda D., 2000, *Cyrkulacyjne uwarunkowania opadów w Polsce*. Wiadomości IMGW, 23 (2), 3–22.
- Walanus A., 2000, *Istotność statystyczna wniosków z analiz ilościowych na przykładzie badań górnego czwartorzędu*. Geologia, Kwartalnik AGH, 26, 4, 1–59.
- Wilks D.S., 2006, *Statistical methods in the atmospheric sciences*. Academic Press, San Diego

Streszczenie

W badaniach klimatologicznych często porównuje się średnie sumy opadów w porach roku lub typach cyrkulacji. Jednak nie zawsze dokonuje się oceny istotności różnicy średnich sum. W pracy zaproponowano dwa odmienne sposoby testowania istotności różnic średnich dobowych opadów wraz z podaniem przykładów stosowania testów przy wykorzystaniu standardowych formuł arkusza MS Excel. Pierwszy ze sposobów polega na transformacji danych za pomocą przekształcenia Boxa-Coxa. Drugi natomiast opiera się na zastosowaniu testów istotności niezależnych od rozkładów, tj. testu serii Walda-Wolfowitza, testu sum rang U Manna-Whitneya i testu zgodności Kołmogorowa-Smirnowa. Podane są algorytmy realizacji wszystkich testów za pomocą standardowych formuł arkusza MS Excel (niezależnych od wersji programu).

Słowa kluczowe: sumy opadów, rozkład prawdopodobieństwa, istotność statystyczna, test statystyczny, przekształcenie Boxa-Coxa

Summary

In climate research it is frequent to search for the difference in precipitation in different circumstances. However, because of high skewness of the probability distributions of the daily precipitation the standard statistical procedures, which assume normality are inappropriate here. Two approaches to the problem are considered, the application of Box-Cox transform to get normal distribution, and application of the standard distribution-independent tests. Namely the Wald-Wolfowitz, Mann-Whitney U, and Kolmogorov-Smirnov tests are considered to be especially useful. The algorithms for all tests are given to perform them in the standard MS Excel spreadsheet (practically independent on the program version, with no use of Visual Basic).

Key words: precipitation totals, probability distribution, statistical significance, statistical test, Box-Cox transform.