Statystyka matematyczna
projekt zaliczeniowy nr 2

Krzysztof Buczyński
matematyka w ekonomii i finansach
III rok
nr indeksu : 254175

Projekt będzie dotyczył badania stanu zdrowia Amerykanów poprzez wybranie kilku interesujących nas kwestii. Na potrzeby pracy zebrano dane 1345 osób.

Będziemy sprawdzać preferencje dotyczące spożywania alkoholu i ich ewentualne związki z płcią badanego, zawartość erytrocytów we krwi respondenta i jej związki z płcią oraz przeprowadzimy podstawową analizę statyczną zawartości hemoglobiny we krwi badanych osób.

Do badań wykorzystamy odpowiednie wykresy, statystki i tabele. Do testowania hipotez użyjemy adekwatnych testów statystycznych na poziomie istotności 0,05. Na koniec, przy użyciu znanych metod statystyki matematycznej, sprawdzimy, do jakiego rozkładu najbardziej zbliżony jest rozkład współczynnika zawartości hemoglobiny we krwi badanych ludzi.

Zadanie 1:

Wykonaj tabelę częstości oraz wykres porównujący odsetek osób unikających i nieunikających spożywania alkoholu w badanej próbie. Stosując test t-Studenta oraz test chi-kwadrat zgodności, zbadaj prawdziwość stwierdzenia, że 15% Amerykanów unika spożywania alkoholu.

Zadanie 1, etap 1 : częstości.

unikanie spożywania alkoholu

		Częstość	Procent	Procent ważnych	Procent skumulowany
Ważne	nie	1126	83,7	83,7	83,7
	tak	219	16,3	16,3	100,0
	Ogółem	1345	100,0	100,0	

Widać wyraźnie, że większość badanych osób nie unika spożywania alkoholu - jedynie około 200 osób (z tabeli odczytujemy, że dokładnie 219 osób).

Zadanie 1, etap 2 : testy.

Test t-Studenta dla jednej średniej:

Hipoteza zerowa H0 : Średni odsetek Amerykanów unikających spożywania alkoholu jest równy 15%.
Hipoteza alternatywna H1: Średni odsetek Amerykanów unikających spożywania alkoholu jest różny od 15%.

Statystyki dla jednej próby

	N	Średnia	Odchylenie standardowe	Błąd standardowy średniej
unikanie spożywania alkoholu	1345	,16	,369	,010

Test dla jednej próby

	Wartość testowana = 0.15					
	t	df	Istotność (dwustronna)	Różnica średnich	95% przedział ufności dla różnicy średnich	
					Dolna granica	Górna granica
unikanie spożywania alkoholu	1,273	1344	,203	,013	-,01	,03

Z tabeli odczytujemy, że p-wartość równa 0,203 jest większa od zakładanej tzn. 0,05. Stąd nie mamy podstaw do odrzucenia hipotezy H0 na rzecz hipotezy H1.

Gdybyśmy przyjęli hipotezy alternatywne postaci H2: średni odsetek Amerykanów unikających spożywania alkoholu jest mniejszy od 15% oraz H3: średni odsetek Amerykanów unikających spożywania alkoholu jest większy niż 15%, to analizując dane z tabelki, otrzymamy p-wartość jednostronną równą : 0,203/2 = 0,1015, co nadal jest większe od zakładanej istotności 0,05. Zatem i w tych wypadkach nie ma podstaw do odrzucenia hipotezy H0.

Na podstawie powyższego rozumowania można przyjąć, że średni odsetek Amerykanów unikających spożywania alkoholu wynosi 15%.

Test chi-kwadrat zgodności.

Hipoteza zerowa H0 : Prawdopodobieństwo, że wybrany Amerykanin unika spożywania alkoholu, wynosi 0,15 (czyli P(X=1)=0,15, zaś P(X=0)=1-0,15=0,75, gdzie X to zmienna losowa opisująca unikanie lub niespożywania alkoholu).
Hipoteza alternatywna H1 : Prawdopodobieństwo, że wybrany Amerykanin unika spożywania alkoholu jest inne niż 0,15.

Podsumowanie testu hipotezy

	Hipoteza zerowa	Test	Istotność	Decyzja
1	Kategorie zmiennej unikanie spożywania alkoholu występują z określonym prawdopodobieństwem	Test chi-kwadrat dla jednej próby	,705	Przyjmij hipotezę zerową.

Przedstawiono asymptotyczne istotności. Poziom istotności wynosi ,05.

Test ten potwierdził naszą wcześniejszą opinię o słuszności przyjęcia hipotezy zerowej.

Zadanie 2:

Wykonaj tabelę krzyżową, w której przedstawiona będzie zależność unikania bądź niespożywania alkoholu wynikająca z płci badanych. Wykonując test chi-kwadrat niezależności sprawdź, czy wspomniana zależność występuje w całej populacji. Wyznacz współczynnik korelacji Pearsona i określ siłę tej zależności. Zaprezentuj rozkład łączny badanych zmiennych na trójwymiarowym wykresie słupkowym.

Zadanie 2, etap 1: płeć a unikanie alkoholu.

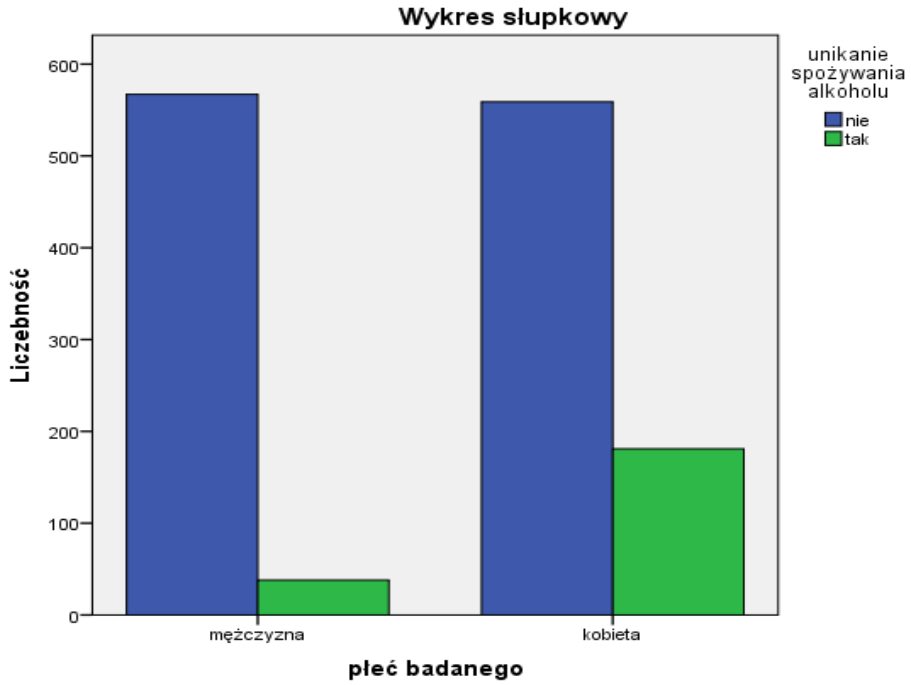
Informacja o analizowanych danych

	Obserwacje					
	Uwzględnione		Wykluczone		Ogółem	
	N	Procent	N	Procent	N	Procent
płeć badanego * unikanie spożywania alkoholu	1345	100,0%	0	0,0%	1345	100,0%

Tabela krzyżowa płeć badanego * unikanie spożywania alkoholu

Liczebność

		unikanie spożywania alkoholu		Ogółem
		nie	tak	
płeć badanego	mężczyzna	567	38	605
	kobieta	559	181	740
Ogółem		1126	219	1345



Widzimy, że więcej kobiet niż mężczyzn unika spożywania alkoholu.

Zadanie 2, etap 2: zależność w całej populacji.

Hipoteza zerowa H0: płeć badanej osoby nie ma wpływu na unikanie przez nią spożywania alkoholu.
Hipoteza alternatywna H1: unikanie spożywania alkoholu zależy od płci (dokładniej: kobiety częściej niż mężczyźni unikają spożywania alkoholu).

Testy Chi-kwadrat

	Wartość	df	Istotność asymptotyczna (dwustronna)	Istotność dokładna (dwustronna)	Istotność dokładna (jednostronna)
Chi-kwadrat Pearsona	80,694[a]	1	,000		
Poprawka na ciągłość[b]	79,366	1	,000		
Iloraz wiarygodności	87,981	1	,000		
Dokładny test Fishera				,000	,000
Test związku liniowego	80,634	1	,000		
N Ważnych obserwacji	1345				

a. ,0% komórek (0) ma liczebność oczekiwaną mniejszą niż 5. Minimalna liczebność oczekiwana wynosi 98,51.

b. Obliczone wyłącznie dla tabeli 2x2.

Istotność testu wynosi 0,00, więc jest mniejsza od zakładanej 0,05. Pozwala to na odrzucenie hipotezy H0 o niezależności unikania spożywania alkoholu od płci respondenta. Wyniki testu można uznać za wiarygodne, gdyż wszystkie komórki mają liczebności oczekiwane większe od 5.

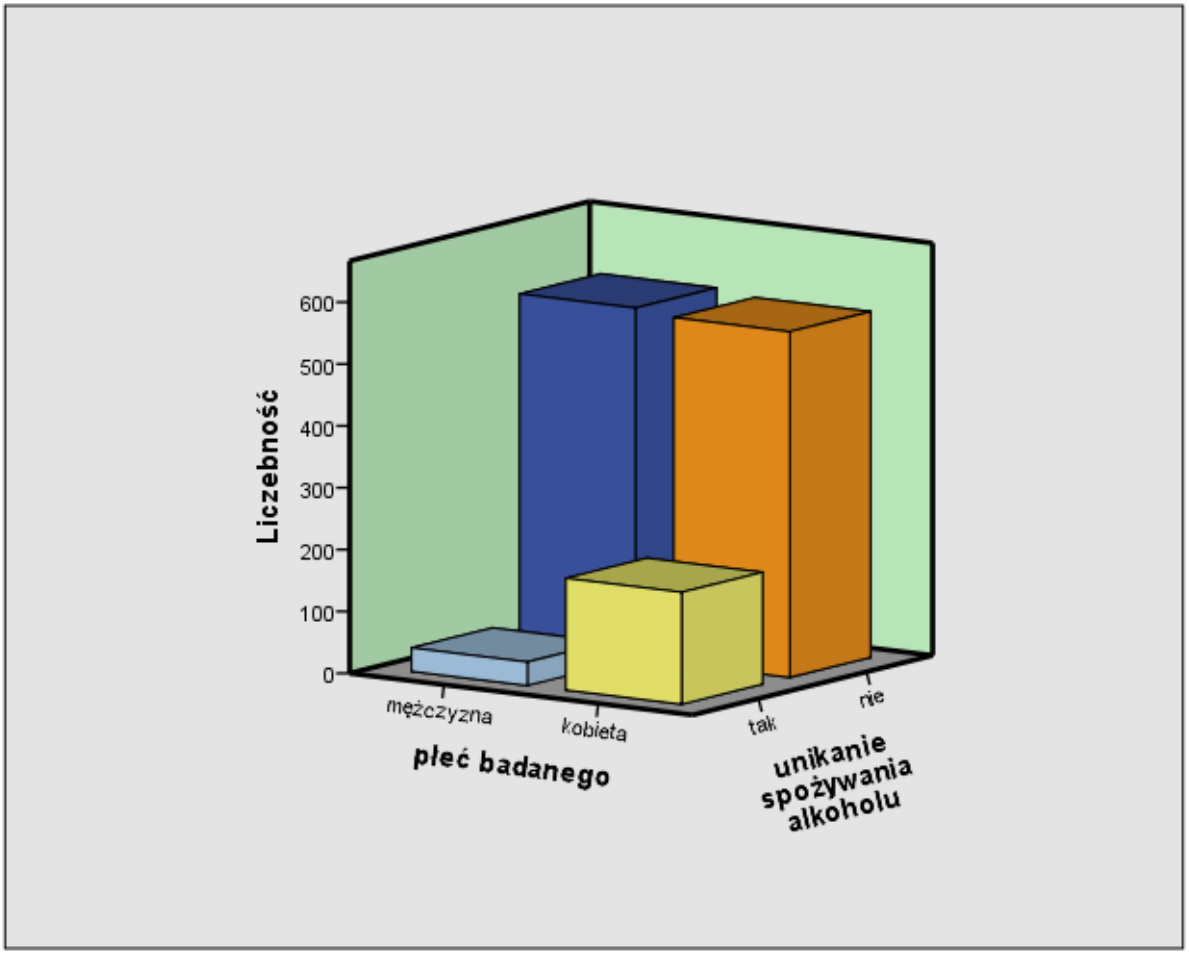
W związku z tym przyjmujemy, że kobiety częściej niż mężczyźni unikają spożywania alkoholu.

Korelacje dla prób zależnych

		N	Korelacja	Istotność
Para 1	płeć badanego & unikanie spożywania alkoholu	1345	,245	,000

Współczynnik korelacji (zależności) wynosi 0,245. Informuje nas to, że występująca zależność jest stosunkowo nieduża (wartość poniżej 0,5).

Zadanie 2, etap 3 : wykres.



Z wykresu można odczytać m.in. potwierdzenie naszych wcześniejszych tez, że więcej kobiet niż mężczyzn unika spożywania alkoholu.

Zadanie 3 :

Przeprowadź analizę liczby erytrocytów we krwi badanych z uwzględnieniem podziału na płeć. Wyznacz podstawowe statystyki opisowe, wykonaj histogramy oraz wykresy skrzynkowe. Określ kształt rozkładów. Na odpowiednim wykresie porównaj średnią liczbę erytrocytów we krwi obu płci. Skonstruuj 95% przedział ufności dla wartości oczekiwanej liczby erytrocytów we krwi Amerykanów i Amerykanek. Wykonując odpowiedni test, sprawdź, czy prawdziwe jest stwierdzenie, że w populacji Amerykanów mężczyźni cechują się wyższą liczbą erytrocytów we krwi niż kobiety.

Zadanie 3, etap 1 : podstawowa analiza.

Podstawowe statystyki. Liczba erytrocytów (czerwonych ciałek krwi) w mln/mm3

płeć badanego			Statystyka	Błąd standardowy
mężczyzna	Średnia		4,8008	,02089
	95% przedział ufności dla średniej	Dolna granica	4,7597	
		Górna granica	4,8418	
	5% średnia obcięta		4,8083	
	Mediana		4,8400	
	Wariancja		,264	
	Odchylenie standardowe		,51394	
	Minimum		2,97	
	Maksimum		6,60	
	Rozstęp		3,63	
	Rozstęp ćwiartkowy		,63	
	Skośność		-,262	,099
	Kurtoza		,418	,198
kobieta	Średnia		4,4517	,01628
	95% przedział ufności dla średniej	Dolna granica	4,4198	
		Górna granica	4,4837	
	5% średnia obcięta		4,4532	
	Mediana		4,4650	
	Wariancja		,196	
	Odchylenie standardowe		,44279	
	Minimum		2,87	
	Maksimum		5,89	
	Rozstęp		3,02	
	Rozstęp ćwiartkowy		,58	
	Skośność		-,074	,090
	Kurtoza		,236	,179

Uwaga ogólna - podane liczby erytrocytów we krwi wyrażone są w mln/mm3.

1. Średnia : średnia liczba erytrocytów we krwi Amerykanina wynosi 4,8008 (przy błędzie standardowym równym 0,02089).U Amerykanek średnia ta jest niższa i wynosi 4,4198 (przy błędzie standardowym 0,1628).

2. Minimum i maksimum : wartości skrajne to odpowiednio - u mężczyzn 2,97 i 6,60 oraz u kobiet 2,87 oraz 5,89. Oba wskaźniki są mniejsze w przypadku kobiet niż mężczyzn.

3. Mediana : u mężczyzn wynosi 4,8400. U kobiet 4,4650. W obu przypadkach średnie są niższe niż mediana, to znaczy, że 50% badanych w obu grupach ma większe niż średnia stężenie erytrocytów we krwi.

4. Odchylenie standardowe : u mężczyzn wynosi 0,51394 i jest większe niż u kobiet, gdzie wynosi 0,44279. Oznacza to, że stężenie erytrocytów we krwi u mężczyzn jest bardziej zróżnicowane niż u kobiet.

5. Skośność : dla obu grup skośność jest ujemna, więc rozkłady będą o lewostronnej asymetrii (wydłużony lewy "ogon" rozkładu). Niemniej u mężczyzn wartość ta wynosi -0,262, więc występuje tylko lekka asymetria. W przypadku kobiet tej asymetrii nie ma prawie w ogóle, gdyż skośność wynosi -0,074. Analiza histogramów i położenia lewej części rozkładu potwierdza wyżej przeprowadzone rozumowanie.

6. Kurtoza : w obu przypadkach kurtozy są nieznacznie dodatnie (odpowiednio : 0,418 oraz 0,236), więc rozkład będzie lekko wysmukły, tzn. skupienie wartości wokół średniej jest większe niż w rozkładzie normalnym.

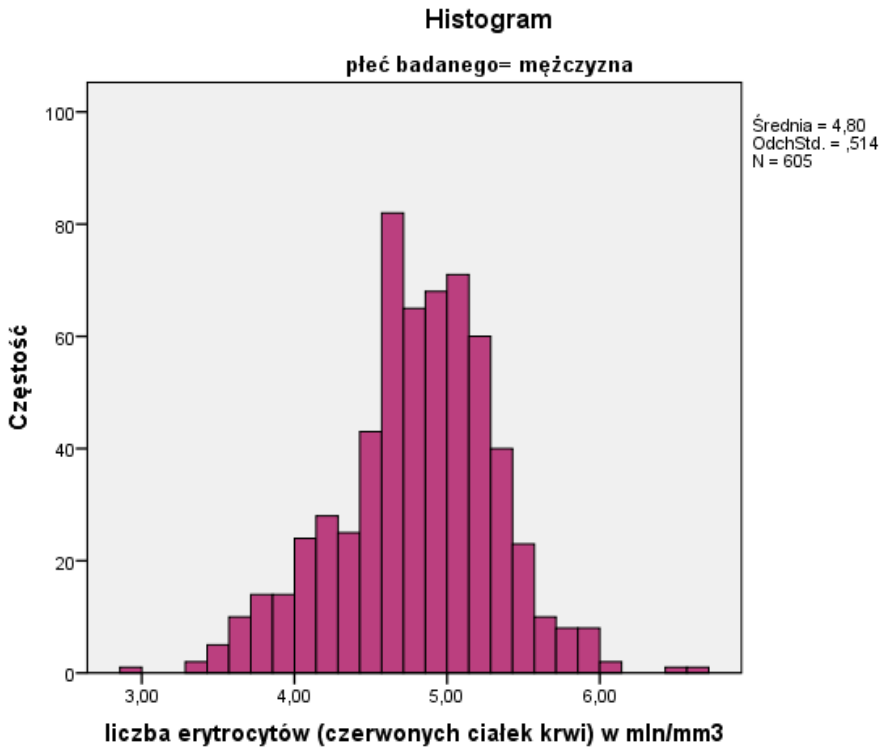
7. Rozstęp : różnica pomiędzy największym a najmniejszym stężeniem erytrocytów we krwi dla obu grup wynosi:

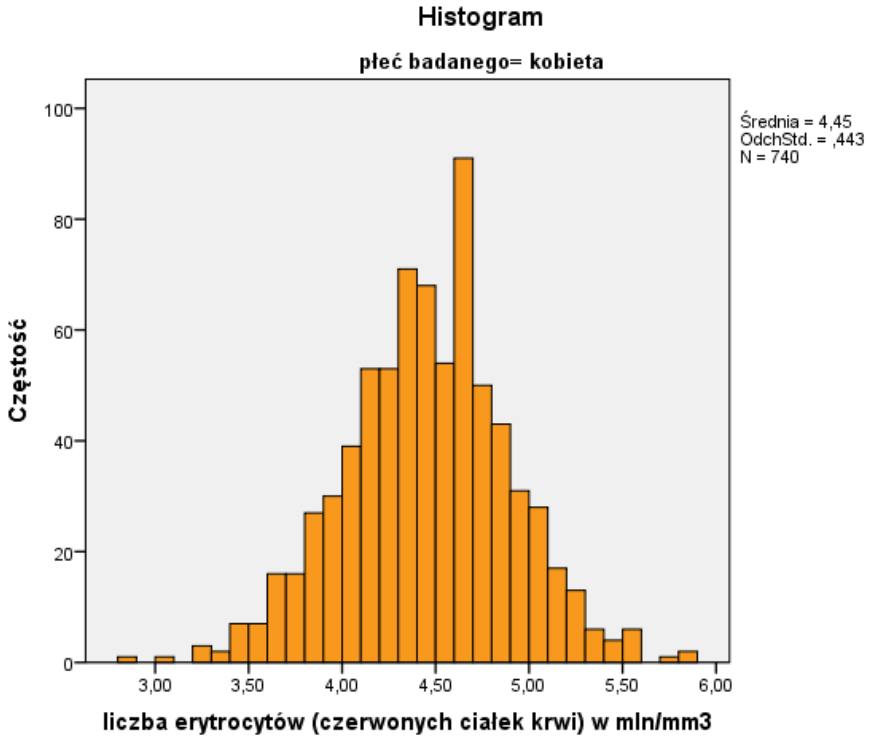
dla mężczyzn : 3,63
dla kobiet : 3,02.

8. Przedziały 95% ufności dla wartości oczekiwanej liczby erytrocytów we krwi Amerykanów i Amerykanek są równe:

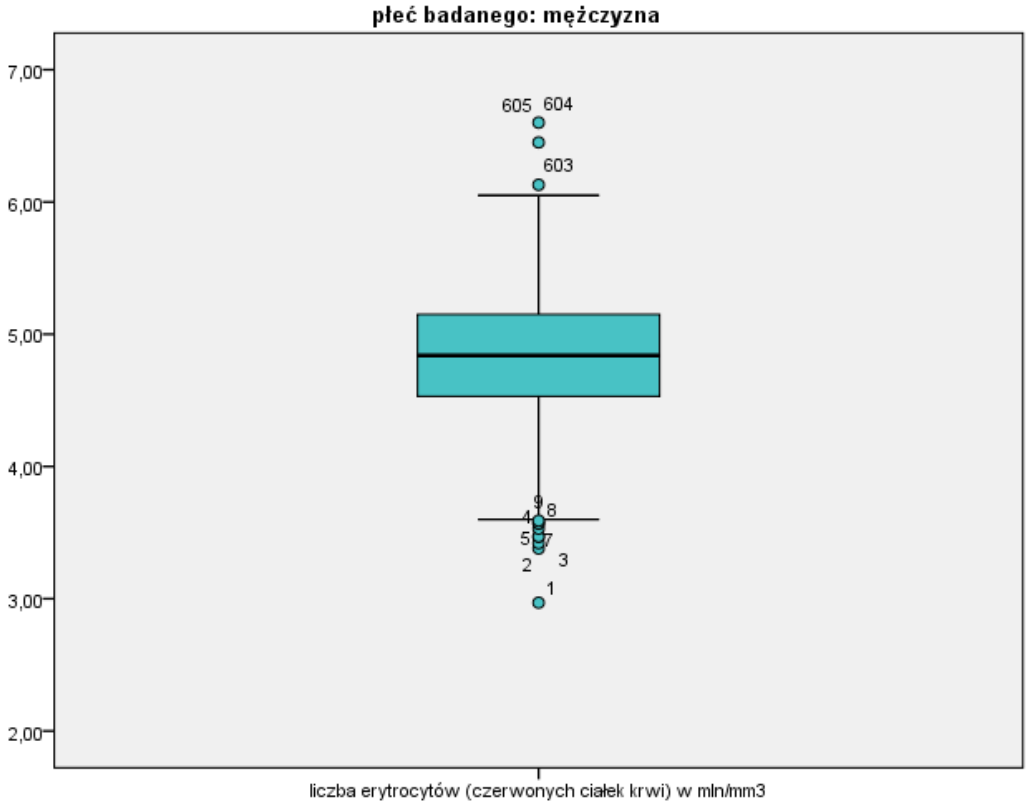
dla mężczyzn : [4,7597 ; 4,8418]
dla kobiet : [4,4198 ; 4,4837]

Uwaga – można było wyznaczyć przedziały, gdyż liczność próbek jest duża (powyżej 30).



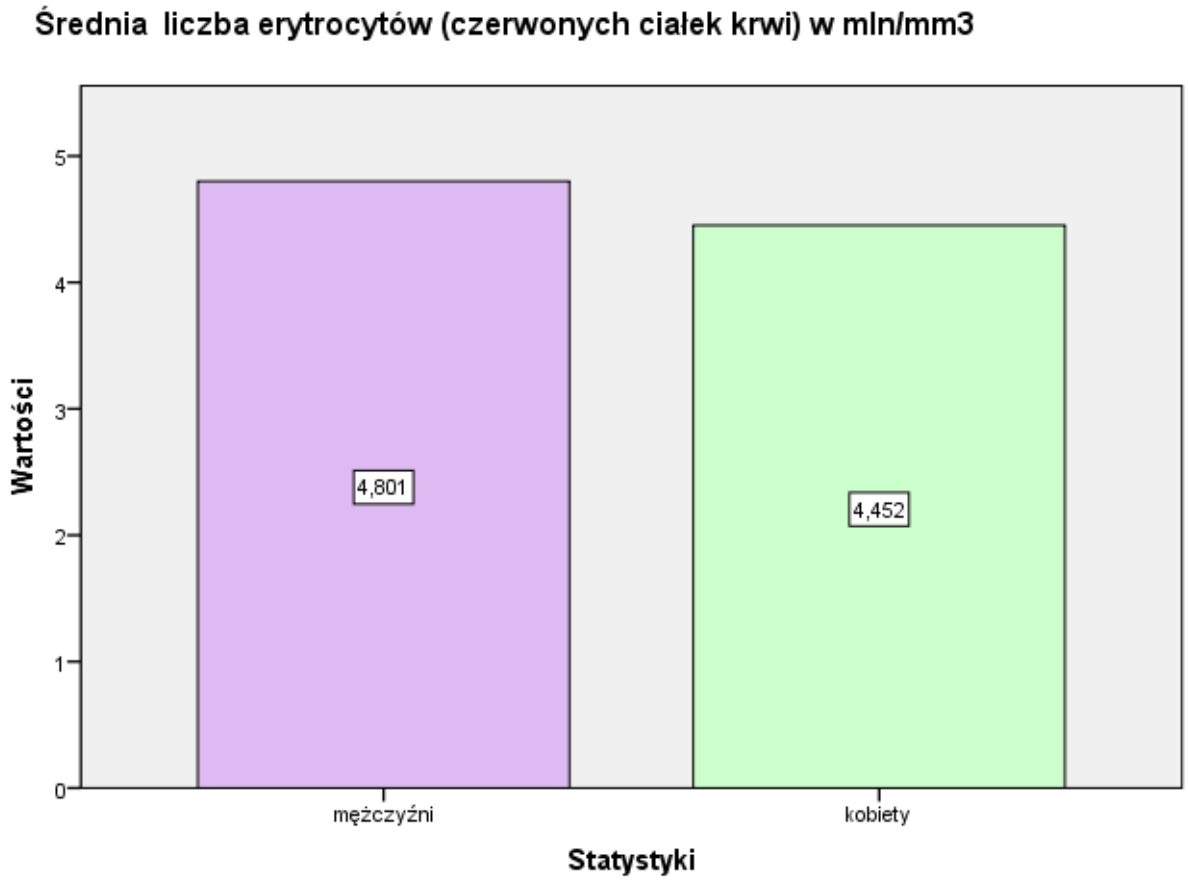


Analiza histogramów potwierdza rozważania na temat kształtów wykresów oparte o analizę kurtozy i skośności.


płeć badanego: mężczyzna
7,00
6,00
5,00
4,00
3,00
2,00
605
604
603
9
8
4
5
7
2
3
1
liczba erytrocytów (czerwonych ciałek krwi) w mln/mm3


płeć badanego: kobieta
6,00
5,00
4,00
3,00
2,00
1 345
1 344
1 343
610
609
607
608
606
liczba erytrocytów (czerwonych ciałek krwi) w mln/mm3

Zadanie 3, etap 2: wykres dla średnich.



Wykres jednoznacznie potwierdza nasze wcześniejsze uwagi o tym, że średnia w przypadku mężczyzn jest wyższa niż u kobiet.

Zadanie 3, etap 3 : test hipotezy.

Hipoteza zerowa H0 : w populacji Amerykanów mężczyźni cechują się taką samą liczbą erytrocytów we krwi jak kobiety.
Hipoteza alternatywna H1 : w populacji Amerykanów mężczyźni cechują się wyższą liczbą erytrocytów we krwi niż kobiety.

Wykonamy test t-Studenta dla dwóch średnich prób niezależnych.

Statystyki dla grup

płeć badanego		N	Średnia	Odchylenie standardowe	Błąd standardowy średniej
liczba erytrocytów (czerwonych ciałek krwi) w mln/mm3	mężczyzna	605	4,8008	,51394	,02089
	kobieta	740	4,4517	,44279	,01628

Test dla prób niezależnych

			liczba erytrocytów (czerwonych ciałek krwi) w mln/mm3	
			Założono równość wariancji	Nie założono równości wariancji
Test Levene'a jednorodności wariancji	F		9,675	
	Istotność		,002	
Test t równości średnich	t		13,375	13,178
	df		1343	1198,720
	Istotność (dwustronna)		,000	,000
	Różnica średnich		,34903	,34903
	Błąd standardowy różnicy		,02610	,02649
	95% przedział ufności dla różnicy średnich	Dolna granica	,29784	,29707
		Górna granica	,40023	,40100

Ze względu na jednostronną hipotezę alternatywną patrzymy na połowę istotności podanej w tabeli, tzn. 0,000/2=0,000 co jest mniejsze od zakładanej 0,05. Odrzucamy zatem hipotezę zerową na rzecz alternatywnej.

Dodatnia wartość statystyki t pozwala przyjąć hipotezę, że w populacji Amerykanów mężczyźni cechują się wyższą liczbą erytrocytów we krwi niż kobiety.

Zadanie 4 :

Wykonaj podstawową analizę statystyczną zawartości hemoglobiny we krwi. Do którego spośród rozkładów : normalnego, jednostajnego i wykładniczego jest najbardziej zbliżony rozkład tego współczynnika? Wykonując odpowiedni test, sprawdź, czy rzeczywiście ma miejsce zgodność rozkładów.

Zadanie 4, etap 1 : podstawowa analiza statystyczna.

Statystyki

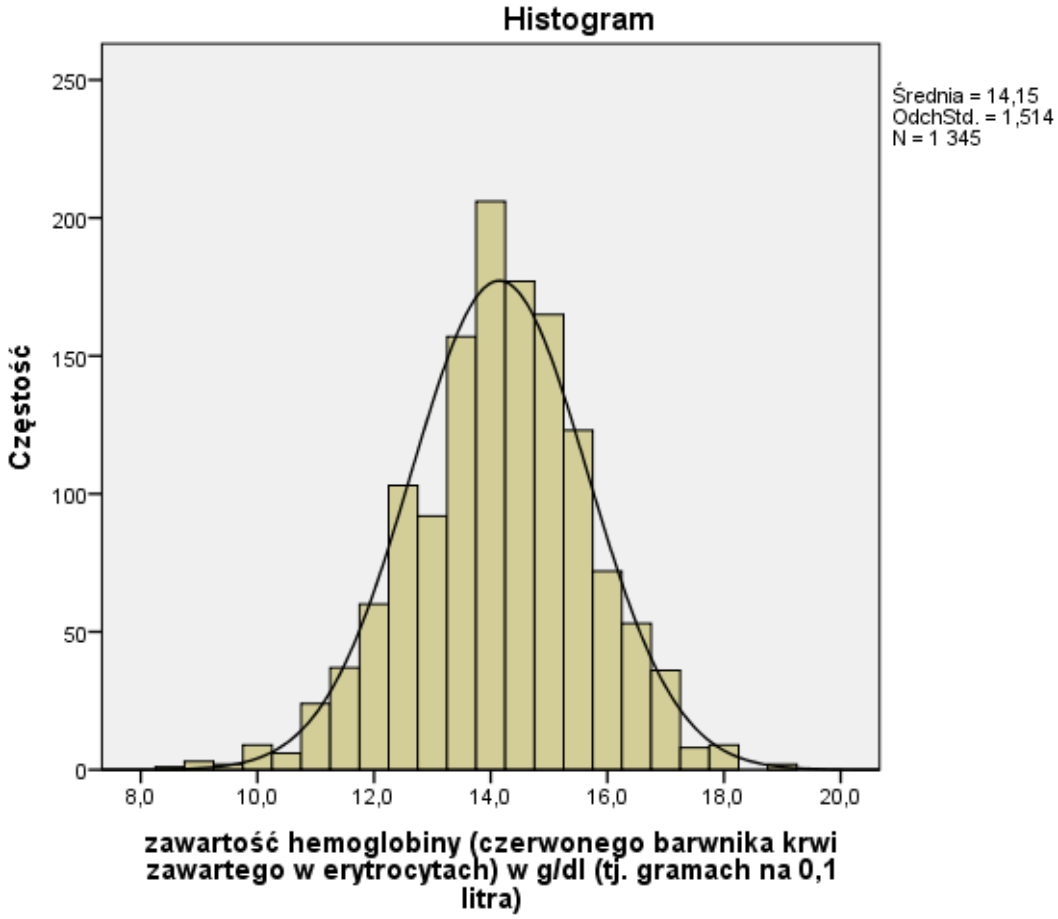
zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra)

N	Ważne	1345
	Braki danych	0
Średnia		14,148
Mediana		14,100
Odchylenie standardowe		1,5135
Wariancja		2,291
Skośność		-,233
Kurtoza		,328
Rozstęp		10,5
Minimum		8,5
Maksimum		19,0

U badanych respondentów średnia zawartość hemoglobiny we krwi wynosi 14,148 i jest większa od mediany równej 14,100, tzn. że 50% badanych ma zawartość hemoglobiny mniejszą niż średnia. Warto zwrócić uwagę na wartości maksymalną i minimalną równe odpowiednio 19,0 oraz 8,5, które znacznie różnią się od średniej. Rozstęp wynosi 10,5. Niemniej odchylenie standardowe na poziomie 1,5135 mówi nam o niezbyt dużych odchyleniach pozostałych danych od wartości średniej.

Skośność równa -0,233 < 0 informuje nas o lekkiej lewostronnej asymetrii. Kurtoza na poziomie 0,328 > 0 wskazuje na wysmukłości rozkładu, tzn. skupienie wartości wokół średniej jest większe niż w rozkładzie normalnym.

Przedstawiony poniżej histogram z krzywą rozkładu normalnego potwierdza powyższe wnioski. Możemy więc przypuszczać, że rozkład współczynnika hemoglobiny we krwi jest zbliżony do rozkładu normalnego. Wykonamy odpowiedni test.



Zadanie 4, etap 2 : test na zgodność rozkładów.

Hipoteza zerowa H0 : badana zmienna ma rozkład normalny.
Hipoteza alternatywna H1 : badana zmienna ma inny rozkład.

Informacja o analizowanych danych

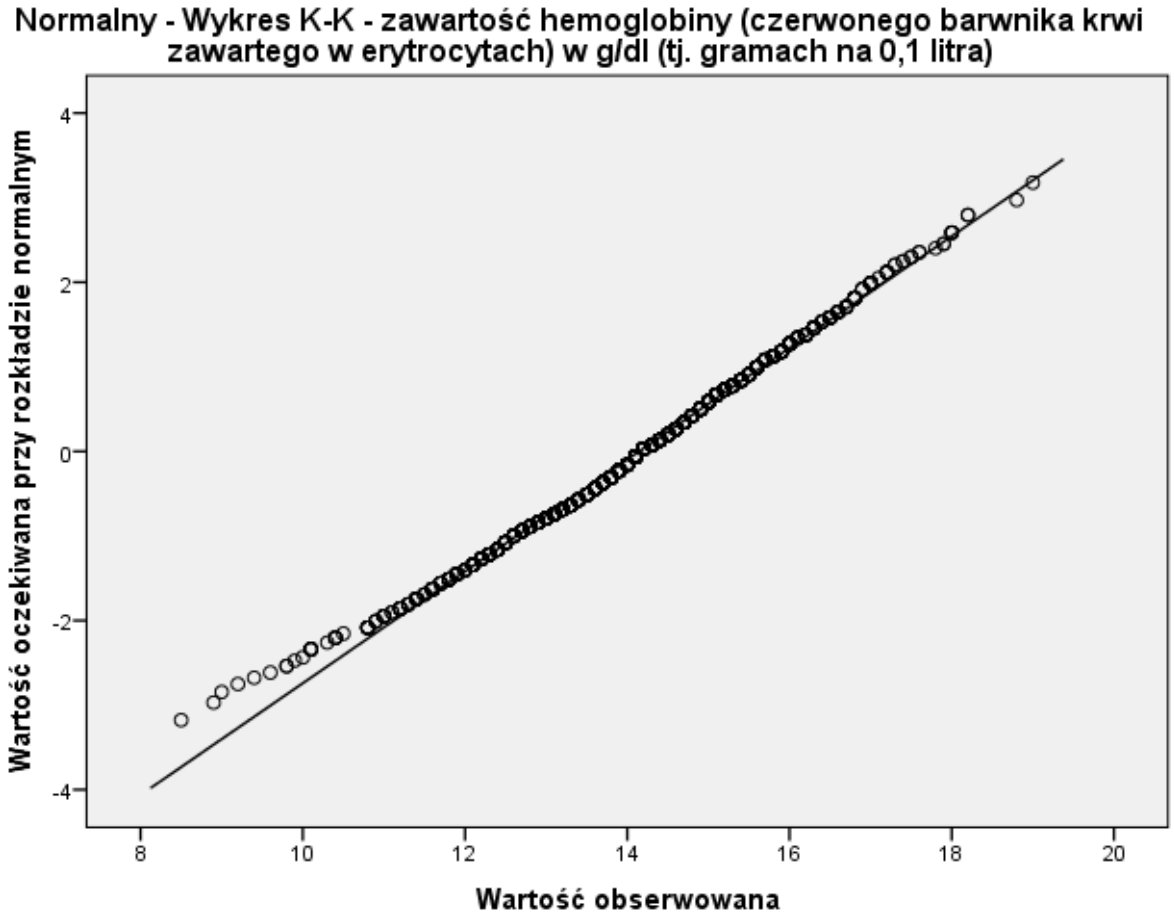
	Obserwacje					
	Uwzględnione		Wykluczone		Ogółem	
	N	Procent	N	Procent	N	Procent
zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra)	1345	100,0%	0	0,0%	1345	100,0%

Testy normalności rozkładu

	Kołmogorow-Smirnow[a]			Shapiro-Wilk		
	Statystyka	df	Istotność	Statystyka	df	Istotność
zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra)	,043	1345	,000	,995	1345	,000

a. Z poprawką istotności Lillieforsa

Z testu wynika, że badany rozkład nie jest normalny (istotność jest równa 0,000, co jest naturalnie mniejsze od zakładanej 0,05, więc odrzucamy hipotezę H0). Potwierdza to też wykres z następnej strony.


Normalny - Wykres K-K - zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra)
Wartość oczekiwana przy rozkładzie normalnym
4
2
0
-2
-4
8
10
12
14
16
18
20
Wartość obserwowana

Przeprowadzamy więc dodatkowy test, w którym porównamy badany rozkład z rozkładami normalnym (raz jeszcze – numer 1), jednostajnym (numer 2) oraz wykładniczym (numer 3).

Podsumowanie testu hipotezy

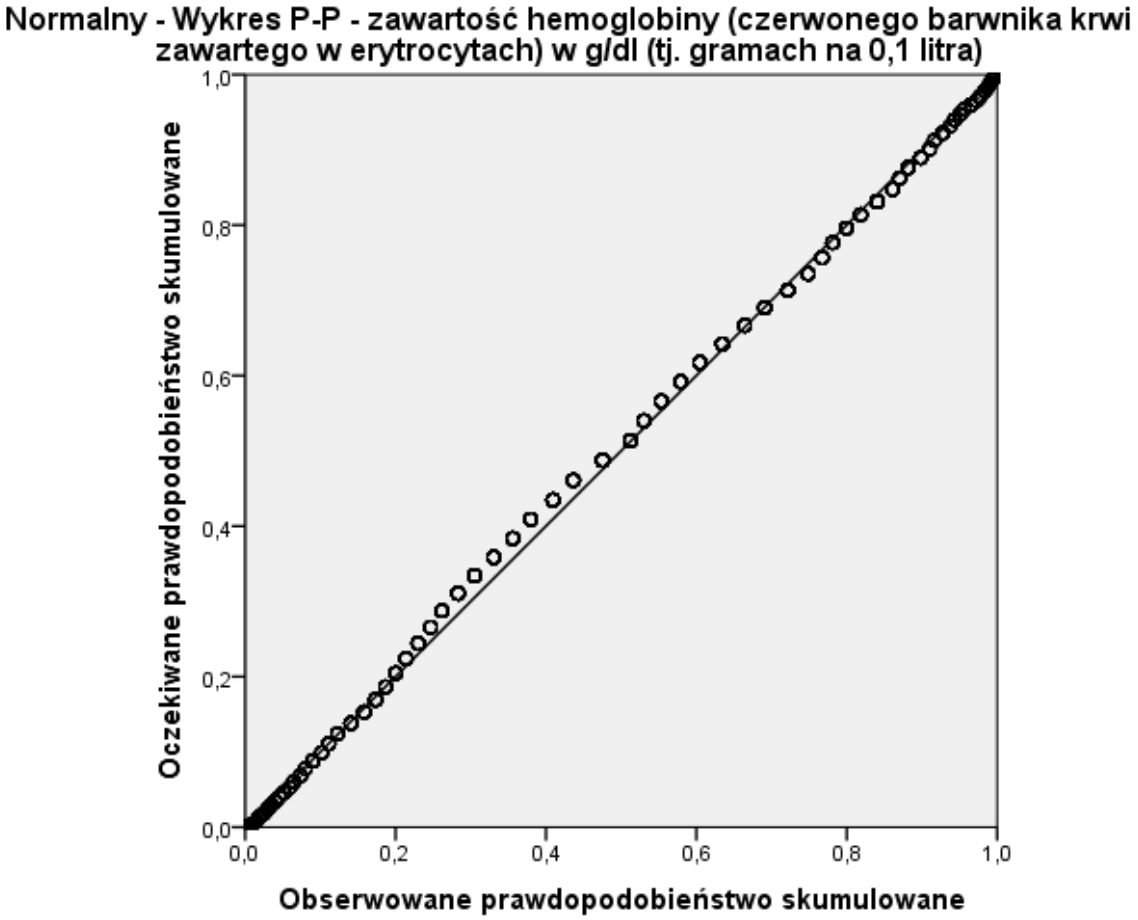
	Hipoteza zerowa	Test	Istotność	Decyzja
1	Rozkład zmiennej zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra) jest normalny ze średnią 14,148 i odchyleniem standardowym 1,51.	Test Kołmogorowa-Smirnowa dla jednej próby	,014	Odrzuć hipotezę zerową.
2	Rozkład zmiennej zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra) jest jednostajny z minimum 8,500 i maksimum 19,000.	Test Kołmogorowa-Smirnowa dla jednej próby	,000	Odrzuć hipotezę zerową.
3	Rozkład zmiennej zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra) jest wykładniczy ze średnią 14,148.	Test Kołmogorowa-Smirnowa dla jednej próby	,000	Odrzuć hipotezę zerową.

Przedstawiono asymptotyczne istotności. Poziom istotności wynosi ,05.

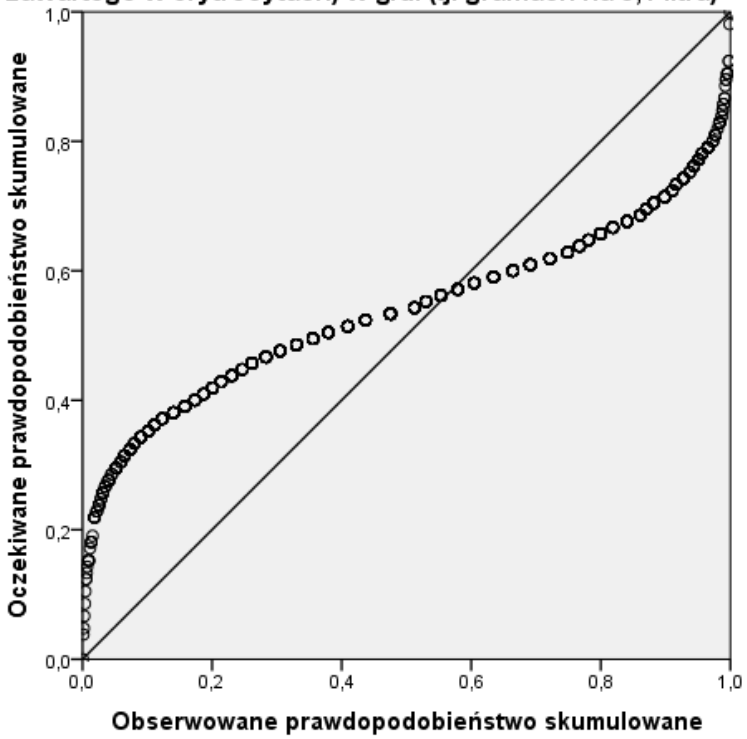
Wszystkie trzy badane rozkłady nie pasują do naszego rozkładu współczynnika hemoglobiny we krwi. Istotności są mniejsze od zakładanej 0,05, więc odrzucamy wszystkie hipotezy i stwierdzamy, że badany współczynnik ma inny niż wymienione rozkład.

Niemniej wracając do naszych wcześniejszych analiz, przypuszczać możemy, że rozkład co prawda nie jest normalny, ale jest do niego zbliżony. Aby przekonać się, czy takie podobieństwo zachodzi, wykonamy wykres P-P z opcją porównywania z rozkładem normalnym. Dodatkowo wykonamy wykresy P-P dla rozkładu jednostajnego i wykładniczego, by zobaczyć, że w tych przypadkach podobieństwo nie zachodzi.

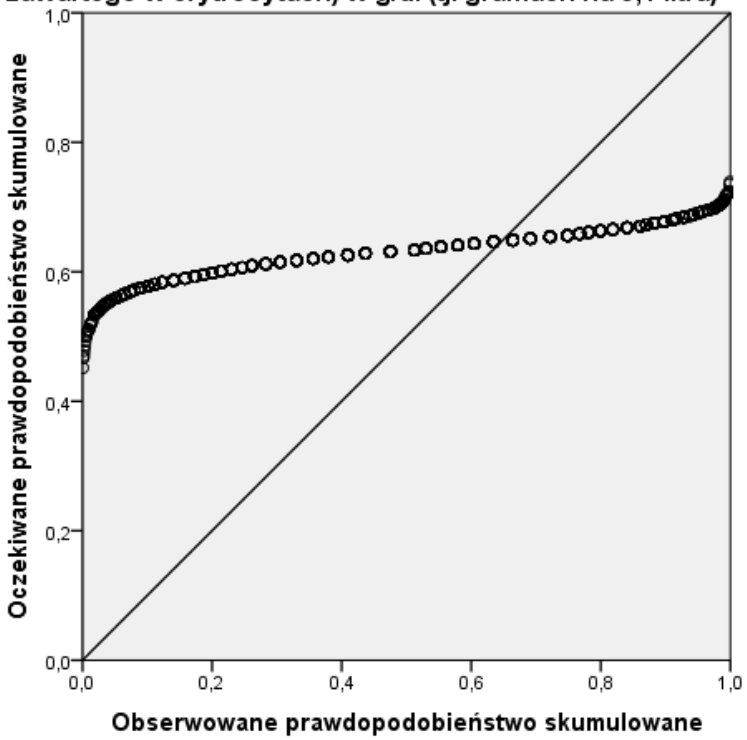
O zbliżoności do danego rozkładu świadczyć będzie podobieństwo do wykresu prostej y=x.



Jednostajny - Wykres P-P - zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra)



Wykładniczy - Wykres P-P - zawartość hemoglobiny (czerwonego barwnika krwi zawartego w erytrocytach) w g/dl (tj. gramach na 0,1 litra)



Na podstawie wykresów stwierdzamy, że rozkład współczynnika zawartości hemoglobiny we krwi jest najbardziej zbliżony do rozkładu normalnego (prawie pokrywa się z prostą y=x --- a im gęściej, tym większa pewność co do zgodności rozkładów).

Konkludując wszystkie rozważania, wnioskujemy, że rozkład współczynnika hemoglobiny we krwi jest najbardziej zbliżony do rozkładu normalnego, niemniej nie ma z nim całkowitej zgodności.