Politechnika Śląska w Gliwicach Wydział Automatyki,
Elektroniki i Informatyki

Projekt z przedmiotu Metody Statystyczne

Temat nr 18

Autorzy:	Krzysztof Ból, Jonatan Chrobak, Łukasz Latusik, Witold Smaga, Michał Stolorz, Dawid Suchy, Andrzej Tenus
Prowadzący	Dr. inż. Marcin Skowronek
Rok akademicki	2018/2019
Kierunek	Informatyka
Rodzaj studiów	SSI
Semestr	4
Grupa	6
Termin oddania sprawozdania	2019-06-05
Data oddania sprawozdania	2019-06-05

1. Temat projektu

Pewien element produkowany jest w nowej i starej hali pewnego zakładu. W ramach badania wydajności pracy (w sztukach na godzinę) przy produkcji tego elementu wylosowano w każdej hali grupę pracowników i wyznaczono ich wydajność pracy. Otrzymano następujące wyniki:

W starej hali zaobserwowano następujące wydajności pracy:
36,4; 41,4; 25,7; 39,6; 40,8; 42,8; 46,4; 49,1; 47,7; 42,1; 46; 39,7; 51,7; 39,4; 39,8; 39,6; 45,2; 34,9; 41,7; 46,7; 39,8; 35; 35,8; 49,3; 42,1; 31,7; 53,3; 48,7; 47,2; 48,6; 43,9; 40,3; 39,2; 49; 44,3; 40,9; 31,7; 40,4; 22,6; 42,3; 30,3; 42,8; 54,7; 45,6; 49,8; 38,9

Wydajności pracy w nowej hali były następujące:
41,6; 43,9; 35,7; 49; 39,5; 38,9; 36,7; 29,5; 35,5; 39,3; 20,4; 37,9; 46,8; 47,8; 42,3; 42,7; 48,3; 42,7; 39,5; 48,5; 49,9; 32,9; 36,1; 45,6; 32,1; 42,7; 36,9; 59,9; 50,9; 59,5; 29,6; 50,2; 24,4; 37,8; 38,3; 39,2; 42

2. Rozwiązania

Zadanie 1.

Dokonać analizy wydajności pracy przy produkcji elementu, wyznaczając miary przeciętne, zróżnicowania, asymetrii i koncentracji. Opracować histogramy rozkładów empirycznych. Miary wyznaczyć dwoma sposobami: a) na podstawie szeregu szczegółowego, b) na podstawie szeregu rozdzielczego.

Wzory wykorzystane przy obliczaniu miar:

1) Miary przeciętne

- Średnia arytmetyczna:

$$\acute{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

- Średnia harmoniczna:

$$\acute{x}=\frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}}$$

- Średnia geometryczna:

$$\acute{x}=\sqrt[n]{\prod_{i=1}^{n}x_i}$$

- Dominanta(Moda):

  a) Dla szeregu szczegółowego:

  - Wyszukanie w szeregu wartości, która występuje najczęściej.

b) Dla szeregu rozdzielczego:

– Wyszukanie tzw. przedziału dominanty (przedziału o największej liczebności), a następnie obliczenie wartości dokładnej zgodnie z wzorem:

$$D=x_0+\frac{n_0-n_{-1}}{(n_0-n_{-1})+(n_0-n_{+1})}\cdot C_0$$

Gdzie:
x_0 – dolna wartość przedziału dominanty.
n_0 – liczebność przedziału dominanty.
n_{-1} – liczebność przedziału poprzedzającego przedział dominanty.
n_{+1} – liczebność przedziału następującego po przedziale dominanty.
c_0 – rozpiętość przedziału dominanty.

Obliczenie dominaty zostało zaimplementowane w funkcjach „dominantaSzczegolowy” oraz „dominantaRozdzielczy”.

2) Miary zróżnicowania

- Rozstęp wyników:

  $$Q_3-Q_1$$

- Rozstęp międzyćwiartkowy:

  $$Q=\frac{Q_3-Q_1}{2}$$

- Wariancja próbkowa:

  $$S^2=\frac{1}{n}\sum_{i=1}^{n}(x¿¿i-\acute{x})^2¿$$

- Odchylenie standardowe:

  $$S=\sqrt{S^2}$$

- Odchylenie od średniej (odchylenie przeciętne):

a) Dla szeregu szczegółowego:

$$d=\frac{1}{n}\sum_{i-1}^{n}|x_i-\acute{x}|$$

b) Dla szeregu rozdzielczego:

$$d=\frac{1}{n}\sum_{i-1}^{n}\left|x_i-\acute{x}\right|*¿n_i¿$$

- Odchylenie od mediany (odchylenie ćwiartkowe):

$$Q=\frac{\left(Q_3-Me\right)+\left(Me-Q_1\right)}{2}=\frac{Q_3-Q_1}{2}$$

- Współczynnik zmienności:

$$V=\frac{S}{\acute{x}}$$

3) Miary asymetrii i koncentracji

- Skośność:

$$As=\frac{1}{n}\frac{\sum_{i=1}^{n}\left(x¿¿i-\acute{x}\right)^3}{S^3}¿$$

- Kurtoza:

$$Krt=\frac{1}{n}\frac{\sum_{i=1}^{n}\left(x¿¿i-\acute{x}\right)^4}{S^4}¿$$

- Excess:

$$Ex=Krt-3$$

Ponadto do obliczenia kwantyla rzędu n dla szeregu rozdzielczego została napisana funkcja zgodnie z poniższym wzorem.

$$Q=x_{i0}+\frac{\left(poz.Q-n_{isk-1}\right)*c_{i0}}{n_{i0}},$$

Gdzie:

x_{i0} – dolna wartość przedziału kwantyla.

poz.Q – pozycja kwantyla.

n_{isk-1} – liczebność skumulowana przedziału poprzedzającego przedział kwantyla.

c_{i0} – rozpiętość przedziału kwantyla.

n_{i0} – liczebność przedziału kwantyla.

Wyniki uzyskane dla szeregu szczegółowego.

	Stara hala	Nowa hala

Średnia arytmetyczna:	41.8791667	40.93243243
Średnia harmoniczna:	40.6506551	39.05881322
Średnia geometryczna:	41.3054863	40.04055569
Kwartyl 0.25:	39.55	36.7
Kwartyl 0.75:	46.475	46.8
Mediana:	41.9	39.5
Dominanta:	brak	42.7
Rozstęp wyników:	32.1	39.5
Rozstęp międzyćwiartkowy:	6.925	10.1
Wariancja obciążona:	42.8828993	68.17948868
Wariancja nieobciążona	43.7953014	70.07336336
Odchylenie standardowe:	6.6178019	8.37098342
Odchylenie od średniej:	4.9208333	6.35222790
Odchylenie od mediany:	3.4625000	5.05000000
Współczynnik zmienności:	0.1580213	0.20450735
Skośność:	-0.6245717	-0.01035532
Kurtoza:	3.5482532	3.22403071
Excess:	0.5482532	0.22403071

Wyniki dla szeregu rozdzielczego.

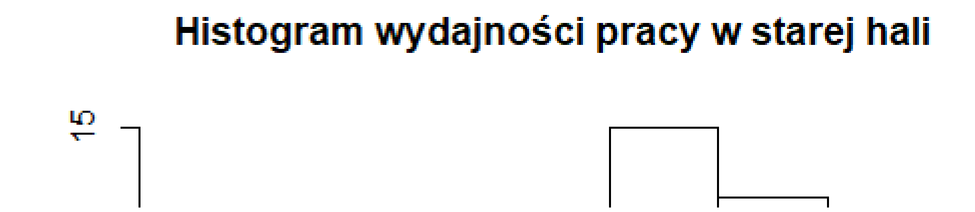
	Stara hala	**Nowa hala**
Średnia arytmetyczna:	41.6666667	40.6081081
Średnia harmoniczna:	40.5032821	38.7526490
Średnia geometryczna:	41.1200481	39.7227039
Kwartyl 0.25:	36.4705882	35.8552632
Kwartyl 0.75:	45.4444444	45.1515152
Mediana:	41.09375	38.1578947
Dominanta:	43.57142	38.235294

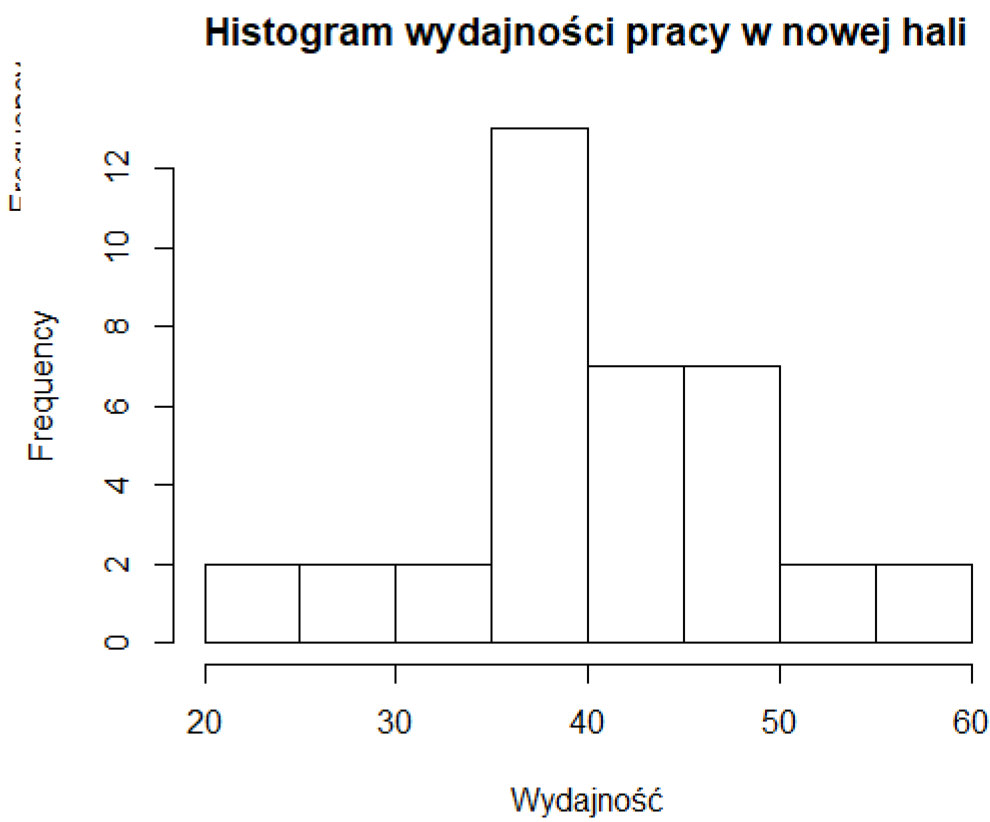
	86	1
Rozstęp wyników:	35	40
Rozstęp międzyćwiartkowy:	8.9738562	9.4098884
Wariancja obciążona:	40.9722222	66.6910153
Wariancja nieobciążona	41.8439716	68.5435435
Odchylenie standardowe:	6.4009548	8.1664567
Odchylenie od średniej:	5.0347222	6.4353543
Odchylenie od mediany:	4.4869281	4.7049442
Współczynnik zmienności:	0.1536229	0.2011041
Skośność:	-0.6399116	-0.1237806
Kurtoza:	3.3498995	3.05447
Excess:	0.3498995	0.0544700

Histogramy:

Zada

Spraw zkład
norm ności
0,95).





Tablica rozkładu wartości dla testu Kołmogorowa-Smirnowa z poprawką Lillieforsa:

	poziom α	
n	**0,01**	**0,05**
31	0,1852	0,1591
32	0,1823	0,1566
33	0,1795	0,1542
34	0,1768	0,1519
35	0,1743	0,1498
36	0,1717	0,1477
37	0,1695	**0,145**
38	0,1673	0,1437
39	0,1651	0,1419
40	0,163	0,1401
41	0,161	0,1384
42	0,1591	0,1367
43	0,1572	0,1351
44	0,1554	0,1336
45	0,1537	0,1321
46	0,152	0,1306
47	0,1504	0,1292
48	0,1488	**0,127**
49	0,1473	0,1266
50	0,1458	0,1253
51	0,1444	0,1241
52	0,143	0,1229
53	0,1416	0,1217
54	0,1403	0,1206
55	0,139	0,1193
60	0,1331	0,1144
65	0,1279	0,1099
70	0,1232	0,1059
75	0,119	0,1023
80	0,1153	0,0991
85	0,1118	0,0961
90	0,1087	0,0934
95	0,1058	0,0909
100	0,1031	0,0886

Korzystając z poniższych wzorów obliczamy wartość D:

$$D = max\, D^{+¿,D^{-¿¿}¿}$$

$$D^{+¿=\max_{i=1,\dots,n}\frac{i}{n}-p(i)¿}$$

$$D^{-¿=\max_{i=1,\dots,n}p(i)-\frac{(i-1)}{n}¿}$$

$p(i)$ – funkcja rozkładu normalnego

Porównujemy otrzymane wartości D z wartościami k uzyskanymi z tabeli – pogrubione,
w czerwonych ramkach.
Wynik działania programu:

```
Dane 1 (stara hala):
wartosc D wynosi: 0.1387923
wartosc k wynosi: 0.1279
Wydajnosci pracy nie maja rozkladu normalnego.
Dane 2 (nowa hala):
wartosc D wynosi: 0.09601999
wartosc k wynosi: 0.1457
Wydajnosci pracy maja rozklad normalny.
```

Zadanie 3.

Oszacować przedziałowo (współczynnik ufności 95) wartość przeciętną wydajności pracy produkcji elementu w starej hali. Obliczyć względną precyzję oszacowania i sprawdzić, czy mamy podstawy do uogólnienia otrzymanego przedziału ufności na całą populację wydajności pracy przy produkcji elementu w starej hali.

Ponieważ odchylenie standardowe dla całej populacji jest nieznane skorzystamy
z następującego wzoru na estymację przedziałową:

$$P\left(\bar{x}-t_{\alpha,n-1}\frac{s}{\sqrt{n}}<\mu<\bar{x}+t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right)=1-\alpha,$$

gdzie $t_{\alpha,\ n-1}$ jest wartością z tablic t-Studenta dla n-1 stopni swobody, spełniającą warunek $P(|t|< t_{\alpha,\ n-1}) = 1-\alpha$.

Obliczamy potrzebne wartości za pomocą funkcji:

- sd(stara) – aby otrzymać próbkowe oszacowanie odchylenia standardowego
- qt(0.975,n-1) – aby otrzymać wartość z tablic t-Studenta dla współczynnika ufności równego 95% przy n-1 stopniach swobody
- mean(stara) – aby otrzymać średnią wszystkich wartości zawartych w podanych danych

Następnie, aby obliczyć względną precyzję naszego przybliżenia, korzystamy ze wzoru:

$\delta=\frac{d}{\acute{x}}*100\%,$

Gdzie:
d - bezwzględny błąd szacunku.

Wynik działania programu:

```
> zad3(stara)
[1] "Interwal estymacji przedzialowej o dokladnosci 95%:"
[1] 39.9576 43.8008
[1] "Wzgledna precyzja oszacowania:"
[1] 4.5885
```

Ponieważ względna precyzja estymacji przedziałowej jest mniejsza od 5%, uprawnione jest uogólnienie wyniku na całą populację.

Zadanie 4.

Oszacować przedziałowo (współczynnik ufności 95) odchylenie standardowe wydajności pracy produkcji elementu w nowej hali. Obliczyć względną precyzję oszacowania i sprawdzić, czy mamy podstawy do uogólnienia otrzymanego przedziału ufności na całą populację wydajności pracy przy produkcji elementu w starej hali.

Średnia próby:

$$\acute{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

Wariancja próbkowa:

$$S^2=\frac{1}{n}\sum_{i=1}^{n}(x¿¿i-\acute{x})^2¿$$

Odchylenie standardowe:

$$S=\sqrt{S^2}$$

Przedział ufności:

$$P\left\{\acute{x}-\frac{u_\alpha s}{\sqrt{n}}<m<\acute{x}+\frac{u_\alpha s}{\sqrt{n}}\right\}=1-\alpha$$

Gdzie dla $1-\alpha = 0.95$: $u_\alpha = 1.96$

Względna precyzja oszacowania:

$$\delta=\frac{d}{\acute{x}}*100\%$$

Obliczamy potrzebne wartości za pomocą funkcji:
•sd(nowa) – aby otrzymać probkowe oszacowanie odchylenia standardowego
•qt(0.975,n-1) – aby otrzymać wartość z tablic t-Studenta dla współczynnika ufności równego 95% przy n-1 stopniach swobody
•mu (nowa) – aby otrzymać średnią wszystkich wartości zawartych w podanych danych

Przy pomocy funkcji *sigma<-sd(nowa)* wyznaczamy odchylenie standardowe.
Następnie przechodzimy do wyznaczenia interwału estymacji przedziałowej o dokładności 95%.
Robimy to przy pomocy funkcji:

*round(mu+c(-1,1)*sigma/sqrt(n)*qnorm(.975),2)*

Oraz względną precyzja oszacowania:

interval=mu+c(-d,d)

Wynik działania programu:

```
> zad4(nowa)
[1] "Interwal, estymacji przedzialowej o dokladnosci 95%:"
[1] 38.1414 43.7235
[1] "wzgledna precyzja oszacowania:"
[1] 6.8186
```

Ponieważ względna precyzja estymacji przedziałowej jest większa od 5% nie możemy uogólnić wyniku na całą populację.

Zadanie 5.
Czy na poziomie istotności 0,05 można twierdzić, że wartości wydajności pracy przy produkcji elementu w starej hali są większe (sformułować i zweryfikować odpowiednią hipotezę)?

Zaczynamy od wykonania testu Fishera w celu przetestowania czy wariancje rozkładów zmiennych losowych dla starej i nowej hali są sobie równe. W zależności od wyniku dobieramy odpowiedni test dla wartości oczekiwanej.

Test Fishera został przeprowadzony według wzoru:
Statystyka testowa F:

$$F=\frac{S_2^2}{S_1^2}$$

gdzie: $S^2_{¿2}$, $S^2_{¿1}$ – nieobciążone estymatory wariancji z populacji
zakładamy, że $S^2_{¿2}$, $>S^2_{¿1}$
Obszar krytyczny testu Fishera:

$K_0 = \left(f\left(0.95, n_1 - 1, n_2 - 1\right), +\infty \right)$

gdzie: $f\ (0.95\ ,\ n1-1, n2-1)$ – kwantyl rzędu 0.95 rozkładu F ze stopniami swobody n1 – 1 oraz n2 - 1

TEST FISHERA
H0: Wariancje wydajności pracy sa sobie rowne
H1: Wariancje wydajności pracy sa rozne od siebie
Statystyka testowa F = 0.6249921
Obszar krytyczny K_0 = (0.5435032 ; 1.839916
Wartosc statystyki zawiera sie w obszarze krytycznym.
Odrzucamy hipoteze zerowa na rzecz hipotezy alternatywnej.
Na poziomie istotnosci 0.05 mozna przyjac hipoteze alternatywna.

Na podstawie testu Fishera nie odrzucamy hipotezy zerowej mówiącej, że wariancje wydajności pracy w starej i nowej hali są sobie równe. Przyjmujemy, że wariancje są takie same zatem do testowania hipotezy o wartościach oczekiwanych stosujemy test T – Studenta.

Test T-Studenta został przeprowadzony według wzoru:

$$T = \frac{\acute{X}_1 - \acute{X}_2}{\sqrt{\frac{n_1 * S_1^2 + n_2 * S_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

gdzie $S_1^2\, i\, S_2^2$są nieobciążononymi estymatorami wariancji.
Natomiast obszar krytyczny jest postaci:

$$K_0 = \left(-\infty, -t\left(1 - \frac{\alpha}{2}, n_1 + n_2 - 2\right)\right) \cup \left(t\left(1 - \frac{\alpha}{2}, n_1 + n_2 - 2\right), +\infty \right)$$

Gdzie t to kwantyl rzędu 0.95 rozkładu T ze stopniem swobody $n_1 + n_2 - 2$

Weryfikujemy hipotezę H0: Średnie wydajności pracy w obu halach są sobie równe, przeciw hipotezie H1: Średnia wydajność pracy w starej hali jest większa.

TEST T-STUDENTA
H0: Srednia wydajnosc pracy w hali starej i nowej jest taka sama
H1: Srednia wydajnosc pracy w hali starej jest wieksza
Statystyka = 0.5753955
Obszar krytyczny K_0 = < 1.66342 , +oo)
Wartosc statystyka NIE miesci sie w obszarze krytycznym.
Brak podstaw do odrzucenia hipotezy zerowej.

Zatem nie można stwierdzić, że wydajność pracy w starej hali jest większa niż w nowej.

Alternatywą dla testu T-Studenta jest test Cochrana-Coxa. Stosujemy go, gdy w teście Fishera wartość statystyki nie zawiera się w obszarze krytycznym. Jest on określony wzorem:

$$T = \frac{\acute{X}_1 - \acute{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

gdzie $S_1^2 \, i \, S_2^2$ są nieobciążononymi estymatorami wariancji.
Natomiast obszar krytyczny jest postaci:

$$K_0 = ¿ \cup ¿$$

Gdzie t to kwantyl rzędu 0.95 rozkładu T ze stopniem swobody v wyliczanym ze wzoru:

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1+1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2+1}\left(\frac{S_2^2}{n_2}\right)^2}$$

Źródła:
„Wykłady z Metod Statystycznych dla Informatyków z przykładami w języku R”
Katarzyna Stąpor, Wydawnictwo Politechniki Śląskiej
„Przewodnik po pakiecie R”
Przemysław Biecek, Oficyna Wydawnicza GiS
www.rdocumentation.org