



Fundusze Europejskie
dla Rozwoju Społecznego

Dofinansowane przez
Unię Europejską



Metody analizy danych z programem Statistica

Statystyki opisowe i ich interpretacja. Graficzna prezentacja danych.

Przykład 1. Na jednym z wydziałów UP w Poznaniu przeprowadzono badanie wzrostu wśród studentów. Do pomiaru wybrano 35 losowych osób i uzyskano następujące wyniki w cm:

160	164	165	166	168	169	170	171	196
182	183	184	184	184	185	186	187	198
175	176	177	180	181	190	192	194	196
172	173	174	200	188	189	189	202	

- Wyznaczyć szereg rozdzielczy.
- Narysować histogram liczebności.
- Wyznaczyć poznane miary położenia i rozproszenia.
- Wyznaczyć 95% przedział ufności dla średniego wzrostu studentów UP w Poznaniu.
- Narysować wykres pudełkowy.

Rozwiązanie

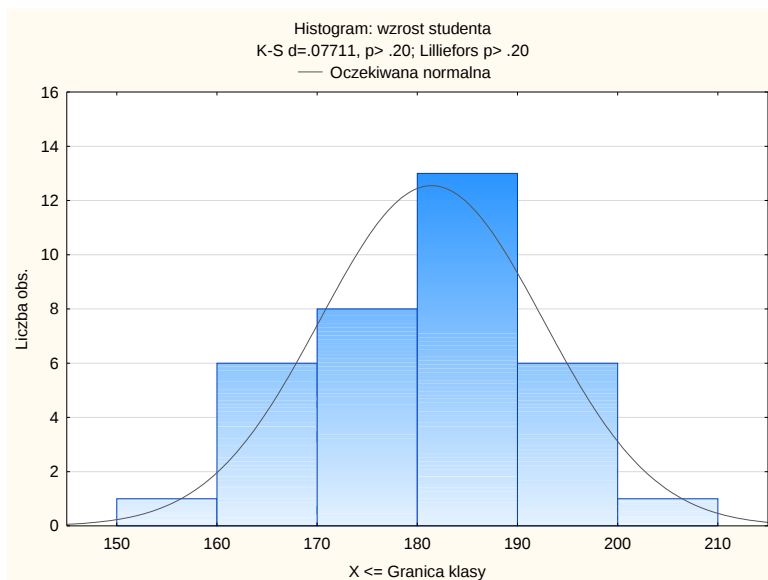
Wpisujemy dane do arkusza pakietu STATISTICA według poniższego wzorca:

	1 wzrost studenta
1	160
2	164
3	165
4	166
5	168
6	169
7	170
8	171
9	196

a) **Postępujemy według schematu:** Statystyka → Statystyki podstawowe i tabele → Statystyki opisowe → OK → deklarujemy zmienną *wzrost studenta* → Normalność → Tabele liczebności

Tabela liczebności: wzrost studenta (Arkusz1)						
K-S d=.07711, p> .20; Lilliefors p> .20						
Klasa	Liczba	Skumulow. Liczba	Procent Ważnych	Skumul. % Ważnych	% ogółu Przypadki	Skumulow. % Ogółu
150.0000<x<=160.0000	1	1	2.85714	2.8571	2.85714	2.8571
160.0000<x<=170.0000	6	7	17.14286	20.0000	17.14286	20.0000
170.0000<x<=180.0000	8	15	22.85714	42.8571	22.85714	42.8571
180.0000<x<=190.0000	13	28	37.14286	80.0000	37.14286	80.0000
190.0000<x<=200.0000	6	34	17.14286	97.1429	17.14286	97.1429
200.0000<x<=210.0000	1	35	2.85714	100.0000	2.85714	100.0000
Braki	0	35	0.00000		0.00000	100.0000

b) **Postępujemy według schematu:** Statystyka → Statystyki podstawowe i tabele → Statystyki opisowe → OK → deklarujemy zmienną *wzrost studenta* → Normalność → Histogramy



c) **Postępujemy według schematu:** Statystyka → Statystyki podstawowe i tabele → Statystyki opisowe → OK → deklarujemy zmienną *wzrost studenta* → Więcej → zaznaczamy interesujące nas okienka → Podsumowanie

Zmienna	Statystyki opisowe																
	Nważnych	Średnia	Ufność -95.000%	Ufność 95.000%	Mediana	Moda	Liczność Mody	Minimum	Maksimum	Dolny Kwartyl.	Gómy Kwartyl.	Rozstęp	Wariancja	Odch.std	Wsp.zmn.	Skośność	Kurtoza
wzrost studenta	35	181.43	177.61	185.25	183	184	3	160	202	172	189	42	123.78	11.13	6.13	-0.0318	-0.8712

Interpretacja otrzymanych wyników:

Średni wzrost studenta UPP wynosi 181,43 cm ± 11,13 cm.

Mediana (wartość środkowa) wynosi 183 cm. Oznacza to, że co drugi student UPP ma wzrost równy co najwyżej 183 cm.

Kwartyl dolny wynosi 172 cm. Oznacza to, że co czwarty student UPP ma wzrost równy co najwyżej 172 cm.

Kwartyl górny wynosi 189 cm. Oznacza to, że co czwarty student UPP ma wzrost równy co najmniej 189 cm.

Wariancja i odchylenie standardowe wzrostu studenta UPP wynoszą odpowiednio 123,78 cm² i 11,13 cm.

Współczynnik zmienności wynosi 6,13% ($\leq 10\%$), co wskazuje, że próba jest stabilna, czyli wyniki są powtarzalne i dobrze odzwierciedlają całą populację.

Najmniejszą obserwowaną wartością wzrostu studenta UPP wśród 30 obserwacji było 160 cm, a największą obserwowaną wartością było 202 cm.

Rozstęp, czyli zakres zmienności obserwacji wynosi 42 cm.

Uwaga

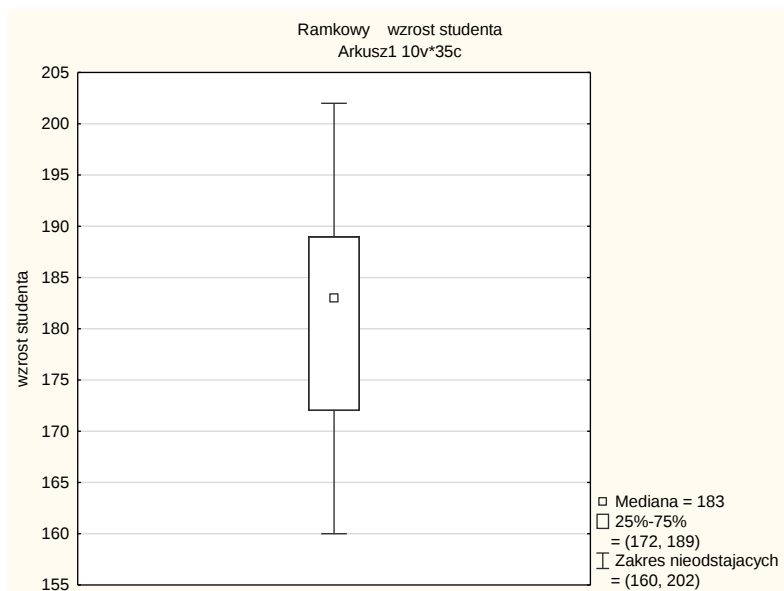
Jeżeli współczynnik zmienności przekracza 10%, oznacza to, że próba losowa nie jest stabilna. W takim przypadku uzyskane wyniki mogą nie być powtarzalne i mogą silnie zależeć od konkretnego doboru badanych, co ogranicza możliwość uogólnienia wyników na całą populację.

d) 95% przedział ufności dla średniego wzrostu studentów UP w Poznaniu jest postaci (177,61 cm; 185,25 cm).

lub

Z prawdopodobieństwem 0,95 przedział postaci (177,61 cm; 185,25 cm) pokrywa rzeczywisty średni wzrost studenta UP w Poznaniu.

e) Wykresy → Wykresy 2W → Wykresy ramka-wąsy → deklarujemy zmienną *wzrost studenta* → Jako wartość w Punkcie środkowym wybieramy **Medianę** → Ok



Zad. 1. Na jednym z wydziałów Uniwersytetu Przyrodniczego w Poznaniu przeprowadzono badanie wzrostu studentek. W ramach badania dokonano pomiaru wzrostu 30 losowo wybranych studentek, uzyskując następujące wyniki (w cm):

180	168	170	161	160	179	181	176	188	194
168	149	189	165	165	189	171	158	178	173
175	180	170	184	172	180	190	181	170	158

- Wyznaczyć szereg rozdzielczy.
- Narysować histogram liczebności.
- Wyznaczyć poznane miary położenia i rozproszenia.
- Wyznaczyć 95% przedział ufności dla średniego wzrostu studentki UP w Poznaniu.
- Narysować wykres pudełkowy.

Odp. c) i d)

Zmienna	Statystyki opisowe															
	Nw. a żnych	Średnia	Ufność -95.000%	Ufność 95.000%	Mediana	Moda	Liczność Mody	Minimum	Maksimum	Dolny Kw. artyl.	Górny Kw. artyl.	Wariancja	Odch.std	Wsp.zmn.	Skośność	Kurtoza
wzrost studentki	30	174.73	170.74	178.73	175.5	Wielokr.	3	149	194	168	181	114.34	10.69	6.12	-0.3481	-0.1349

Testowanie hipotez dla jednej lub dwóch populacji.

- Pakiet Statistica podaje nam wartość statystyki W oraz wartość „ p ”.
- Wartość „ p ” jest to najmniejszy poziom istotności, przy którym można odrzucić hipotezę zerową. Jeżeli chcemy podjąć decyzję o odrzuceniu, bądź nie, hipotezy zerowej na danym z góry ustalonym poziomie istotności α to musimy porównać wartość p z α .
- Ponieważ $p < \alpha$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności α . Stwierdzamy zatem, że...
- Ponieważ $p > \alpha$, to brak podstaw na poziomie istotności α do odrzucenia hipotezy zerowej, orzekającej, że

Lub

- Ponieważ $p > \alpha$, to brak podstaw na poziomie istotności α do odrzucenia hipotezy zerowej. Wyniki doświadczenia nie przeczą hipotezie, że ...

Test hipotezy dla średniej populacji

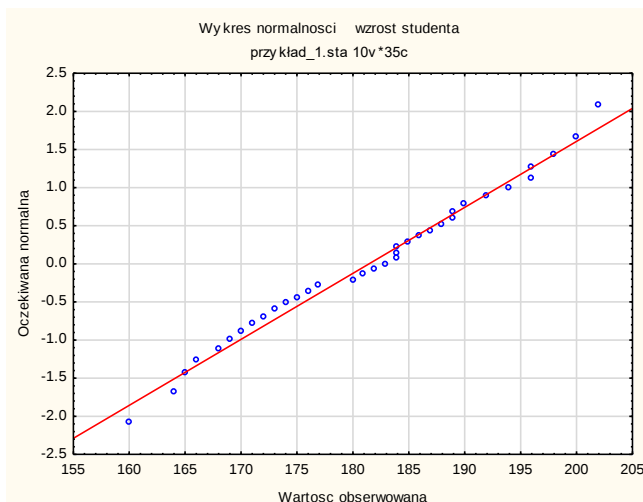
$$\begin{array}{lll} H_0: \mu = \mu_0 & H_0: \mu \geq \mu_0 & H_0: \mu \leq \mu_0 \\ H_1: \mu \neq \mu_0 & H_1: \mu < \mu_0 & H_1: \mu > \mu_0 \end{array}$$

Przykład 2.

1. Dla danych z Przykładu 1. wyznaczyć wykres normalności, wykres kwantyl - kwantyl oraz wykres prawdopodobieństwo - prawdopodobieństwo.
2. Dla danych z Przykładu 1. zweryfikuj na poziomie istotności $\alpha = 0,05$ następujące hipotezy:
 - a) Empiryczny rozkład wzrostu studentów jest zgodny z hipotetycznym rozkładem normalnym.
 - b) Średnia wysokość studenta UP w Poznaniu różni się od 175 cm.
 - c) Średnia wysokość studenta UP w Poznaniu jest większa niż 175 cm.

Rozwiązanie:

1. Aby wyznaczyć wykres normalności (wykres kwantyl - kwantyl oraz wykres prawdopodobieństwo - prawdopodobieństwo) postępujemy w następujący sposób: Wykresy → Wykresy $\underline{2}W$ → Wykresy normalności (lub Wykresy kwantyl - kwantyl, Wykresy prawdopodob. - prawdopodob.) → wprowadzamy zmienną dla rozważanej cechy → OK → OK.



Sprawdzenie, czy zebrane wyniki pomiarów mają rozkład normalny, powinno być jedną z pierwszych czynności przy rozpoczynaniu analizy zachowania się monitorowanego procesu. Bardzo skutecznym środkiem graficznego sprawdzania, czy cecha ma rozkład normalny jest jeden z wyznaczonych przez nas wykresów. Odpowiadający otrzymanemu rozkładowi wykres normalności, kwantyl - kwantyl lub wykres prawdopodob. - prawdopodob. pozwalają stwierdzić, że nasze przypuszczenia dotyczące

rozkładu normalnego są najprawdopodobniej trafne (punkty układają się wzdłuż odpowiednio wykreślonej prostej).

2.

Uwaga

Test normalności Shapiro-Wilka stosuje się do weryfikacji hipotezy o normalności rozkładu zmiennej losowej, w przypadku nieznanymi parametrów rozkładu hipotetycznego.

a)

Formułujemy hipotezy:

H_0 : Rozkład empiryczny jest zgodny z rozkładem normalnym hipotetycznym,

$H_1: \sim H_0$

STATISTICA: Statystyka → Statystyki podstawowe i tabele → Tabele licznosci → OK → deklarujemy zmienną *wzrost studenta* → wybieramy zakładkę NORMALNOŚĆ → zaznaczamy test W Shapiro-Wilka → Testy normalności

Zmienna	Testy normalności		
	N	W	p
wzrost studenta	35	0.977606	0.679962

Odp. Ponieważ $p = 0,679962 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej orzekającej, że empiryczny rozkład wzrostu studentów jest zgodny z hipotetycznym rozkładem normalnym.

b)

Formułujemy hipotezy:

$H_0: \mu = \mu_0 \quad \mu_0 = 175$

$H_1: \mu \neq \mu_0$

STATISTICA: Statystyka → Statystyki podstawowe i tabele → Test t dla pojedynczej próby → Ok → deklarujemy zmienną *wzrost studenta* → Opcje → W oknie pt. „Testuj średnie względem” wpisujemy wartość $\mu_0 (= 175)$ → Podsumowanie

Zmienna	Test średnich względem stałej wartości odniesienia							
	Średnia	Odch.st.	Ważnych	Bł. std.	Odniesienie Stała	t	df	p
wzrost studenta	181.4286	11.12571	35	1.880589	175.0000	3.418382	34	0.001652

Odp. Ponieważ $p = 0,001652 < \alpha = 0,05$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy zatem, że średnia wysokość studenta UP w Poznaniu różni się od 175 cm.

c)

$H_0: \mu \leq \mu_0 \quad \mu_0 = 175$

$H_1: \mu > \mu_0$

Uwaga: Jeżeli hipoteza mówi o większym/mniejszym, to stosujemy test jednostronny. W takim przypadku wartość p porównujemy z 2α .

Odp. Ponieważ $p = 0,001652 < 2\alpha = 0,1$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy zatem, że średnia wysokość studenta UP w Poznaniu jest większa niż 175 cm.

Zad. 2. Dla danych z Przykładu 2. zweryfikuj na poziomie istotności $\alpha = 0,05$ następujące hipotezy:

- Empiryczny rozkład wzrostu studentek jest zgodny z hipotetycznym rozkładem normalnym.
- Średnia wysokość studentki UP w Poznaniu różni się od 175 cm.
- Średnia wysokość studentki UP w Poznaniu jest mniejsza niż 175 cm.

Odp.

Zmienna	Testy normalności		
	N	W	p
wzrost studentek	29	0.971587	0.603455

Zmienna	Test średnich względem stałej wartości odniesienia							
	Średnia	Odch.st.	Ważnych	Bł. std.	Odniesienie Stała	t	df	p
wzrost studentek	171.2069	9.313505	29	1.729474	175.0000	-2.19321	28	0.036761

Test hipotezy dla dwóch średnich (przeciętnych) populacji

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$H_0: \mu_1 \geq \mu_2$$

$$H_1: \mu_1 < \mu_2$$

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Przykład 3.

Porównano dwa województwa A i B pod względem wielkości rocznej emisji SO₂ (w t/km²). W województwie A wybrano 9 zakładów szczególnie uciążliwych dla środowiska, natomiast w województwie B zakładów tych wybrano 7. Wyniki emisji poszczególnych zakładów w obu województwach przedstawia poniższa tabela:

Województwo A	5,60	5,82	5,77	5,89	5,84	5,69	6,01	5,87	6,10
Województwo B	6,00	5,77	5,69	5,82	5,79	5,88	5,42		

- Sprawdź, czy rozważane cechy mają rozkład normalny.
- Sprawdź, czy wariancje są jednorodne.
- Zweryfikować hipotezę, że rzeczywiste średnie emisje zanieczyszczeń SO₂ w obu województwach są jednakowe.

Przyjąć poziom istotności $\alpha = 0,05$.

Rozwiązanie:

a)

Formułujemy hipotezy:

H₀: Rozkład empiryczny rocznej emisji SO₂ w województwie A jest zgodny z rozkładem normalnym hipotetycznym,

H₁: ~ H₀

H₀: Rozkład empiryczny rocznej emisji SO₂ w województwie B jest zgodny z rozkładem normalnym hipotetycznym,

H₁: ~ H₀

Zmienna	Testy normalności		
	N	W	p
A	9	0.982831	0.977243
B	7	0.922540	0.489385

Odp. Ponieważ $p = 0,977243 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej orzekającej, że empiryczny rozkład rocznej emisji SO₂ w województwie A jest zgodny z hipotetycznym rozkładem normalnym.

Odp. Ponieważ $p = 0,489385 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej orzekającej, że empiryczny rozkład rocznej emisji SO₂ w województwie B jest zgodny z hipotetycznym rozkładem normalnym.

b)

Formułujemy hipotezy:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

STATISTICA: Statystyka → Statystyki podstawowe i tabele → Test t dla prób niezależnych (wzgl. zmn.) → Ok → Deklarujemy zmienne dla 1 i 2 grupy → Podsumowanie

		Testy dla prób niezależnych									
		Uwaga: Zmienne traktowane są jako niezależne próby.									
Grupa 1 wz. Grupy 2		Średnia Grupa 1	Średnia Grupa 2	t	df	p	Nważnych Grupa 1	Nważnych Grupa 2	Odch.std Grupa 1	Odch.std Grupa 2	iloraz F Wariacje
A vs. B		5.843333	5.767143	0.915326	14	0.375517	9	7	0.152151	0.181081	1.416435

Odp. Ponieważ $p = 0,631725 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej. Wyniki doświadczenia nie przeczą hipotezie, że wariacje są jednorodne.

c)

Do porównania średnich zastosujemy test t-Studenta.

Formułujemy hipotezy:

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Odp. Ponieważ $p = 0,375517 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej orzekającej, że rzeczywiste średnie emisje zanieczyszczeń SO_2 w obu województwach są jednakowe.

Przykład 4.

Badaniu statystycznemu poddano czas trwania dwóch reakcji chemicznych (A i B). Wyniki (w sekundach) były następujące:

Reakcja A	32	25	36	43	42	46	19	32	42	26	29	22
Reakcja B	39	41	40	41	34	31	37	40	37	39		

Zakładając, że czas trwania reakcji jest zmienną losową o rozkładzie normalnym, odpowiedz na pytanie czy na poziomie istotności $\alpha = 0,05$ można uważać, że średni czas trwania reakcji chemicznej A jest istotnie mniejszy od średniego czasu trwania reakcji chemicznej B?

Rozwiązanie:

Z treści zadania nie wynika, czy można przyjąć założenie o równości wariacji, dlatego rozpoczniemy od testowania wariacji.

Formułujemy hipotezy zerową i alternatywną:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

		Testy dla prób niezależnych (Arkusz25)									
		Uwaga: Zmienne traktowane są jako niezależne próby.									
Grupa 1 wz. Grupy 2		Średnia Grupa 1	Średnia Grupa 2	t	df	p	Nważnych Grupa 1	Nważnych Grupa 2	Odch.std Grupa 1	Odch.std Grupa 2	iloraz F Wariacje
Reakcja A vs. Reakcja B		32.83333	37.90000	-1.68825	20	0.106897	12	10	8.983149	3.247221	7.653032

Odp. Ponieważ $p = 0,004956 < \alpha = 0,05$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy zatem, że wariacje czasu trwania reakcji A i B różnią się między sobą istotnie.

Do porównania średnich zastosujemy test t-Studenta z poprawką Cochran-Coxa.

Formułujemy hipotezy:

$$H_0: \mu_1 \geq \mu_2$$

$$H_1: \mu_1 < \mu_2$$

STATISTICA: Statystyka → Statystyki podstawowe i tabele → Test t dla prób niezależnych (wzgl. zmn.) → Ok → Deklarujemy zmienne dla 1 i 2 grupy → Następnie wybieramy dodatkowo kartę OPCJE i w niej zaznaczamy Test t z oddzielną oceną wariancji → Podsumowanie

Testy dla prób niezależnych													
Uwaga: Zmienne traktowane są jako niezależne próby.													
Grupa 1 wz. Grupy 2	Średnia Grupa 1	Średnia Grupa 2	t	df	p	t oddz. estwar.	df	p dwustronny	Nważnych Grupa 1	Nważnych Grupa 2	Odch.std Grupa 1	Odch.std Grupa 2	iloraz F Wariacje
Reakcja A vs. Reakcja B	32.83333	37.9	-1.68825	20	0.106897	-1.81658	14.29063	0.090311	12	10	8.983149	3.247221	7.653032

Mamy w niej podane wartości „p” jednocześnie dla trzech testów dwustronnych.

Uwaga: Ponieważ przeprowadzony test dla średnich jest jednostronny, stąd wartość p porównujemy z 2α .

Odp. Ponieważ $p = 0,090311 < 2\alpha = 0,1$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy zatem, że średni czas trwania reakcji chemicznej A jest istotnie mniejszy od średniego czasu trwania reakcji chemicznej B.

Przykład 5.

Decydując się na wybór ogrzewania domu podczas jego budowy lub modernizacji stajemy przed dylematem - jakim paliwem ogrzewać? Mamy do wyboru wiele paliw. Korzystanie z gazu ziemnego jest uwarunkowane istnieniem sieci, ogrzewanie drewnem będzie bardziej opłacalne gdy mamy dostęp do taniego drewna drzew owocowych z okolicznych sadów. Wybór paliwa powinien uwzględniać również wartość opałową paliw. Dla wybranych dwóch rodzajów paliw pobrano po 8 próbek i zbadano ich wartość opałową uzyskując wyniki:

1 m ³ gazu ziemnego	1 l oleju opałowego
34,4	38,8
34,2	35,4
33,9	37,5
35,1	40,1
34,6	39,9
22,1	44,1
23,2	33,9
35,6	39,1

Stosując test mediany odpowiedzieć na pytanie, czy na poziomie istotności $\alpha = 0,05$ można uważać, że średnia wartość opałowa wybranych paliw jest jednakowa.

ROZWIĄZANIE:**KROK 1: BADANIE NORMALNOŚCI ROZKŁADU**

Układ hipotez: (dwa układy hipotez)

H_0 : Populacja „1 m³ gazu ziemnego” ma rozkład normalny

$H_1: \sim H_0$

H_0 : Populacja „1 l oleju opałowego” ma rozkład normalny

$H_1: \sim H_0$

Zmienna	Testy normalności		
	N	W	p
1 m ³ gazu ziemnego	8	0.668046	0.000975
1 l oleju opałowego	8	0.963486	0.842580

Odp. Ponieważ $p = 0,000975 < \alpha = 0,05$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy zatem, że próba pochodząca z populacji „1 m³ gazu ziemnego” nie ma rozkładu normalnego.

Ponieważ $p = 0,842580 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej orzekającej, że próba pochodząca z populacji „1 l oleju opałowego” ma rozkład normalny.

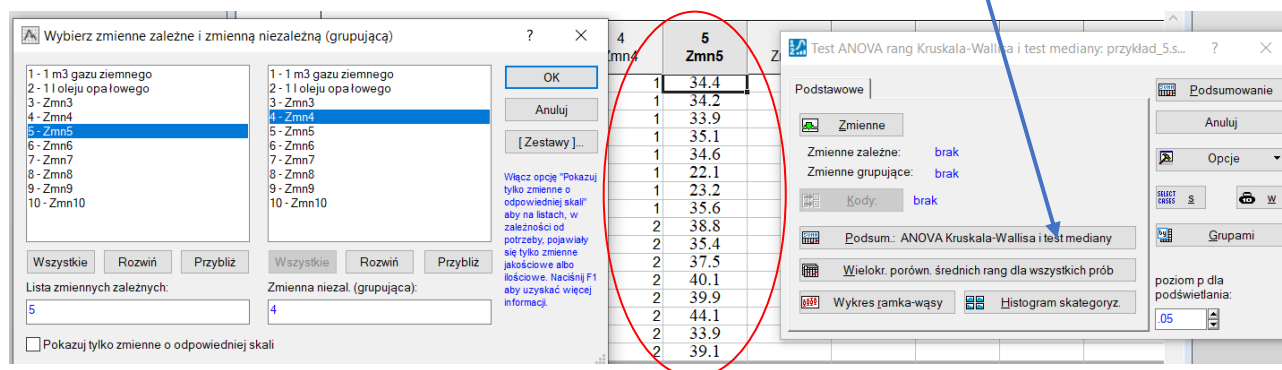
KROK 2: PORÓWNANIE DWÓCH ŚREDNICH W POPULACJACH

Układ hipotez:

H_0 : Średnia wartość opałowa „1 m³ gazu ziemnego” i „1 l oleju opałowego” jest jednakowa

$H_1: \sim H_0$

STATISTICA: Statystyka → Statystyki nieparametryczne → Porównanie wielu prób niezależnych (grup) → Wypełniamy wg wzoru → OK → Podsum: ANOVA Kruskala-Wallisa i test mediany



Odp. Ponieważ $p = 0,0027 < \alpha = 0,05$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy zatem, że średnia wartość opałowa „1 m³ gazu ziemnego” i „1 l oleju opałowego” nie jest jednakowa.

Zad. 3. Badano wpływ dwóch diet niskokalorycznych na spadek wagi ciała kobiet w wieku od 25-30 lat. Obserwacje przeprowadzono w dwóch grupach. Po czterech tygodniach stosowania diety odnotowano następujące spadki wagi ciała (w kg):

A – dieta pierwsza	5,3	5,3	5,0	4,7	4,6	4,6	4,9	5,0	5,2	4,7	5,1
B – dieta druga	4,7	4,7	5,0	4,4	4,6	5,0	4,0	4,0	4,4	4,5	4,4

Czy dieta A jest skuteczniejsza od diety B? Tzn. czy stosując dietę A średni spadek wagi będzie większy niż stosując dietę B? Przyjąć poziom istotności $\alpha = 0,05$.

Odp.

Zmienna	Testy normalności		
	N	W	p
dieta A	11	0.907975	0.230744
dieta B	11	0.925146	0.363853

Grupa 1 wz. Grupy 2	Testy dla prób niezależnych										
	Uwaga: Zmienne traktowane są jako niezależne próby.										
	Średnia Grupa 1	Średnia Grupa 2	t	df	p	Nważnych Grupa 1	Nważnych Grupa 2	Odch.std Grupa 1	Odch.std Grupa 2	iloraz F Wariancje	p Wariancje
dieta A vs. dieta B	4.945455	4.518182	3.318428	20	0.003428	11	11	0.265946	0.334120	1.578406	0.483349

Zad. 4. Porównywano zawartość magnezu (w mg/l) w wodach mineralnych dwóch rodzajów dostępnych na polskim rynku. Otrzymane obserwacje zapisano w tabeli:

Muszynianka	150	151	149	152	148	151	152	149	150	151
Galicjanka	145	156	147	145	147	149	145	148	145	146

Za pomocą testu mediany, na poziomie istotności $\alpha = 0,05$, zweryfikować hipotezę, że średnia zawartość magnezu w Muszyniance jest większa niż w Galicjance.

Odp. $p = 0,0062$

Korelacja i regresja liniowa cech ilościowych (mierzalnych).

Przykład 6.

Badając zanieczyszczenie terenów wokół pewnego obiektu przemysłowego, odsłonięto siedem profili glebowych. W powierzchniowej warstwie badanych profili zawartości ołowiu i cynku (w mg/kg) przedstawiały się następująco:

ołów (X)	355	190	345	316	269	210	275	200	300	320
cynk (Y)	82	53	93	82	67	46	80	51	76	78

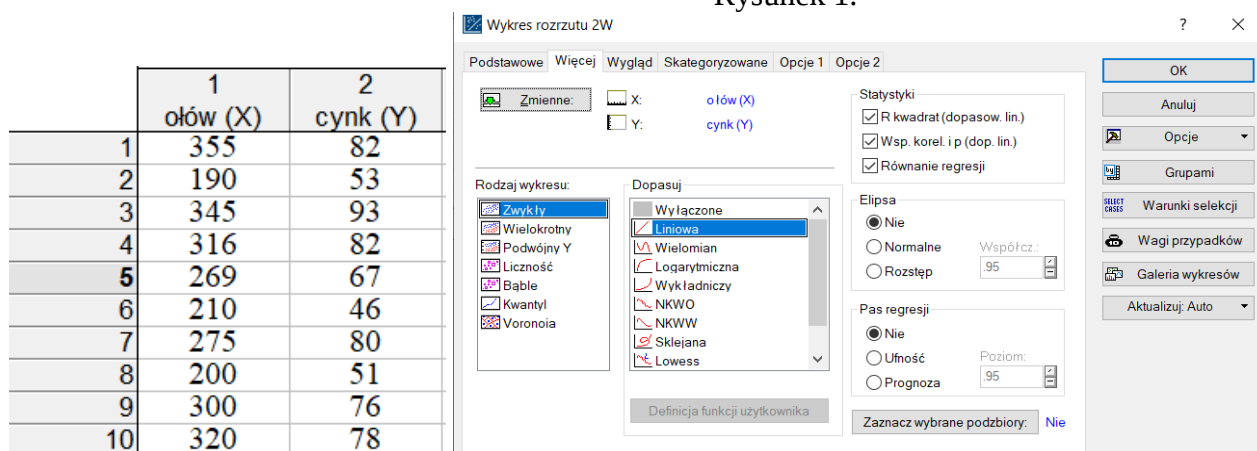
Przy prawdziwości założeń analizy regresji:

- Sprawdź hipotezę o braku korelacji między zawartością ołowiu i cynku w powierzchniowej warstwie badanych profili. Przyjmij poziom istotności 0,05.
- Oblicz i zinterpretuj współczynnik korelacji między cechami X i Y.
- Wyznacz równanie regresji liniowej zawartości cynku względem zawartości ołowiu w powierzchniowej warstwie badanych profili. Zinterpretuj współczynnik regresji.
- Oblicz i zinterpretuj współczynnik determinacji.
- Na poziomie istotności $\alpha = 0,05$ zweryfikuj, za pomocą analizy wariancji, hipotezę o braku regresji liniowej zawartości cynku względem zawartości ołowiu.
- Wyznacz na podstawie równania regresji zawartość cynku w powierzchniowej warstwie badanych profili, gdy zawartość ołowiu wynosi 270 (mg/kg).
- Zbuduj 95% przedział ufności dla oczekiwanej zawartości cynku w powierzchniowej warstwie badanych profili, jeśli zawartość ołowiu wynosi 270 mg/kg.

Rozwiązanie:

Dane należy wpisać w dwóch kolumnach

Rysunek 1.



Aby wyznaczyć współczynnik korelacji dla zmiennych Y (zawartości cynku) i X (zawartości ołowiu), zweryfikować hipotezę o braku korelacji, wyznaczyć równanie regresji i współczynnik determinacji, z menu wybieramy **Wykresy** → **Wykresy rozrzutu** → w oknie **Zmienne** wybieramy odpowiednie zmienne → **OK** → **Więcej** → **Dopasuj** → wybieramy **Liniowa** oraz statystyki nas interesujące tzn. **R kwadrat, wsp. korel. i p, równanie regresji** (Rysunek 1.) → **OK**.

- $H_0 : \rho = 0$ (korelacja między cechami jest nieistotna)
 $H_1 : \rho \neq 0$ (korelacja między cechami jest istotna)

Odp. Ponieważ $p = 0,00008 < \alpha = 0,05$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy, że korelacja między zawartościami ołowiu i cynku w glebie jest istotna.

b) Współczynnik korelacji $r = 0,9338$ jest dodatni, czyli korelacja między cechami jest dodatnia: im większa jest zawartość ołowiu w glebie, tym większa jest zawartość cynku. Wartość współczynnika korelacji jest większa niż 0,9, a zatem zależność liniowa między rozważanymi cechami jest bardzo silna.

c) Prosta regresji wyraża się wzorem $y = 2,7218 + 0,2449x$.

Jeżeli zawartość ołowiu w glebie wzrośnie o jednostkę, czyli o 1 mg/kg, to przeciętna zawartość cynku wzrośnie średnio o 0,2449 mg/kg.

d) Współczynnik determinacji $R^2 = r^2 \cdot 100\%$, czyli u nas $R^2 = 0,8720 \cdot 100\% = 87,2\%$. Oznacza to, że otrzymany model jest w 87,2% dopasowany do danych.

Lub

Współczynnik determinacji $R^2 = 87,2\%$ oznacza, że ponad 87% zmienności zawartości cynku w powierzchniowej warstwie badanych profili można wyjaśnić zmiennością zawartości ołowiu w tej warstwie, zgodnie z modelem regresji $y = 2,7218 + 0,2449x$.

e) $H_0 : \beta_1 = 0$ (regresja liniowa jest nieistotna)

$H_1 : \beta_1 \neq 0$ (regresja liniowa jest istotna)

(To testowanie przeprowadzamy wówczas, gdy wcześniej nie weryfikujemy hipotezy z punktu a)

Aby zweryfikować tę hipotezę należy z menu **Statystyka** wybrać opcję **Regresja wieloraka**, wskazać zmienną zależną Y i zmienną niezależną X → OK → OK. W oknie **Wyniki regresji wielorakiej**, należy wybrać przycisk **Więcej**, a następnie opcję **ANOVA (sum. dobroć dopasow.)**. Na ekranie pojawi się tabela analizy wariancji i wyniki testu F (Rysunek 2.):

Rysunek 2.

Efekt	Analiza wariancji ; DV: cynk (Y)				
	Suma kwadrat.	df	Średnia kwadrat.	F	p
Regres.	1958.104	1	1958.104	54.48716	0.000078
Reszta	287.496	8	35.937		
Razem	2245.600				

Rysunek 3.

Zmienna	Obliczanie wartości zmiennej: cynk (Y)		
	Wagi b	Wartość	Wagi b * Wartość
ołów (X)	0.244885	270.0000	66.12
W. wolny			2.72
Przewidyw.			68.84
-95.0%GU			64.43
+95.0%GU			73.26

Odp. Ponieważ $p = 0,000078 < \alpha = 0,05$, to odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną na poziomie istotności $\alpha = 0,05$. Stwierdzamy, że regresja liniowa zawartości cynku względem zawartości ołowiu jest istotna.

f) Wracamy do okna **Wyniki regresji wielorakiej**, wybieramy przycisk **Reszty, założenia, predykcja**, a następnie **Predykcja zmiennej zależnej** (przy aktywnym poleceniu **Oblicz granice ufności**).

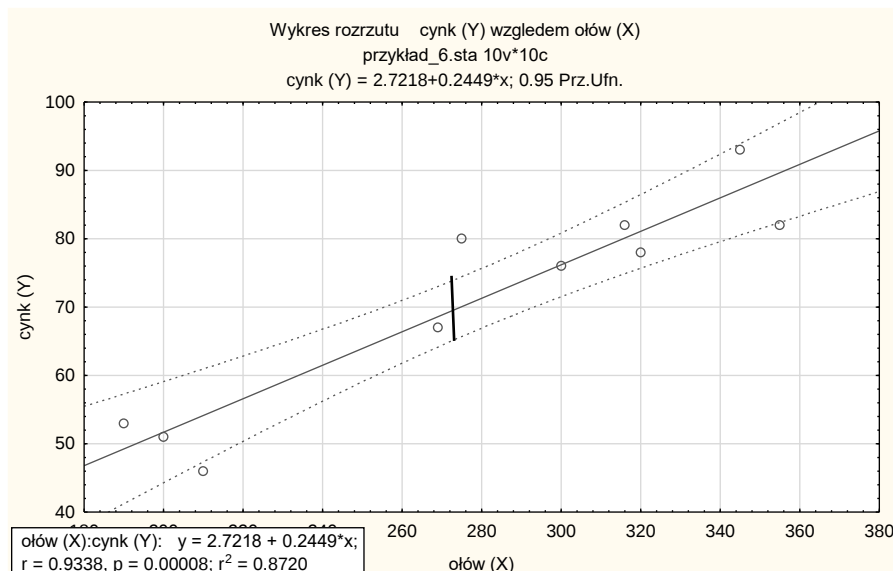
W oknie **Określ wartości zmiennych niezależnych** wpisujemy liczbę 270 i naciskamy OK. Otrzymujemy ocenę punktową prognozy oraz poszukiwane granice ufności wartości prognozowanej.

Odp. Jeśli zawartość ołowiu w powierzchniowej warstwie badanych profili wynosi 270 mg/kg, to przewidywana zawartość cynku wyniesie 68,84 mg/kg (Rysunek 3.).

g) Z tej samej tabeli (Rysunek 3.) odczytujemy lewy i prawy kraniec przedziału ufności.

Odp. Z prawdopodobieństwem 0,95 stwierdzamy, że przedział o końcach (64,43; 73,26) pokrywa oczekiwaną zawartość cynku (w mg/kg) w powierzchniowej warstwie badanych profili przy zawartości ołowiu 270 mg/kg.

Dodatkowo można wyznaczyć obszar ufności dla oczekiwanej zawartości cynku przy różnych zawartościach ołowiu i ew. zilustrować („ręcznie”) przedział ufności dla $x = 270$. Postępujemy podobnie jak w punkcie a), dodatkowo zaznaczając **Pas regresji** → **ufność**.



Przykład 7.

Dla danych zawartych w pliku o IndustrialEvaporator (STATISTICA → Pomoc → Otwórz Przykłady → Datasets) zweryfikuj hipotezę o nieistotności korelacji między badanymi parami cech.

Rozwiązanie

STATISTICA: Statystyka → Statystyki podstawowe i tabele → Macierze korelacji → Ok → w oknie: **Jedna lista zmien.** deklarujemy rozważane cechy → Ok → Podsumowanie

Zad. 5. W pewnej oczyszczalni ścieków zbadano wpływ temperatury (w $^{\circ}\text{C}$) na zawartość lotnych kwasów tłuszczowych LKT w ściekach bez zawiesin (w mg/l)

X-temperatura	16,5	29,5	30,1	30,5	20,6	23,0	25,3	18,8	26,1
Y-zawartość LKT	37,5	15,0	9,0	7,5	32,2	27,0	22,8	23,0	16,2

- Narysować wykres regresji liniowej na tle diagramu korelacyjnego.
- Wyznaczyć i zinterpretować współczynnik korelacji.
 Odp. $r = -0.914651$
- Wyznaczyć równanie regresji liniowej zawartości lotnych kwasów tłuszczowych LKT w ściekach bez zawiesin względem temperatury.
 Odp. $y = 65,6097 - 1,8162x$
- Podać interpretację współczynnika regresji b_1 .
- Podać interpretację współczynnika determinacji.
 Odp. $R^2 = 83,66\%$
- Sprawdzić, czy Przyjąć współczynnik istotności $\alpha = 0,05$.
 Odp. $p = 0,00055$
- Określić przewidywaną średnią zawartość LTK w ściekach bez zawiesin, gdy temperatura wynosi 18°C oraz wyznaczyć granice ufności, przyjąć $\alpha = 0,05$.
 Odp. 32,92; (27,12; 38,72)

Korelacje cech jakościowych (niemierzalnych).

Przykład 8. (Kala 2005, s. 87)

Podczas badania jakości jabłek oceniono owoce ze względu na uszkodzenia spowodowane przez owocówkę jabłkoweczkę (U – owoce uszkodzone, N – owoce nieuszkodzone) oraz porażenie parchem jabłoniowym (C – owoce bez plam, P – owoce z plamami). W wyniku klasyfikacji owoców uzyskano następujące liczebności:

Parch	Owocówka	
	U	N
C	29	194
P	17	68

Czy na poziomie istotności 0,05 można uznać, że badane zmienne są niezależne?

Rozwiązanie:

Formułujemy hipotezy:

H_0 : zmienne X i Y są niezależne

H_1 : zmienne X i Y są zależne

STATISTICA: Wprowadzamy dane do arkusza w dwóch kolumnach. Wybieramy: Statystyka → Statystyki podstawowe → Tabele wielodzidelcze → Zbiorcza → Określ tabele → wprowadzamy zmienne → OK → Opcje → zaznaczamy okienko „Liczebności oczekiwane” → Podsumowanie

Podsumowanie: Wylizanie liczebności			
Liczebność oznacz. komórek > 10			
Chi ² Pearsona: 2.37058, df=1, p=.123641			
Parch	Owocówka U	Owocówka N	Wiersz Razem
C	33.30519	189.6948	223.0000
P	12.69481	72.3052	85.0000
Ogół	46.00000	262.0000	308.0000

Odp. Ponieważ $p = 0,123641 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej. Wyniki doświadczenia nie przeczą hipotezie, że nie ma zależności pomiędzy uszkodzeniami spowodowanymi owocówką a występowaniem porażenia parchem jabłoniowym.

Kala R. (2005): Statystyka dla przyrodników. Wydawnictwo Akademii Rolniczej w Poznaniu, Poznań

Analiza danych ankietowych.

Przykład 9. (W. Jaszczur, K. Deręgowski, Edukacja zdalna w czasie pandemii w świetle wyników analizy statystycznej, Studia Kaliskie 9 (2021), 301-318)

Przeanalizuj poniższe pytanie ankietowe, które dotyczy oceny jakości kontaktu dydaktycznego w czasie edukacji zdalnej:

Czy w Pan(i)a opinii kontakt dydaktyczny studenta, ucznia szkoły średniej z nauczycielem w okresie edukacji zdalnej w porównaniu do zajęć tradycyjnych:

- a) był utrudniony, sporadyczny,
- b) nie było problemów z możliwością kontaktu,
- c) był na tym samym poziomie, co w edukacji tradycyjnej.

Dane znajdują się w pliku: przykład_9.

Rozwiązanie

STATISTICA: Statystyka → Statystyki podstawowe i tabele → Tabele licznosci → OK → deklarujemy zmienną: kontakt dydaktyczny → Podstawowe → Podsumowanie tabela licznosci

1. Początkowa analiza dotyczy krotności występowania poszczególnych odpowiedzi

Odpowiedź „nie było problemów z możliwością kontaktu” podano 123 razy co stanowi 40,07% osób ankietowanych.

Odpowiedź: „był utrudniony, sporadyczny” podano 140 razy co stanowi 45,60% osób ankietowanych.

Odpowiedź „był na tym samym poziomie, co w edukacji tradycyjnej” podano 44 razy co stanowi 14,33% osób ankietowanych.

Klasa	Tabela licznosci: kontakt dydaktyczny (przykład_9.sta)			
	Liczba	Skumulow. Liczba	Procent	Skumulow. Procent
nie było problemów z możliwością kontaktu,	123	123	40.07	40.07
był utrudniony, sporadyczny,	140	263	45.60	85.67
był na tym samym poziomie, co w edukacji tradycyjnej.	44	307	14.33	100.00
Braki	0	307	0.00	100.00

STATISTICA: Statystyka → Statystyki podstawowe → Tabele wielodzzielcze → Zbiorcza → Określ tabele → wprowadzamy zmienne → OK → OK → Opcje → zaznaczamy okienko „Licznosci oczekiwane” i „Procenty w kolumnach” → Podsumowanie

2. Badanie zależności pomiędzy udzieloną odpowiedzią na pytanie i płcią respondenta.

Test sprawdzający niezależność pomiędzy podaną odpowiedzią na pytanie i płcią respondenta dał wartość parametru chi-kwadrat równą 7,24 przy p-wartości równej 0,026732 to świadczy, że badane cechy są od siebie zależne, czyli istnieje statystyczna zależność pomiędzy opinią respondenta na temat kontaktu dydaktycznego z nauczycielem podczas nauki zdalnej a płcią respondenta.

Warto zauważyć, że panie rzadziej wskazywały odpowiedź „nie było problemów z możliwością kontaktu” niż panowie (73 panie co stanowi 36,32% ogółu kobiet wobec 50 panów co stanowi 47,17% ogółu mężczyzn). Co więcej, panie częściej twierdziły, że kontakt z nauczycielem był na tym samym poziomie co w edukacji tradycyjnej – odpowiedź c wskazało 36 pań co stanowi 17,91% ogółu kobiet wobec 8 panów co stanowi 7,55% ogółu mężczyzn.

Podsumowanie: Wylizanie licznosci (przyklad_9.sta)			
Chi ² Pearsona: 7.24382, df=2, p=.026732			
kontakt dydaktyczny	plec kobieta	plec męczyzna	Wiersz Razem
nie bylo problemow z mozliwoscia kontaktu,	80.53	42.47	123
byl utrudniony, sporadyczny,	91.66	48.34	140
byl na tym samym poziomie, co w edukacji tradycyjnej.	28.81	15.19	44
Ogól	201.00	106.00	307

Tabela licznosci (przyklad_9.sta)				
Tabela:kontakt dydaktyczny(3) x plec(2)				
	kontakt dydaktyczny	plec kobieta	plec męczyzna	Wiersz Razem
Liczba	nie bylo problemow z mozliwoscia kontaktu,	73	50	123
% z kolumny		36.32%	47.17%	
Liczba	byl utrudniony, sporadyczny,	92	48	140
% z kolumny		45.77%	45.28%	
Liczba	byl na tym samym poziomie, co w edukacji tradycyjnej.	36	8	44
% z kolumny		17.91%	7.55%	
Liczba	Ogól	201	106	307

3. Badanie zaleznosci pomiedzy udzielona odpowiedzi na pytanie i miejscem zamieszkania respondenta.

Test sprawdzajacy niezaleznosc pomiedzy podana odpowiedzi na pytanie i miejscem zamieszkania respondenta dal wartosc parametru chi-kwadrat rowna 4,57 przy p-wartosci rownej 0,600327, to swiadczy, ze wyniki doswiadczenia nie przecza hipotezie, ze badane cechy sa od siebie niezalezne.

Podsumowanie: Wylizanie licznosci (przyklad_9.sta)					
Chi ² Pearsona: 4.56769, df=6, p=.600327					
kontakt dydaktyczny	zamieszkanie do 5 tys.	zamieszkanie od 5 - 25 tys.	zamieszkanie powyzej 100 tys.	zamieszkanie od 25 - 100 tys.	Wiersz Razem
nie bylo problemow z mozliwoscia kontaktu,	69.31	19.63	14.42	19.63	123
byl utrudniony, sporadyczny,	78.89	22.35	16.42	22.35	140
byl na tym samym poziomie, co w edukacji tradycyjnej.	24.79	7.02	5.16	7.02	44
Ogól	173.00	49.00	36.00	49.00	307

4. Badanie zaleznosci pomiedzy udzielona odpowiedzi na pytanie i pozycja respondenta.

Test sprawdzajacy niezaleznosc pomiedzy podana odpowiedzi na pytanie i pozycja respondenta dal wartosc parametru chi-kwadrat rowna 18,4013 przy p-wartosci rownej 0,000101 to swiadczy, ze badane cechy sa od siebie zalezne, czyli istnieje statystyczna zaleznosc pomiedzy opinia respondenta na temat kontaktu dydaktycznego z nauczycielem podczas nauki zdalnej a pozycja respondenta w systemie edukacji.

Warto zauwazyć, ze studenci rzadziej wskazywali odpowiedz „byl utrudniony, sporadyczny” niz uczniowie (60 osob co stanowi 34,88% ogolu studentow wobec 80 uczniow co stanowi 59,26% ogolu uczniow). Co wiecej, studenci czesciej wskazywali ze nie bylo problemow z kontaktem z nauczycielem – odpowiedz te wskazalo 84 osob (co stanowi 48,84% wszystkich studentow bioracych udzial w ankiecie) wobec 39 uczniow (co stanowi 28,89% wszystkich uczniow bioracych udzial w ankiecie).

Podsumowanie: Wyliczanie licznosci (przyklad_9.sta)			
Chi ² Pearsona: 18.4013, df=2, p=.000101			
kontakt dydaktyczny	pozycja student	pozycja uczeń	Wiersz Razem
nie było problemów z możliwością kontaktu,	68.91	54.09	123
był utrudniony, sporadyczny,	78.44	61.56	140
był na tym samym poziomie, co w edukacji tradycyjnej.	24.65	19.35	44
Ogół	172.00	135.00	307

Tabela licznosci (przyklad_9.sta)				
Tabela:kontakt dydaktyczny(3) x pozycja(2)				
	kontakt dydaktyczny	pozycja student	pozycja uczeń	Wiersz Razem
Liczba	nie było problemów z możliwością kontaktu,	84	39	123
% z kolumny		48.84%	28.89%	
Liczba	był utrudniony, sporadyczny,	60	80	140
% z kolumny		34.88%	59.26%	
Liczba	był na tym samym poziomie, co w edukacji tradycyjnej.	28	16	44
% z kolumny		16.28%	11.85%	
Liczba	Ogół	172	135	307

Analiza wariancji doświadczeń jednoczynnikowych w dwóch układach niezależnych.

Przykład 10.

Porównywano zawartość magnezu (w mg/l) w wodach mineralnych czterech rodzajów (obiektów) dostępnych na polskim rynku. Otrzymane obserwacje zapisano w tabeli:

Muszynianka	Muszynianka plus	Piwniczanka	Galicjanka
155	139	92	91
150	130	90	90
151	132	92	91
152	135	99	99
150	137	93	101

Przy prawdziwości założenia analizy wariancji (założenia o rozkładzie normalnym zmiennej obserwowanej) :

- Na poziomie istotności $\alpha = 0,05$ zweryfikować hipotezę o jednorodności wariancji zawartości magnezu w badanych wodach. Zastosować test Bartletta.
- Sformułować odpowiednią hipotezę ogólną w układzie całkowicie losowym.
- Zweryfikować postawioną hipotezę za pomocą testu F w analizie wariancji na poziomie istotności $\alpha = 0,05$.
- Wykorzystać test Tukeya do porównań wielokrotnych średnich zawartości magnezu w wodach mineralnych (**Istotne różnice**). Zweryfikować odpowiednie hipotezy szczegółowe na poziomie istotności $\alpha = 0,05$.

Rozwiązanie

Sprawdzanie założenia o jednorodności wariancji :

Stawiamy hipotezy: zerową i alternatywną

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ rzeczywiste wariancje zawartości magnezu w badanych wodach mineralnych nie różnią się między sobą istotnie

$H_1: \sim H_0$ co najmniej dwie wariancje zawartości magnezu w badanych wodach mineralnych różnią się między sobą istotnie

Do arkusza wprowadzamy dane w następujący sposób:

	1 wody (obiekty)	2 zawartość magnezu
1	1	155
2	1	150
3	1	151
4	1	152
5	1	150
6	2	139
7	2	130
8	2	132
9	2	135
10	2	137
11	3	92
12	3	90
13	3	92
14	3	99
15	3	93
16	4	91
17	4	90
18	4	91
19	4	99
20	4	101

albo

	1 obiekty	2 zawartość magnezu
1	1	155
2	1	150
3	1	151
4	1	152
5	1	150
6	2	139
7	2	130
8	2	132
9	2	135
10	2	137
11	3	92
12	3	90
13	3	92
14	3	99
15	3	93
16	4	91
17	4	90
18	4	91
19	4	99
20	4	101

Weryfikacja hipotezy H_0 przeciwko hipotezie H_1 przy pomocy pakietu STATISTICA:

Statystyka → ANOVA → Rodzaj analizy: **Jednoczynnikowa ANOVA**; Sposób definiowania analizy:

Szybkie definiowanie → OK → Zmienne → W oknie, które się otworzy, w obrębie *Zmienne zależne* zaznaczamy zmienną: **zawartość magnezu**, a w obrębie *Czynniki jakościowe* zaznaczamy zmienną: **wody** → OK → OK → wybieramy zakładkę **WIĘCEJ WYNIKÓW** → następnie zakładkę **ZAŁOŻENIA** i w niej wybieramy test Bartletta, klikając na przycisk między innymi z tym testem.

	Testy jednorodności wariancji				
	Efekt: "wody (obiekty)"				
	Hartleya F-maks	Cochrana C	Bartlett Chi-kw.	a df	p
zawartość magnezu	6.232558	0.477718	2.785797	3	0.425843

Odp. Ponieważ $p = 0,425843 > \alpha = 0,05$, to brak podstaw na poziomie istotności $\alpha = 0,05$ do odrzucenia hipotezy zerowej, orzekającej o jednorodności wariancji.

b) Układ hipotez

H_0 : $\mu_1 = \mu_2 = \mu_3 = \mu_4$ (średnie zawartości magnezu w badanych wodach mineralnych są jednakowe)

H_1 : co najmniej dwie średnie zawartości magnezu różnią się między sobą istotnie

c) Za pomocą programu STATISTICA

wracamy do początkowego okna **ANOVA – Wyniki 1** (klikając przycisk **Mniej**) → Więcej → Wyniki jednowym.

Wyniki jednowymiarowe dla każdej ZZ Parametryzacja z sigma-ograniczeniami Dekompozycja efektywnych hipotez					
Efekt	Stopnie swobody	zawartość magnezu SS	zawartość magnezu MS	zawartość magnezu F	zawartość magnezu p
Wyraz wolny	1	280608.0	280608.0	20007.70	0.000000
wody (obiekty)	3	12878.6	4292.9	306.09	0.000000
Błąd	16	224.4	14.0		
Ogół	19	13103.0			

Wartość p

Odp. Ponieważ $p = 0,000000 < \alpha = 0,05$, to odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej na poziomie istotności $\alpha = 0,05$. Stwierdzamy, zatem, że co najmniej dwie średnie zawartości magnezu w badanych wodach mineralnych różnią się między sobą istotnie.

Ad d)

Hipotezy szczegółowe

$$H_0 : \mu_i - \mu_j = 0, \quad i, j = 1, 2, 3, 4, \quad i \neq j$$

$$H_1 : \sim H_0$$

Test Tukeya:

ANOVA - Wyniki 1 (okno podświetlone na pasku) → Więcej wyników → Post hoc → Po otwarciu się okna wybieramy opcję: **Pokaż Istotne różnice** → Test Tukeya (HSD).

Test HSD Tukeya; zmienna zawartość magnezu Przybliżone prawdopodobieństwa dla testów post hoc Błąd: MS międzygrupowe = 14.025, df = 16.000				
Nr podkl.	{1}	{2}	{3}	{4}
	151.60	134.60	93.200	94.400
1		0.000192	0.000185	0.000185
2	0.000192		0.000185	0.000185
3	0.000185	0.000185		0.956386
4	0.000185	0.000185	0.956386	

Decyzje:

Na poziomie istotności $\alpha = 0,05$, na podstawie testu Tukeya, odrzucamy hipotezy zerowe szczegółowe dla następujących par wód: Muszynianka-Muszynianka plus (bo $p = 0,000192 < \alpha = 0,05$), Muszynianka-Piwniczanka (bo $p = 0,000185 < \alpha = 0,05$), Muszynianka-Galicjanka (bo $p = 0,000185 < \alpha = 0,05$), Muszynianka plus-Piwniczanka ($p = 0,000185 < \alpha = 0,05$), Muszynianka plus-Galicjanka ($p = 0,000185 < \alpha = 0,05$),

Wnioski:

Na poziomie istotności 0,05 stwierdzamy, że prawdziwe średnie zawartości magnezu dla wypisanych powyżej par wód różnią się między sobą istotnie.

Przykład 11.

Porównywano źródła wody mineralnej m.in. pod względem zawartości kationów magnezowych (w mg/l). Doświadczenie założono w układzie **bloków losowanych kompletnych**, gdyż oznaczenia wykonywano trzema metodami. Dane są w pliku o nazwie przyklad_6_2

		źródła				
		A	B	C	D	E
Blok (metody)	m1	5,5	7,3	7,2	12,4	8,0
	m2	6,5	7	7	11,6	8,4
	m3	6	6,7	7,8	12	7,2

Przy prawdziwości założeń analizy wariancji:

- a) Zweryfikować hipotezę, że prawdziwe średnie zawartości kationów magnezowych w badanych źródłach nie różnią się istotnie. Przyjąć $\alpha=0,05$.
- b) Wykorzystać test Tukeya do utworzenia grup jednorodnych. Zweryfikować odpowiednie hipotezy szczegółowe na poziomie istotności $\alpha = 0,05$.

Rozwiązanie

a)

Układ hipotez

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ (średnie zawartości kationów magnezowych w badanych źródłach nie różnią się istotnie)

H_1 : co najmniej dwie średnie zawartości kationów magnezowych różnią się między sobą istotnie

Do arkusza wprowadzamy dane w następujący sposób:

	1 Blok	2 Obiekt	3 zawartość kationów magnezowych
1	1	1	5,5
2	2	1	6,5
3	3	1	6
4	1	2	7,3
5	2	2	7
6	3	2	6,7
7	1	3	7,2
8	2	3	7
9	3	3	7,8
10	1	4	12,4
11	2	4	11,6
12	3	4	12
13	1	5	8
14	2	5	8,4
15	3	5	7,2

Weryfikacja hipotezy H_0 przeciwko hipotezie H_1 przy pomocy pakietu STATISTICA:

Statystyka → ANOVA → Rodzaj analizy: **Jednoczynnikowa ANOVA**; Sposób definiowania analizy: **Kreator analizy** → OK → Zmienne → W oknie, które się otworzy, w obrębie **Zmienne zależne** zaznaczamy zmienną **zawartość kationów magnezowych**, a w obrębie **Czynniki jakościowe** zaznaczamy zmienne **Bloki** oraz **Obiekty** → OK → OK → Więcej → Wyniki jednowym.

Wyniki jednowymiarowe dla każdej ZZ (przykład_6_2)					
Parametryzacja z sigma-ograniczeniami					
Dekompozycja efektywnych hipotez					
Efekt	Stopnie swobody	zawartość kationów magnezowych SS	zawartość kationów magnezowych MS	zawartość kationów magnezowych F	zawartość kationów magnezowych p
Wyraz wolny	1	969,6240	969,6240	3845,171	0,000000
Bloki	2	0,0760	0,0380	0,151	0,862496
Obiekty	4	64,3627	16,0907	63,810	0,000004
Błąd	8	2,0173	0,2522		
Ogół	14	66,4560			

Wartość p

Odp. Ponieważ $p = 0,000004 < \alpha = 0,05$, to odrzucamy hipotezę zerową na korzyść hipotezy alternatywnej na poziomie istotności $\alpha = 0,05$. Stwierdzamy, zatem, że co najmniej dwie średnie zawartości kationów magnezowych dla badanych źródeł różnią się między sobą istotnie.

b)

Hipotezy szczegółowe

$$H_0: \mu_i - \mu_j = 0, \quad i, j = 1, 2, 3, 4, 5, \quad i \neq j$$

$$H_1: \sim H_0$$

ANOVA - Wyniki 1 (okno podświetlone na pasku) → Więcej wyników → Post hoc → Po otwarciu się okna jako **Efekt** wybieramy **Obiekty** → w okienku **Pokaż** wybieramy **Jednorodne grupy** → Test Tukeya (HSD).

Test HSD Tukeya; zmienna zawartość kationów magnezowych (przykład_6_2)						
Grupy jednorodne, alfa = ,05000						
Błąd: MS międzygrupowe = ,25217, df = 8,0000						
Nr podkl.	Obiekty	zawartość kationów magnezowych Średnie	1	2	3	
1	1	6,00000	****			
2	2	7,00000	****	****		
3	3	7,33333	****	****		
5	5	7,86667		****		
4	4	12,00000			****	

Interpretacja: Z tabeli wynika, że mamy trzy grupy jednorodne. Do pierwszej grupy należą obiekty 1, 2, 3. Do drugiej grupy obiekty 2, 3, 5. W trzeciej grupie jest tylko obiekt 4. Średnia zawartość kationów magnezowych dla źródła wody mineralnej D różni się istotnie od średniej zawartości kationów magnezowych dla źródła A, źródła B, źródła C oraz źródła E i jest najwyższa.

Średnia zawartość kationów magnezowych dla źródła wody mineralnej A różni się istotnie od średniej zawartości kationów magnezowych dla źródła E i jest najniższa.

Natomiast średnia zawartość kationów magnezowych dla źródła wody mineralnej A nie różni się istotnie od średniej zawartości kationów magnezowych dla źródła B oraz źródła C.

Średnia zawartość kationów magnezowych dla źródła wody mineralnej E nie różni się istotnie od średniej dla źródła B oraz źródła C.

Zad. 6. Porównywano zawartość wapnia w wodzie pitnej w pięciu miastach A, B, C, D i E.

Przeprowadzono odczyty i otrzymano:

A	9,5	9,6	9,6	9,5
B	9,0	9,4	9,4	9,2
C	8,9	8,1	8,8	9,0
D	8,8	9,2	9,0	9,0
E	6,0	8,3	7,8	7,4

Przy prawdziwości założeń analizy wariancji:

- Sformułować odpowiednią hipotezę ogólną w tym układzie.
- Zweryfikować postawioną hipotezę za pomocą testu F w analizie wariancji na poziomie istotności $\alpha = 0,05$.
- W przypadku odrzucenia hipotezy zerowej, zastosować do porównań par średnich test Tukeya. Przyjąć $\alpha = 0,05$.

Odp b) $p = 0,000159$; H_0 odrzucamy

c) A-E ($p = 0,000259 < \alpha$), B-E ($p = 0,000735 < \alpha$), D-E ($p = 0,002465 < \alpha$)

Zad. 7. Porównywano wpływ kilku rodzajów proszków grzybowych (G1, G2, G3, G4, G5) na ocenę sensoryczną konserwy mięsnej (oceniano barwę, związanie plastra, zapach i smak w punktach od 0 do 5). Ocenę sensoryczną dokonało 6 zespołów studenckich (**bloki**). Otrzymane wyniki są średnimi z każdej grupy:

		proszki grzybowe				
		G1	G2	G3	G4	G5
bloki	zespół 1	3,9	4,2	4,8	5	4,2
	zespół 2	3,6	4	5	5	4,5
	zespół 3	3,7	4,4	4,3	4,5	4,4
	zespół 4	3,6	4	4,7	4,9	4,2
	zespół 5	3,8	4,4	4,6	4,8	4,5
	zespół 6	3,4	4	4,6	4,8	4,2

Przy prawdziwości założeń analizy wariancji:

- Wykonać analizę wariancji. Wnioskować na poziomie istotności $\alpha = 0,05$.
- W przypadku odrzucenia hipotezy zerowej, zastosować do porównań par średnich test Tukeya.