Sprawozdanie nr 2 od Lab 8 do Lab 14

Krystian Baran 145000

5 czerwca 2021

Spis treści

Część I

Laboratoria 07 - Estymacja punktowa

Celem tych laboratoriów było zapoznanie się z metodami estymacji punktowej i wyznaczanie estymatorów metodą MM oraz NNW. MM to metoda momentów polegająca na wyznaczenie parametrów poprzez znajomości wzorów momentów rozkładu zależnych od parametrów. MNW to metoda największej wiarygodności w której wyznacza się funkcję wiarygodności i poszukuje się w których punktach osiąga ona maksimum, zatem sprowadza się do układu liniowego z pochodnymi częściowymi tej funkcji przyrównane do 0.

1 Zadanie 1

Wyznaczyć estymator parametru p w rozkładzie Bernoulliego.

Rozkład Bernoulliego ma rozkład następujący:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, \ldots, n$$

Aby wyznaczyć estymator parametru p skorzystamy z metody momentów. Dla rozkładu Bernoulliego $\mathbb{E}X = np$ i $\mathbb{D}^2(X) = np(1-p)$. Oznaczymy momenty punktowe jako:

$$\mathbb{E}X = \overline{X}$$
$$\mathbb{D}^2(X) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = s^2$$

Wtedy:

$$\begin{cases} np = \overline{X} \\ np(1-p) = s^2 \end{cases}$$
$$\begin{cases} np = \overline{X} \\ \overline{X}(1-p) = s^2 \end{cases}$$
$$\begin{cases} np = \overline{X} \\ 1 - p = \frac{s^2}{\overline{X}} \end{cases}$$
$$\begin{cases} np = \overline{X} \\ p = 1 - \frac{s^2}{\overline{X}} \end{cases}$$
$$\begin{cases} n = \frac{\overline{X}^2}{\overline{X} - s^2} \\ p = 1 - \frac{s^2}{\overline{X}} \end{cases}$$

Zatem estymator parametru p jest $1 - \frac{s^2}{\overline{X}}$.

2 Zadanie 2

Wyznaczyć MM oraz MNW estymatory parametrów rozkładu normalnego.

Rozkład normalny definiowany jest w następujący sposób:

$$N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Parametr $\mu = \mathbb{E}X$ natomiast $\sigma = \mathbb{D}X$. Drugi moment zwykły tego rozkładu jest następujący:

$$\mathbb{E}(X^2) = \sigma^2 + \mu^2$$

Wyznaczymy estymatory parametrów metodą momentów. Niech $\mathbb{E}X = \overline{X}$ i $\mathbb{E}(X^2) = \frac{\sum_{i=1}^{n} X_i^2}{n}$, wtedy:

$$\begin{cases} \overline{X} = \mu \\ \frac{\sum_{i=1}^{n} X_i^2}{n} = \sigma^2 + \mu^2 \end{cases}$$
$$\begin{cases} \overline{X} = \mu \\ \frac{\sum_{i=1}^{n} X_i^2}{n} = \sigma^2 + \overline{X}^2 \end{cases}$$
$$\begin{cases} \overline{X} = \mu \\ \frac{\sum_{i=1}^{n} X_i^2}{n} = \sigma^2 + \mu^2 \end{cases}$$

Metodą największej wiarygodności natomiast potrzebujemy wyznaczyć funkcje wiarygodności.

$$\begin{aligned} L(x_1, x_2, \ldots, x_n | \mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}} \\ \ln(L) &= -n\ln(\sigma) - \frac{n}{2}\ln(2\pi) - \sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2} \end{aligned}$$

Obliczymy najpierw estymator parametru μ.

$$\begin{aligned}
\frac{\partial(\ln(L))}{\partial\mu} &= \sum_{i=1}^{n} \frac{2(x_i-\mu)}{2\sigma^2} \\
&= \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0 \\
\sum_{i=1}^{n}(x_i-\mu) &= 0 \\
\sum_{i=1}^{n} x_i - n\mu &= 0 \\
\mu &= \frac{\sum_{i=1}^{n} x_i}{n}
\end{aligned}$$

Dla parametru σ natomiast:

$$\begin{aligned}
\frac{\partial(\ln(L))}{\partial\sigma} &= -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{2(x_i-\mu)^2}{2\sigma^3} \\
&= -\frac{n}{\sigma} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = 0 \\
\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} &= \frac{n}{\sigma} \\
\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n} &= \sigma^2 \\
\sigma &= \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}
\end{aligned}$$

Widzimy zatem że parametry μ i σ są, odpowiednio, średnią i odchyleniem standardowym populacji szeregu punktowego.

3 Zadanie 3

Wyznaczyć MNW estymator parametru rozkładu Poissona.

Rozkład Poissona definiowany jest w następujący sposób:

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 0, 1, 2, \ldots$$

Jest to rozkład jedno parometrowy. Aby wyliczyć estymator parametru λ potrzebujemy obliczyć funkcję wiarygodności.

$$\begin{aligned} L(k_1, k_2, \ldots, k_n|\lambda) &= \prod_{i=1}^{n} P(X = k_i) = \prod_{i=1}^{n} \frac{\lambda^{k_i} e^{-\lambda}}{k_i!} \\ &= \lambda^{\sum_{i=1}^{n} k_i} e^{-n\lambda} \prod_{i=1}^{n} \frac{1}{k_i!} \\ \ln(L) &= \ln(\lambda) \sum_{i=1}^{n} k_i - n\lambda - \sum_{i=1}^{n} \ln k_i! \end{aligned}$$

Estymator parametru jest maksimum tej funkcji po zmiennej λ, zatem przyrównamy pierwszą pochodną do zera i znajdziemy szukany estymator.

$$\begin{aligned} \frac{\partial(\ln(L))}{\partial\lambda} &= \frac{\sum_{i=1}^{n} k_i}{\lambda} - n = 0 \\ n &= \frac{\sum_{i=1}^{n} k_i}{\lambda} \\ \lambda &= \frac{\sum_{i=1}^{n} k_i}{n} \end{aligned}$$

Sprawdzimy teraz drugą pochodną.

$$\frac{\partial^2(\ln(L))}{\partial\lambda^2} = -\frac{\sum_{i=1}^{n} k_i}{\lambda^2} < 0, \quad \forall\lambda$$

Zatem estymator parametru λ jest średnia arytmetyczna populacji.

4 Zadanie 4

Celem sprawdzenia dokładności wskazań pewnego przyrządu pomiarowego dokonano 10 pomiarów tej samej wielkości fizycznej X i otrzymano następujące wyniki:

9,01; 9,00; 9,02; 8,99; 8,98; 9,00; 9,00; 9,01; 8,99; 9,00.

Dokonać przekształcenia pomiarów według wzoru:

$$Y = 100(X - 9)$$

Dla wielkości X i Y oszacować ich wartości oczekiwane i wariancje.

Na początku sporządzimy tabele wartości X i Y korzystając z podanego wzoru.

Lp.	X	Y
1	9.01	1
2	9	0
3	9.02	2
4	8.99	-1
5	8.98	-2
6	9	0
7	9	0
8	9.01	1
9	8.99	-1
10	9	0

Oszacujemy wartość oczekiwaną jako średnia z podanych wartości, czyli:

$$\mathbb{E}X = \overline{X} = \frac{\sum x_i}{n} = \frac{90}{10} = 9$$

$$\mathbb{E}Y = \overline{Y} = \frac{\sum y_i}{n} = \frac{9}{10} = 0$$

Aby obliczyć odchylenie standardowe potrzebujemy sumę kwadratów obniżonych o średnią.

Lp.	$(x_i - \overline{X})^2$	$(y_i - \overline{Y})^2$
1	0.0001	1
2	0.0000	0
3	0.0004	4
4	0.0001	1
5	0.0004	4
6	0.0000	0
7	0.0000	0
8	0.0001	1
9	0.0001	1
10	0.0000	0
SUM	0.0012	12

Wtedy można łatwo obliczyć wartość odchylenia standardowego:

$$\sigma_x = \sqrt{\frac{\sum(x_i - \overline{X})^2}{n}} = \sqrt{\frac{0.0012}{10}} \approx 0.010954451$$

$$\sigma_y = \sqrt{\frac{\sum(y_i - \overline{Y})^2}{n}} = \sqrt{\frac{12}{10}} \approx 1.095445115$$

Wartość oczekiwana zmiennej X wynosi 9, gdzie X jest mierzona długość, więc możemy przyjąć że jest to długość mierzonego obiektu.
Wartość oczekiwana zmiennej Y, która wskazuje nam błąd procentowy względem wartości rzeczywistej 9, wynosi 0; zatem obiekt zmierzony został poprawnie.

Odchylenie standardowe zmiennej X wynosi w przybliżeniu 0.01, oznacza to że rzeczywistsza długość obiektu, z uwzględnieniem błędu pomiarowego wynosi 9.00 ± 0.01.
Odchylenie standardowe zmiennej Y wynosi w przybliżeniu 1, zatem rzeczywista wartość procentowego błędu jest $\pm 1\%$.

5 Zadanie 5

Wygenerować 50 elementową próbę prostą z populacji, w której cecha X ma rozkład o gęstości $f(x) = \frac{x}{8}\mathbf{1}_{(0;4)}(x)$

a) Sporządzić histogram.

b) Wyznaczyć wartość oczekiwaną i wariancję oraz ich oceny na podstawie wygenerowanej próby.

5.1 a)

Aby wygenerować próbę korzystając z twierdzenia obrócenia dystrybuanty musimy obliczyć dystrybuantę

$$\begin{aligned} F_X(x) &= \int_{\mathbb{R}} \frac{t}{8}\mathbf{1}_{(0;4)}(t)dt = \int_0^x \frac{t}{8}dt \\ &= \frac{t^2}{16}\Big|_0^x \\ &= \frac{x^2}{16} \end{aligned}$$

Zatem dystrybuanta jest następująca:

$$F_X(x) = \begin{cases} 0 & , \quad x \leqslant 0 \\ \frac{x^2}{16} & , \quad 0 < x < 4 \\ 1 & , \quad x \geqslant 4 \end{cases}$$

Odwrócimy dystrybuantę.

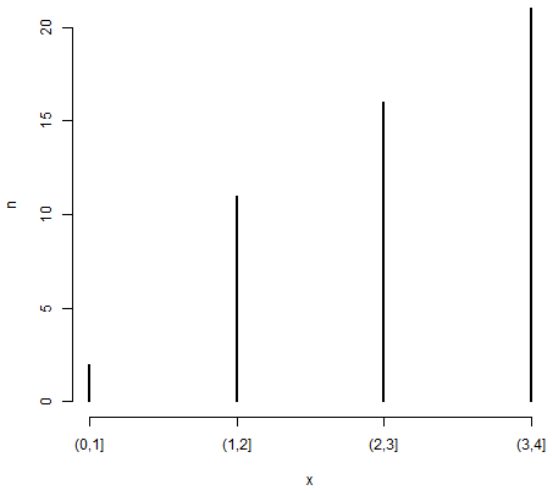
$$\begin{aligned} y &= \frac{x^2}{16} \\ 16y &= x^2 \\ x &= 4\sqrt{y} \end{aligned}$$

Poniżej przedstawiony został histogram przedziałowy wygenerowanej próby:

5.2 b)

Obliczymy teraz wartość oczekiwaną i wariancję z podanej funkcji i z wygenerowanej próby:

$$\begin{aligned} \mathbb{E}X &= \int_{\mathbb{R}} \frac{x^2}{8}\mathbb{I}_{(0,4)}(x)dx = \int_0^4 \frac{x^2}{8}dx \\ &= \frac{x^3}{24}\Big|_0^4 \\ &= \frac{64}{24} \approx 2.666666667 \end{aligned}$$



$$\begin{aligned}\mathbb{E}(X^2) &= \int_{\mathbb{R}} \frac{x^3}{8}\mathbb{I}_{(0,4)}(x)dx = \int_0^4 \frac{x^3}{8}dx \\ &= \frac{x^4}{32}\Big|_0^4 \\ &= \frac{64}{32} = 8\end{aligned}$$

$$\mathbb{D}^2(X) = \mathbb{E}(X^2) - \mathbb{E}X^2 = 8 - \frac{4096}{576} \approx 0.88888888889$$

Wartość oczekiwana z próby będzie średnią z próby, natomiast wariancje oznaczymy następującym wzorem:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

Obliczymy szukane wartości w R, gdzie *prob* jest tablicą zawierającą 50-elementową próbę.

$$\overline{X} \stackrel{R}{=} mean(prob) \approx 2.751196$$

$$s^2 \stackrel{R}{=} var(prob) \approx 0.8295317$$

Widzimy że że oba wartości są do siebie blisko, zatem możemy stwierdzić że obliczyliśmy poprawnie, gdzie $(X) = 2.7 \pm 0.1$ i $s^2 = 0.8 \pm 0.1$.

6 Zadanie 6

Korzystając z dostępnego oprogramowania wygenerować 100 elementową próbę według rozkładu

a) bin(20; 0,8),

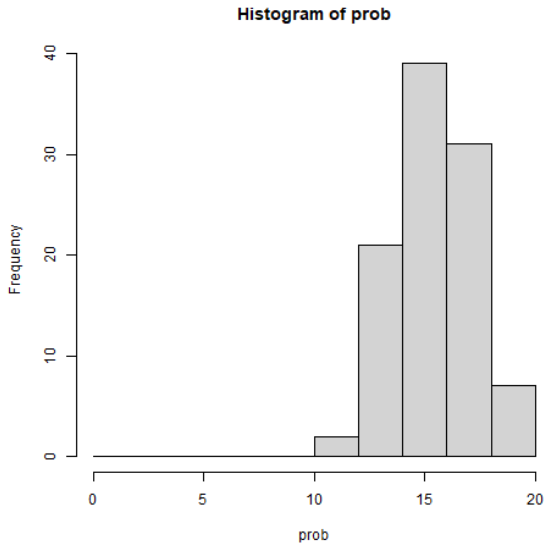
b) nbin(3;0,1),

c) Poisson(5).

Sporządzić histogram i dokonać ocenę punktową parametrów.

6.1 a)

Aby wygenerować losową próbę 100 elementową rozkładu Dwumianowego skorzystamy z funkcji R-owskiej *rbinom()*. Poniżej przestawiony został histogram dla losowej próby.



Wartość oczekiwana i wariancja teoretyczna wynoszą:

$$\mathbb{E}(X) = np = 20 \cdot 0.8 = 16$$

$$\mathbb{D}^2(X) = np(1-p) = 20 \cdot 0.8 \cdot 0.2 = 3.2$$

Natomiast, korzystając z wygenerowanej próby i z funkcji na średnią i wariancje w R otrzymujemy następujące wartości:

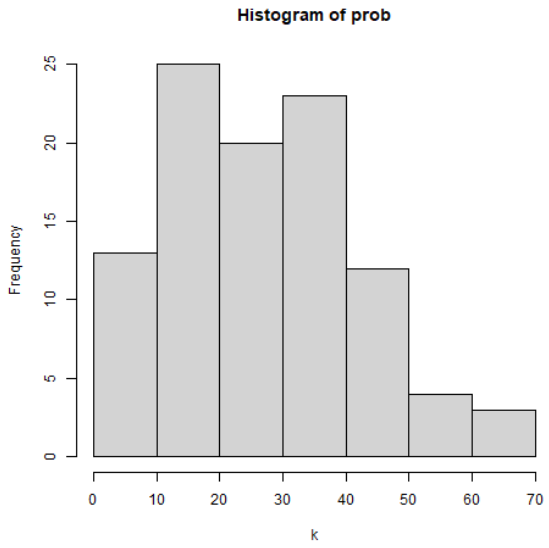
$$\overline{X} \overset{R}{=} mean(prob) \approx 15.87$$

$$s^2 \stackrel{R}{=} var(prob) \approx 3.104141$$

Widzimy że wartości tę są blisko wartości teoretycznej, zatem można stwierdzić że estymowana wartość oczekiwana wynosi 16 ± 0.2 a wariancja wynosi 3.1 ± 0.1.

6.2 b)

Podobnie jak w podpunkcie **a** wygenerujemy losową próbę 100-elementową w R za pomocą funkcji wbudowanej *rnbinom()*. Poniżej przedstawiono histogram.



Wartość oczekiwana i wariancja teoretyczna wynoszą:

$$\mathbb{E}(X) = \frac{(1-p)r}{p} = \frac{3 \cdot 0.9}{0.1} \approx 27$$

$$\mathbb{D}^2(X) = \frac{(1-p)r}{p^2} = \frac{3 \cdot 0.9}{0.01} \approx 270$$

Natomiast, korzystając z wygenerowanej próby i z funkcji na średnią i wariancje w R otrzymujemy następujące wartości:

$$\overline{X} \stackrel{R}{=} mean(prob) \approx 27.75$$

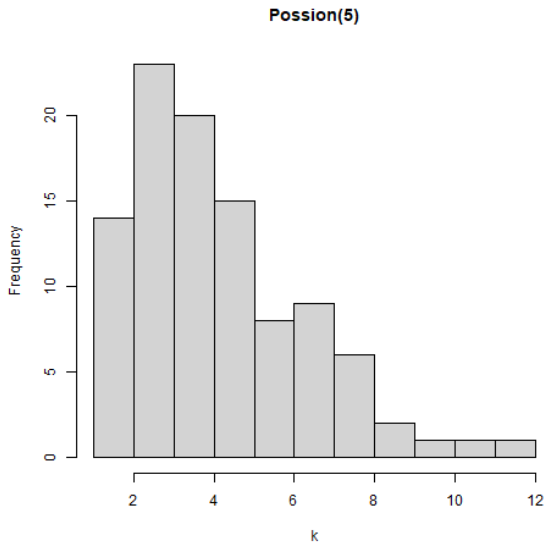
$$s^2 \stackrel{R}{=} var(prob) \approx 216.8965$$

Dla wartości oczekiwanej widzimy że wartości tę są blisko wartości teoretycznej, zatem można stwierdzić że estymowana wartość oczekiwana wynosi 27 ± 0.7.

Natomiast wariancja jest znacznie inna niż wartość teoretyczna; może wynikać to z tego że dla dużych wartości nie uzyskujemy znaczną dokładność, zatem na pewno pierwsza liczba została obliczona dokładnie a reszta już nie.

6.3 c)

Podobnie jak w podpunkcie **a** i **b** wygenerujemy losową próbę 100-elementową w R za pomocą funkcji wbudowanej *rpois()*. Poniżej przedstawiono histogram.



Wartość oczekiwana i wariancja teoretyczna wynoszą:

$$\mathbb{E}(X) = \lambda = 5$$

$$\mathbb{D}^2(X) = \lambda = 5$$

Natomiast, korzystając z wygenerowanej próby i z funkcji na średnią i wariancje w R otrzymujemy następujące wartości:

$$\overline{X} \overset{R}{=} mean(prob) \approx 4.59$$

$$s^2 \overset{R}{=} var(prob) \approx 4.870606$$

W tym przypadku widzimy że przy aproksymacji do liczby całkowitej uzyskamy dobrą wartość estymowanych parametrów. Zatem uznajemy że estymowana wartość oczekiwana wynosi 5 ± 1 a wariancja tak samo.

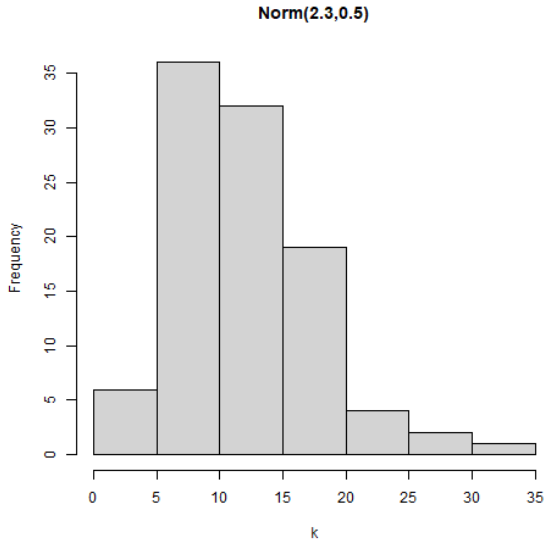
7 Zadanie 7

Wygenerować 100 elementową próbę według rozkładu logarytmiczno-normalnego z parametrami $\mu = 2.3$ i $\sigma = 0.5$.

a) Sporządzić histogram.

b) Dokonać estymacji parametrów, ocenić wartość oczekiwaną i wariancję oraz porównać te wartości z wartościami teoretycznymi.

Aby wygenerować losową próbę 100-elementową skorzystamy z dostępnej funkcji R-owskiej dla rozkładu logarytmiczno-normalnego *rlnorm()*. Poniżej przedstawiony został histogram z wygenerowanej próby:



Dla rozkładu logarytmiczno-normalnego wartość oczekiwana i wariancja są następujące:

$$\mathbb{E}(X) = e^{\mu} = e^2.3 \approx 9.974182455$$

$$\mathbb{D}^2(X) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} = (e^{0.25} - 1)e^{4.6+0.25} \approx 36.28151745$$

Korzystając z wygenerowanej próby i z funkcji na średnią i wariancje w R otrzymujemy następujące wartości:

$$\overline{X} \overset{R}{=} mean(prob) \approx 11.72756$$

$$s^2 \overset{R}{=} var(prob) \approx 30.83267$$

Wartości te różnią się nie wiele od wartości teoretycznych.

Część II
Laboratoria 8 - Estymacja przedziałowa

Celem tych laboratoriów było zapoznanie się z metodami estymacji parametrów dla danych przedziałowych. Zapoznaliśmy się także z wyznaczaniem przedziałów ufności dla wartości oczekiwanej oraz wariancji.

8 Zadanie 2

Korzystając z dostępnego oprogramowania wybrać rozkład i wygenerować małą oraz dużą próbę i na ich podstawie dokonać estymacji punktowej przedziałowej parametrów.

Niech prędkość wiatru w danej miejscowości ma rozkład Weibulla z danymi parametrami $k = 2$ i $\lambda = 8$ i niech mała próba losowa będzie się składać z 20 elementów, natomiast duża próba będzie zawierała 100 elementów. n-elementowa próba rozkładu Weibulla została wykonana z pomocą funkcji R-owskiej *rweibull()*.

8.1 20 - elementów

Poniżej przedstawiono tabele przedziałową próby.

Lp	Przedz	Licz
1	(0;2]	1
2	(2;4]	4
3	(4;6]	3
4	(6;8]	0
5	(8;10]	5
6	(10;12]	6
7	(12;14]	1
8	(14;16]	0
9	(16;18]	0
10	(18;20]	0

Aby obliczyć średnią korzystaliśmy ze wzoru poniżej, gdzie x_i jest środkiem przedziału. n_i natomiast jest liczebnością przedziału.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i \cdot n_i}{n} = \frac{152}{20} \approx 7.6$$

Natomiast dla wariancji skorzystaliśmy ze wzoru poniżej.

$$S_n^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X})^2 \cdot n_i}{n-1} = \frac{256.8}{19} \approx 13.51578947$$

8.2 100 - elementów

Poniżej przedstawiono tabele przedziałową próby.

Lp	Przedz	Licz
1	(0;2]	9
2	(2;4]	10
3	(4;6]	16
4	(6;8]	23
5	(8;10]	14
6	(10;12]	13
7	(12;14]	7
8	(14;16]	6
9	(16;18]	0
10	(18;20]	2

Jak w podpunkcie **a** obliczono, odpowiednio, średnią i wariancję.

$$\overline{X} = \frac{768}{100} \approx 7.68$$

$$S_n^2 = \frac{1689.76}{99} \approx 17.06828283$$

Widzimy zatem że przy zwiększeniu próby uzyskaliśmy tylko lekką poprawę wartości średniej, natomiast wariancje znacznie różnią się od siebie.

9 Zadanie 3

Rozkład wyników pomiarów głębokości morza w pewnym rejonie jest normalny. Dokonano 5 niezależnych pomiarów głębokości morza w tym rejonie i otrzymano następujące wyniki (w [m]): 871, 862, 870, 876, 866. Na poziomie ufności 0,90 wyznaczyć CI dla wartości oczekiwanej oraz dla wariancji głębokości morza w badanym rejonie.

Obliczymy najpierw średnią i wariancję z podanej próby:

$$\overline{X}_5 = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{4345}{5} = 869$$

$$S_5^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{112}{4} = 28$$

$$S_5 = \sqrt{S_5^2} \approx 5.291502622$$

Wyznaczymy teraz α wiedząc że poziom ufności jest 0.90.

$$1 - \alpha = 0.90 \Rightarrow \alpha = 0.1$$

Ponieważ pomiary głębokości morza mają rozkład normalny gdzie nie znane są parametry m i σ skorzystamy najpierw z przedziału ufności dla σ^2 podany poniżej:

$$R = \Big(\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}; \frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2};n-1}}\Big)$$

Obliczymy teraz wartości kwantyli rozkładu χ^2:

$$\chi^2_{1-0.05;4} \overset{R}{=} qchisq(0.95, 4) \approx 9.487729$$

$$\chi^2_{0.05;4} \overset{R}{=} qchisq(0.05, 4) \approx 0.710723$$

Wtedy szukany przedziały ufności dla σ^2:

$$R = (11.80472166; 157.5860075)$$

Teraz możemy obliczyć przedziały ufności dla wartości oczekiwanej z poniższego wzoru:

$$\overline{X}_n \mp t_{1-\frac{\alpha}{2};n-1}\frac{S_n}{\sqrt{n}}$$

$t_{1-\frac{\alpha}{2};n-1}$ to kwantyl rozkładu Studenta z $n-1$ stopniami swobody.

$$t_{0.95;4} \overset{R}{=} qt(0.95, 4) \approx 2.131847$$

Wtedy szukany przedział to:

$$R = (863.9551292; 874.0448708)$$

10 Zadanie 5

Linia lotnicza chce oszacować frakcję Polaków, którzy będą korzystać z nowo otwartego połączenia między Poznaniem a Londynem. Wybrano losową próbę 347 pasażerów korzystających z tego połączenia, z których 201 okazało się Polakami.

a) Wyznaczyć 90% przedział ufności dla frakcji Polaków wśród pasażerów korzystających z nowo otwartego połączenia.

b) Wygenerować 347 elementową próbę według rozkładu $B(0, 58)$ identyfikującą polskich pasażerów i na tej podstawie wyznaczyć 90% przedział ufności.

10.1 a)

Wyznaczmy α widząc ze szukamy przedział ufności 90%.

$$1 - \alpha = 0.9 \Rightarrow \alpha = 0.1$$

Niech każdy pasażer ma narodowość niezależną od innych pasażerów, wtedy każdy pasażer X_i będzie miał rozkład Bernouilliego z nieznanym parametrem p opisując czy jest polakiem czy nie. Zatem $X = \sum X_i$ będzie także rozkładem Bernoulliego i będzie opisywało liczbę polaków z pośród pasażerów.
Aby wyznaczyć przedział ufności dla parametru p sprawdzimy poniższy warunek:

$$1 < \overline{p}_n \mp 3\sqrt{\frac{\overline{p}_n(1-\overline{p}_n)}{n}} < 1$$

$$\overline{p}_n \mp 3\sqrt{\frac{\overline{p}_n(1-\overline{p}_n}{n}}) = \frac{201}{347} \mp 3\sqrt{\frac{201/347 \cdot 146/347}{347}} \approx 0.579251 \mp 0.079506$$

Spełniony jest warunek dla obu wartości, zatem możemy wyznaczyć szukany przedział ufności ze wzoru.

$$\overline{P}_n \mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}_n(1-\overline{P}_n)}{n}}$$

$$z_{1-\frac{\alpha}{2}} \overset{R}{=} qnorm(0.95, 0, 1) \approx 1.644854$$

$$\overline{P}_n \mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}_n(1-\overline{P}_n}{n}} = 0.579251 \mp 0.043592$$

$$R = \Big(0.535659; 0.622843\Big)$$

Podany powyżej jest szukany przedział z 90% ufności.

10.2 b)

Aby wygenerować losową próbę wykorzystamy funkcję R-owską dla rozkładu Bernoulliego *rbinom()*. Następne kroki jak w poprzednim przypadku. Oznaczmy próbę jako:

$$\text{prob} \overset{R}{=} rbinom(1, 347, 0.58) = 189$$

Dla takiej liczby sprawdzimy warunek:

$$\overline{p}_n \mp 3\sqrt{\frac{\overline{p}_n(1-\overline{p}_n}{n}} = \frac{189}{347} \mp 3\sqrt{\frac{189/347 \cdot 158/347}{347}} \approx 0.544669 \mp 0.026734$$

Warunek jest spełniony zatem obliczamy jak poprzednio:

$$\overline{P}_n \mp z_{1-\frac{\alpha}{2}}\sqrt{\frac{\overline{P}_n(1-\overline{P}_n)}{n}} = 0.544669 \mp 0.043974$$

Wtedy przedział ufności wynosi:

$$R = \Big(0.500695; 0.588643\Big)$$

11 Zadanie 6

Frekwencja widzów na seansie filmowym w jednym z kin ma rozkład $N(\mu = ?; \sigma = 30)$. Na podstawie rejestru liczby widzów na 25 losowo wybranych seansach filmowych oszacowano przedział liczbowy (184; 216) dla nieznanej przeciętnej frekwencji na wszystkich seansach.

a) Obliczyć średnią liczbę widzów w badanej próbie.

b) Jaki poziom ufności przyjęto przy estymacji?

11.1 a)

Dla rozkładu normalnego ze znanym parametrem σ przedział ufności dla wartości oczekiwanej jest następujący:

$$\overline{X}_n \mp z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Gdzie n jest liczebność próby i α jest parametrem ufności.
Można wtedy wyznaczyć szukany parametr $\overline{X}_n$ i α.

$$\begin{cases} 184 = \overline{X}_{25} - z_{1-\frac{\alpha}{2}} \frac{30}{\sqrt{25}} \\ 216 = \overline{X}_{25} + z_{1-\frac{\alpha}{2}} \frac{30}{\sqrt{25}} \end{cases}$$

$$\begin{cases} \overline{X}_{25} = 184 + z_{1-\frac{\alpha}{2}} 6 \\ \overline{X}_{25} = 216 - z_{1-\frac{\alpha}{2}} 6 \end{cases}$$

$$\begin{cases} \overline{X}_{25} = 184 + z_{1-\frac{\alpha}{2}} 6 \\ 184 + z_{1-\frac{\alpha}{2}} 6 = 216 - z_{1-\frac{\alpha}{2}} 6 \end{cases}$$

$$\begin{cases} \overline{X}_{25} = 184 + z_{1-\frac{\alpha}{2}} 6 \\ 12 z_{1-\frac{\alpha}{2}} = 32 \end{cases}$$

$$\begin{cases} \overline{X}_{25} = 184 + z_{1-\frac{\alpha}{2}} 6 \\ 1 - \frac{\alpha}{2} = \Phi(2.66666667) \end{cases}$$

$$\begin{cases} \overline{X}_{25} = 184 + z_{1-\frac{\alpha}{2}} 6 \\ \alpha = 2 - 2\Phi(2.66666667) \overset{R}{=} 2 - 2 * pnorm(2.67, 0, 1) \approx 0.007585125 \end{cases}$$

$$\begin{cases} \overline{X}_{25} \overset{R}{=} 184 + qnorm(1 - 0.007585125/2, 0, 1) * 6 \approx 200.02 \\ \alpha \approx 0.007585125 \end{cases}$$

Zatem szukana średnia wynosi 200.

11.2 b)

Jak wyliczono w podpunkcie **a** α wynosi około 0.008, zatem procent ufności wynosi 99.2%.

12 Zadanie 15 - Studium przypadku

Z partii kondensatorów wybrano losowo 12 kondensatorów i zmierzono ich pojemności, otrzymując wyniki (w pF): 4,45, 4,40, 4,42, 4,38, 4,44, 4,36, 4,40, 4,39, 4,45, 4,35, 4,40, 4,35.

a) Znaleźć ocenę wartości oczekiwanej $\overline{x}_{12}$ i wariancji s^2_{12} pojemności kondensatora pochodzącego z danej partii.

b) Wygenerować 100 elementową próbę według rozkładu $N(\overline{x}_{12}, s_{12})$.

c) Znaleźć ocenę wskaźnika kondensatorów, które nie spełniają wymagań technicznych, przyjmując, że kondensator nie spełnia tych wymagań, gdy jego pojemność jest mniejsza od 4,39 pF.

d) Znaleźć ocenę wariancji pojemności kondensatorów.

e) Wyznaczyć 90-procentową ocenę przedziału ufności dla wartości oczekiwanej pojemności kondensatora pochodzącego z danej partii.

f) Wyznaczyć 90-procentową realizację przedziału ufności dla wskaźnika kondensatorów, które nie spełniają wymagań technicznych w badanej partii.

12.1 a)

Obliczymy $\overline{x}_{12}$ i s^2_{12} ze wzorów odpowiednio:

$$\overline{x}_{12} = \frac{1}{12}\sum_{i=1}^{12} x_i = \frac{52.79}{12} \approx 4.399166667$$

$$s^2_{12} = \frac{1}{11}\sum_{i=1}^{12}(x_i - \overline{x}_{12})^2 = \frac{0.014091667}{11} \approx 0.001281061$$

Zatem $\overline{x}_{12} = 4.4$ a $s^2_{12} = 0.0013$

12.2 b)

Poniżej przedstawiono wygenerowaną próbę za pomocą funkcji R-owskiej *rnorm()*.

lp	przedz	licz
1	(4.35;4.37]	11
2	(4.37;4.39]	25
3	(4.39;4.41]	21
4	(4.41;4.43]	16
5	(4.43;4.45]	11

12.3 c)

Liczba kondensatorów która nie spełnia wymagania ma rozkład dwumianowy z nieznanym parametrem p. Z 12-elementowej próby możemy obliczyć ile kondensatorów nie spełnia warunki i wyznaczyć wskaźnik struktury:

$$\overline{P}_{12} = \frac{1}{12} K_1 2 = \frac{4}{12} \approx 0.333333$$

Zatem szukany p z próby wynosi około 0.33

12.4 d)

Nie rozwiązane.

12.5 e)

Wyznaczymy parametr α jako:

$$1 - \alpha = 0.9 \Rightarrow \alpha = 0.1$$

Dla rozkładu normalnego przedział ufności wyznacza się w następujący sposób odpowiednio dla σ^2 i m:

$$\Big(\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}; \frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2};n-1}}\Big)$$

$$\overline{X}_n \mp t_{1-\frac{\alpha}{2};n-1} \frac{S_n}{\sqrt{n}}$$

Gdzie $t_{1-\frac{\alpha}{2};n-1}$ to kwantyl rozkładu Studenta z $n-1$ stopniami swobody a $\chi^2_{1-\frac{\alpha}{2};n-1}$ to podobnie kwantyl rozkładu chi kwadrat.

Zatem dla σ^2.

$$\chi^2_{1-\frac{\alpha}{2};n-1} \overset{R}{=} qchisq(0.95, 11) \approx 19.67514$$

$$\chi^2_{\frac{\alpha}{2};n-1} \overset{R}{=} qchisq(0.05, 11) \approx 4.574813$$

$$\Big(\frac{11 \cdot 0.0013}{\chi^2_{0.95;11}}; \frac{11 \cdot 0.0013}{\chi^2_{0.95;11}}\Big) = (0.0007268; 0.003126)$$

Natomiast dla m.

$$t_{1-\frac{\alpha}{2};n-1} \overset{R}{=} qt(0.95, 11) \approx 1.795885$$

$$\Big(4.4 - 1.795885\sqrt{\frac{0.0013}{12}}; 4.4 - 1.795885\sqrt{\frac{0.0013}{12}}\Big) = (4.3813; 4.4187)$$

12.6 f)

Skorzystamy z α obliczone w poprzednim podpunkcie i z następującego wzory na przedział ufności:

$$\overline{P}_n \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{P}_n(1-\overline{P}_n)}{n}}$$

Sprawdzimy najpierw warunek:

$$1 < \overline{p}_n \mp 3\sqrt{\frac{\overline{p}_n(1-\overline{p}_n)}{n}} < 1$$

$$\overline{p}_n \mp 3\sqrt{\frac{\overline{p}_n(1-\overline{p}_n)}{n}} = \frac{4}{12} \mp 3\sqrt{\frac{4/12 \cdot 8/12}{12}} \approx 0.333333 \mp 0.408248$$

Warunek jest spełniony tylko dla wartości dodatniej, zatem przedział ufności jest prawostronny i wynosi:

$$z_{1-\frac{\alpha}{2}} \overset{R}{=} qnorm(0.95, 0, 1) \approx 1.644854$$

$$\overline{P}_n + z_{1-\frac{\alpha}{2}} \sqrt{\frac{\overline{P}_n(1-\overline{P}_n)}{n}} = \frac{4}{12} + 1.644854\sqrt{\frac{4/12 \cdot 8/12}{12}}$$

$$(0; 0.55717)$$

Podany powyżej jest szukany przedział ufności.

13 Zadanie 17

A random sample of 64 observation from a population produced the following summary statistics: $\sum x_i = 500$, $\sum (x_i - \overline{x})^2 = 3,566$.

a) Find 95% confidence interval for m.

b) Interpret the confidence interval you found in part (a).

14 a)

First we calculate α from the confidence level:

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$$

Because the standard deviation is not given we will use the formula given below to calculate our confidence interval borders:

$$\overline{X}_n \mp t_{1-\frac{\alpha}{2};n-1} \frac{S_n}{\sqrt{n}}$$

For $\overline{X}_n$ we just divide the given sum by the number of observations:

$$\overline{X}_6 4 = \frac{500}{64} \approx 7.8125$$

For S_n we need to take the square root of the variance which is given by the formula below:

$$S_n^2 = \frac{1}{n-1} \sum (x_i - \overline{x})^2 = \frac{3.566}{63} \approx 0.56444444$$

$$S_n = \sqrt{0.56444444} \approx 0.751295$$

Because $t_{1-\frac{\alpha}{2};n-1}$ is the quantile function of Student-t distribution with $n-1$ degrees of freedom we can calculate it using the R programming language:

$$t_{1-\frac{\alpha}{2};n-1} \overset{R}{=} qt(0.975, 63) \approx 1.998341$$

Finally we can plug in the calculated values to find the confidence interval:

$$\overline{X}_n \mp t_{1-\frac{\alpha}{2};n-1} \frac{S_n}{\sqrt{n}} = 7.8125 \mp 1.998341 \cdot \frac{0.751295}{8} \approx 7.8125 \mp 0.187668$$

$$(7.624832; 8.000168)$$

The above interval is our confidence interval.

14.1 b)

The purpose of the confidence interval is to find an interval where almost for sure we can find our searched parameter. In this case we can say that our searched m parameter is for sure between 7.7 and 8. We can then take a value from this interval and say that our population is normally distributed with that parameter m.

15 Zadanie 19

Jak liczna powinna być próba, aby na jej podstawie można było z prawd. 0,99 oszacować średni wzrost noworodków przy maksymalnym błędzie szacunku 1cm? Zakładamy, że rozkład wzrostu noworodków jest rozkładem normalnym z odchyleniem standardowym 2,5cm.

Aby obliczyć minimalną liczebność próby skorzystamy z następującego wzoru:

$$n = \left\lceil \frac{z^2_{1-\frac{\alpha}{2}}\sigma^2}{d^2} \right\rceil$$

Gdzie $z_{1-\frac{\alpha}{2}}$ to kwantyl rozkładu $N(0,1)$, $d = 1[cm]$, $\sigma = 2.5[cm]$ i $1 - \alpha = 0.99 \Rightarrow \alpha = 0.01$.

Obliczymy najpierw kwantyl.

$$z^2_{1-\frac{\alpha}{2}} \overset{R}{=} qnorm(0.995, 0, 1)^2 \approx 6.634897$$

Wtedy możemy wyznaczyć n.

$$n = \left\lceil \frac{6.634897 \cdot 6.25}{1} \right\rceil = \lceil 41.468106 \rceil = 42$$

Zatem minimalna liczebność próby aby z prawdopodobieństwem 0.99 oszacować wzrost noworodków przy maksymalnym błędzie szacunku 1 [cm] jest 42.

Część III
Laboratoria 9

Celem tych laboratoriów było zapoznanie się z metodami przeprowadzania testów parametrycznych dla jednej populacji. Przeprowadzane testy były aby ocenić wartości oczekiwane, wariancje i wskaźniki struktury. Zapoznaliśmy się także z metodami wyznaczania hipotez, zerowych i alternatywnych, i na podstawie wyników przeprowadzonych testów jedna z hipotez była odrzucana. Nauczyliśmy się także wyznaczać obszary krytyczne dla tych testów oraz obliczenie *p-value* i jak tę wartość interpretować.

16 Zadanie 2 - Funkcje testów w R

W języku programowania R istnieją różne funkcji do wykonywania testów, natomiast zatrzymamy się na najważniejszych.

16.1 t.test()

Funkcja do wykonania testu t-Studenta dla którego zakłada się że dane pochodzą z rozkładu normalnego z parametrami nieznanymi. Parametry tej funkcj są następujące:

- **x** - wektor wartości próby
- **y** - wektor wartości próby do porównania z próbą x. Opcjonalny
- **alternative** - specyfikacja hipotezy alternatywnej przyjmujące wartości: "two.sided", "greater", "less"
- **mu** - wartość prawdziwej wartości oczekiwanej lub różnica między wartościami oczekiwanymi dla dwóch prób
- **paired** - logiczna wartość dla "paired" testu
- **var.equal** - logiczna mówiąca czy traktować wariancje jako takie same lub nie. Gdy FALSE korzysta się z aproksymacja Welcha.
- **conf.equal** - confidence level
- **formula** - "lhs" dla numerycznej zmiennej oddającej wartości, "rhs" dla two pionową korespondencji grup.
- **data** - Opcjonalna macierz z wartościami użytych do polu *formula*
- **subset** - Opcjonalny wektor z podzbiorem obserwacji do wykorzystania w teście

- **na.action** - funkcja definiująca co się dzieje gdy napotkane zostają wartości *NA*

Oddawane przez tą funkcje wartości są następujące:

- **statistic** - wartość statystyki t
- **parameter** - stopnie swobody statystyki t
- **p.value** - *p value* wykonanego testu
- **conf.int** - przedział ufności dla wartości oczekiwanej
- **estimate** - estymowana wartość oczekiwana lub różnica wartości oczekiwanych dla testu dwóch prób
- **null.value** - podana wartość *mu*
- **stderr** - standardowy błąd wartości oczekiwanej, używany jako mianownik statystyki t
- **alternative** - opis hipotezy alternatywnej
- **method** - typ wykonanego testu t
- **data.name** - imię podanej macierz pod *data*

16.2 wilcox.test()

Test Wilcoxona wykorzystany jest do badania wartości oczekiwanej jak w teście t-Studenta ale nie zakłada się że próba ma rozkład normalny. Funkcja ta przyjmuje następujące parametry:

- **x** - wektor liczb na podstawie której będzie test prowadzony
- **y** - wektor liczb w przypadku testu dwóch prób
- **mu** - wartość oczekiwana hipotezy zerowej
- **paired** - logiczna dla testu "paired"
- **exact** - logiczna mówiąca czy dokładna wartość *p value* powinna być liczona
- **correct** - logiczna mówiąca czy *p value* powinna być poprawiona ze względu na ciągłość
- **conf.int** - logiczna mówiąca czy powinien zostać liczony przedział ufności
- **conf.level** - poziom ufności testu
- **formula** - "lhs" dla numerycznej zmiennej oddającej wartości, "rhs" dla two pionową korespondencji grup.

- **data** - Opcjonalna macierz z wartościami użytych do polu *formula*
- **subset** - Opcjonalny wektor z podzbiorem obserwacji do wykorzystania w teście
- **na.action** - funkcja definiująca co się dzieje gdy napotkane zostają wartości *NA*

Funkcja ta oddaje podobne wartości jak test t-Studenta:

- **statistic** - wartość statystyki z imieniem opisującym
- **parameter** - parametry dokładnego rozkładu statystyki testowej
- **p.value** - *p value* wykonanego testu
- **null.value** - podana wartość *mu*
- **alternative** - opis hipotezy alternatywnej
- **method** - typ wykonanego testu
- **data.name** - imię podanej macierz pod *data*
- **conf.int** - przedział ufności dla wartości oczekiwanej
- **estimate** - estymowana wartość oczekiwana lub różnica wartości oczekiwanych dla testu dwóch prób

16.3 var.test()

Test ten jest testem F Snedecora dla porównania wariancji pomiędzy dwoma populacjami. Funkcja ta przyjmuje następujące wartości:

- **x, y** - wektory liczb dla których przeprowadzony jest test
- **ratio** - hipotetyczny "ratio" pomiędzy badanymi wariancjami
- **alternative** - hipoteza alternatywna, przyjmuje wartości: "two.sided", "greater", "less"
- **conf.level** - poziom ufności testu
- **formula** - "lhs" dla numerycznej zmiennej oddającej wartości, "rhs" dla two pionową korespondencji grup.
- **data** - Opcjonalna macierz z wartościami użytych do polu *formula*
- **subset** - Opcjonalny wektor z podzbiorem obserwacji do wykorzystania w teście
- **na.action** - funkcja definiująca co się dzieje gdy napotkane zostają wartości *NA*

Funkcja ta zwraca listę typu "htest" zawierająca następujące komponenty:

- **statistic** - wartość statystyki F-test
- **parameter** - stopnie swobody rozkładu F dla testu
- **p.value** - *p value* wykonanego testu
- **conf.int** - przedział ufności dla wartości oczekiwanej
- **estimate** - estymowana wartość oczekiwana lub różnica wartości oczekiwanych dla testu dwóch prób
- **null.value** - "ratio" wariancji populacji podane
- **alternative** - opis hipotezy alternatywnej
- **method** - typ wykonanego testu
- **data.name** - imię podanej macierz pod *data*

16.4 ks.test()

Funckja do wykonania testu na dwóch próbach w celu sprawdzenia czy mają ten sam rozkład. Funkcja ta przyjmuje następujące wartosci:

- **x** - wektor wartości
- **y** - wektor wartości lub łańcuch opisujący dystrybuantę rozkładu
- **alternative** - hipoteza alternatywna, przyjmuje wartości: "two.sided", "greater", "less"
- **exact** - logiczna mówiąca czy dokładna wartość *p value* powinna być liczona
- **tol** - górny koniec przedziału dla błędu zaokrąglania
- **simulate.p.value** - logiczna mówiąca czy symulować *p value* według Monte Carlo
- **B** - liczba replikatów dla testu Monte Carlo

Funkcja ta zwraca listę typu "htest" zawierająca następujące komponenty:

- **statistic** - wartość statystyki testowej
- **p.value** - *p value* wykonanego testu
- **alternative** - opis hipotezy alternatywnej
- **method** - typ wykonanego testu
- **data.name** - imię podanej macierz pod *data*

17 Zadanie 3

Wytwórnia cukierków paczkuje w torebki po około 200 sztuk mieszankę złożoną z dwóch rodzajów cukierków, przy czym paczkowane są dwa typy mieszanek. Mieszanka typu A zawiera 40% cukierków pierwszego rodzaju i 60% drugiego rodzaju, natomiast mieszanka typu B zawiera jednakowe liczby cukierków obydwu rodzajów.
Do weryfikacji hipotezy $H_0 : p = 40\%$, że mieszanka jest typu A, wobec hipotezy alternatywnej $H_1 : p = 50\%$, zaproponowano następującą procedurę:
jeśli wśród 5 cukierków wylosowanych z torebki znajdą się więcej niż 3 cukierki pierwszego rodzaju, to odrzuca się hipotezę zerową na rzecz hipotezy alternatywnej. W przeciwnym przypadku przyjmuje się hipotezę zerową.
Przy tak określonej procedurze testowej, znaleźć prawdopodobieństwa błędów obydwu rodzajów oraz moc testu.

Ponieważ znamy ilość cukierków w torebce i procentowość każdego rodzaju cukierków w obu typu paczek, możemy zastosować rozkład hypergeometryczny z parametrami $n = 200 * 40/100 = 80, m = 200 * 60/100 = 120$ dla typu A i $n = 100, m = 100$ dla typu B. Nazwijmy je odpowiednio X i Y.
Aby sprawdzić błąd pierwszego rodzaju oznaczony jako α potrzebujemy przedział krytyczny dla typu A. Skoro podany został typ testu, jeżeli wylosujemy więcej niż 3 cukierki pierwszego rodzaju jest to typ B, zatem przedział krytyczny dla H_0 jest:

$$R = \{4, 5\}$$

Wtedy, zakładając że hipoteza zerowa jest prawdziwa, czyli korzystamy z rozkładu X, można obliczyć α jako:

$$\begin{aligned}\alpha &= P(U_n \in R|H_0) = P(X > 3) = 1 - P(X \leqslant 3) \\ &\overset{R}{=} 1 - phyper(3, 80, 120, 5) \approx 0.08432931\end{aligned}$$

Aby obliczyć błąd drugiego rodzaju β zakładamy ze hipoteza alternatywna jest prawdziwa, zatem korzystamy z rozkładu Y. Wtedy można obliczyć szukane β z następującego wzoru:

$$\begin{aligned}\beta &= 1 - P(U_n \in R|H_1) = 1 - P(Y > 3) = P(Y \leqslant 3) \\ &\overset{R}{=} phyper(3, 100, 100, 5) \approx 0.8156646\end{aligned}$$

Moc testu oznacza się jako $1 - \beta = 1 - 0.8156646 = 0.1843354$. Zatem moc testu wynosi około 0.184.

18 Zadanie 4 - Studium przypadku

Pascal jest językiem programowania wysokiego poziomu, stosowanym często do oprogramowywania mikrokomputerów. W celu zbadania wskaźnika p zmiennych pascalowych typu tablicowego został przeprowadzony eksperyment. Dwadzieścia zmiennych zostało losowo wybranych ze zbioru programów pascalowych i liczba X zmiennych typu tablicowego została odnotowana. Celem poznawczym jest zweryfikowanie hipotezy, że pascal jest językiem o większej wydolności (tj. ma większy udział zmiennych typu tablicowego) niż algol, dla którego, jak pokazało doświadczenie, jedynie 20% zmiennych jest typu tablicowego.

a) Skonstruować test statystyczny do zweryfikowania postawionej hipotezy.

b) Znaleźć α dla zbioru odrzuceń $X \geqslant 8$.

c) Znaleźć α dla zbioru odrzuceń $X \geqslant 5$.

d) Znaleźć β dla zbioru odrzuceń $X \geqslant 8$, jeżeli $p = 0,5$ (doświadczenie pokazuje, że około połowa zmiennych w programach pascalowskich jest typu tablicowego).

e) Znaleźć β dla zbioru odrzuceń $X \geqslant 5$, jeżeli $p = 0,5$.

f) Który ze zbiorów odrzuceń $X \geqslant 8$ czy $X \geqslant 5$ jest bardziej pożądany, jeżeli minimalizowany jest:

   A) błąd I rodzaju?

   B) błąd II rodzaju?

g) Znaleźć jednostronny zbiór odrzuceń postaci $X \geqslant a$, tak aby poziom ufności był w przybliżeniu równy $\alpha = 0,01$.

h) Dla zbioru odrzuceń wyznaczonego w poprzednim punkcie znaleźć moc testu, jeżeli $p = 0,4$.

i) Dla zbioru odrzuceń wyznaczonego w punkcie g) znaleźć moc testu, jeżeli $p = 0,7$.

18.1 a)

Oznaczmy jako X liczbę zmiennych typu tablicowego w Pascal z losowo wybranych, wtedy X ma rozkład dwumianowy z nieznanym parametrem p. Załóżmy że liczby typu tablicowego wylosowane z programu Algog ma także rozkład dwumianowy gdzie natomiast jest znany parametr $p = 0.2$. Wtedy hipoteza że Pascal jest językiem programowania o większej zdolności niż Algol, czyli że p liczb typu tablicowego w Pascal jest większa niż p dla liczb tablicowych w Algol. Jest to hipoteza alternatywa ponieważ wstępuje ostra nierówność, zatem stosując normy statystyki można wyznaczyć hipotezę zerowa.

H_0	$p \leqslant p_0 = 0.2$
H_1	$p > p_0 = 0.2$

Znany jest rozkład ale nie znany jest parametr p, musimy sprawdzić poniższy warunek aby móc zastosować statystykę.

$$0 < p_0 \mp \sqrt{\frac{p_0(1-p_0)}{n}} = 0.2 \mp \sqrt{\frac{0.16}{20}} < 1$$

Spełniony jest warunek zatem możemy zastosować statystykę która jest zbliżona do rozkładu $N(0,1)$:

$$Z = \frac{\overline{P}_n - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

18.2 b)

Zbiór odrzuceń jest zbiorem dla którego, jeżeli liczba liczb tablicowych się znajdzie odrzucamy hipotezę zerową tj że Pascal jest mniej wydajny. Zatem dla $X \geqslant 8$ zbiór krytyczny jest następujący:

$$R = \{8, 9, \ldots, 20\}$$

Wtedy błąd pierwszego rodzaju, zakładając że dla H_0 $p = 0.2$ jest następujący:

$$\begin{aligned}\alpha &= P(U_n \in R | H_0) = P(X \geqslant 8) = 1 - P(X \leqslant 7) \\ &\overset{R}{=} 1 - pbinom(7, 20, 0.2) \approx 0.03214266\end{aligned}$$

18.3 c)

Podobnie jak w poprzednim podpunkcie wyznaczymy zbiór krytyczny:

$$R = \{5, 6, \ldots, 20\}$$

Wtedy, zakładając jak poprzednio, błąd pierwszego rodzaju wynosi:

$$\begin{aligned}\alpha &= P(U_n \in R | H_0) = P(X \geqslant 5) = 1 - P(X \leqslant 4) \\ &\overset{R}{=} 1 - pbinom(4, 20, 0.2) \approx 0.3703517\end{aligned}$$

18.4 d)

Jak dla podpunktu b, zbiór wartości krytycznych jest następujący:

$$R = \{8, 9, \ldots, 20\}$$

Natomiast potrzebujemy obliczyć błąd drugiego rodzaju, to znaczy że zakładamy że hipoteza alternatywna jest prawdziwa, czyli że Pascal jest bardziej

wydajnym programem, i zakładamy że $p = 0.5$. Wtedy szykane β wyraża się następującym wzorem:

$$\begin{aligned}\beta &= 1 - P(U_n \in R|H_1) = 1 - P(X \geqslant 8) = P(X \leqslant 7) \\ &\overset{R}{=} pbinom(7, 20, 0.5) \approx 0.131588\end{aligned}$$

Zatem β wynosi około 0.1326.

18.5 e)

Zbiór wartości krytycznych jest jak w podpunkcie c:

$$R = \{5, 6, \ldots, 20\}$$

Natomiast obliczamy błąd drugiego rodzaju jak dla poprzedniego podpunktu, czyli:

$$\begin{aligned}\beta &= 1 - P(U_n \in R|H_1) = 1 - P(X \geqslant 5) = P(X \leqslant 4) \\ &\overset{R}{=} pbinom(4, 20, 0.5) \approx 0.005908966\end{aligned}$$

Zatem β wynosi około 0.0059.

18.6 f)

Nie rozwiązane.

18.7 g)

Szukamy wartość a taka aby błąd pierwszego rodzaju był równy 0.01. Zatem, korzystając z poprzednich podpunktów można tę wartość wyznaczyć:

$$\begin{aligned}\alpha = 0.01 &= 1 - P(X \leqslant a - 1) \\ P(X \leqslant a - 1) &= 0.99 \\ a - 1 = F^{-1}(0.99) &\overset{R}{=} qbinom(0.99, 20, 0.2) = 8 \\ a &= 9\end{aligned}$$

Zatem dla $X \geqslant 9$ błąd pierwszego rodzaju wynosi 0.01.

18.8 h)

Jak w poprzednich podpunktach obliczymy błąd drugiego rodzaju i na jego podstawie moc testu. Przyjmujemy wartość $p = 0.4$.

$$\beta = 1 - P(X \leqslant 8) \overset{R}{=} 1 - pbinom(8, 20, 0.4) \approx 0.4044013$$

Wtedy moc testu wynosi $1 - \beta = 0.5955987$.

18.9 i)

Nie rozwiązane. Jak w poprzednim podpunkcie obliczymy moc testu na podstawie błędu drugiego rodzaju przyjmując $p = 0.7$.

$$\beta = 1 - P(X \leqslant 8) \overset{R}{=} 1 - pbinom(8, 20, 0.7) \approx 0.9948618$$

Wtedy moc testu wynosi $1 - \beta = 0.005138162$.

19 Zadanie 7

Zbadano czułość 80 telewizorów i uzyskano $\overline{x} = 348[mV]$ i $s = 107[mV]$. Na poziomie istotności $\alpha = 0,05$ zweryfikować hipotezę, że odchylenie standardowe czułości jest większe od nominalnej wartości wynoszącej 100[mV].

Wyznaczmy najpierw hipotezę zerową. Ponieważ szukamy aby odchylenie standardowe było większe od pewnej wartości przyjmujemy że to hipoteza alternatywna. Zatem szukane hipotezy będą wyglądać następująco:

$$H_0 : \sigma \leqslant \sigma_0 = 100[mV]$$
$$H_1 : \sigma > \sigma_0 = 100[mV]$$

Nie jest znany rozkład czułości telewizora i nieznane są także parametry tego rozkładu. Zatem zastosujemy statystykę:

$$Z = \frac{S_n^2 - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{n}{2}}$$

Dla wartości $\sigma_0 = 100[mV]$, $S_{80} = 107[mV]$ i $n = 80$. Zakładając że n jest wystarczająco duże rozkład ten jest zbliżony do rozkładu $N(0, 1)$.

Obliczymy teraz wartość Z_0 która pozwoli nam obliczyć wartość *p value* aby sprawdzić prawdziwość hipotezy alternatywnej.

$$Z_0 = \frac{107^2 - 100^2}{100^2}\sqrt{\frac{80}{2}} \approx 0.916428$$

Wtedy można obliczyć *p value* następująco:

$$\text{p value} = 1 - \Phi(Z_0) \overset{R}{=} 1 - pnorm(0.916428, 0, 1) \approx 0.1797212$$

Ponieważ wartość p jest większa od α nie mamy podstaw żeby odrzucić hipotezę zerową. Zatem odchylenie standardowe może być mniejsze od wartości nominalnej.

20 Zadanie 10

Wzrost losowo wybranej osoby z pewnej populacji ma rozkład normalny o nieznanych parametrach. Pobrano próbę losową o liczności $n = 26$ i po obliczeniu przedziału ufności na poziomie 0,9 otrzymano następujący wynik: (162;178)(cm). Wygenerować próbę złożoną z 26 pomiarów według rozkładu $N(\overline{x}, s_{26})$ i na poziomie istotności 0,05 zweryfikować hipotezy

a) średni wzrost ludzi z badanej populacji jest większy od 178 cm.

b) odchylenie standardowe wzrostu ludzi z badanej populacji jest mniejsze od 24 cm.

Wyznaczymy najpierw parametr α. Ponieważ ufność wynosi 0.9 wtedy $\alpha = 0.1$.
Następnie za pomocą tabeli na przedział ufności wartości oczekiwanej wyznaczymy średnią i wariancję z próby.

$$\overline{X}_n \mp t_{1-\frac{\alpha}{2};n-1}\frac{S_n}{\sqrt{n}}$$

$$t_{0.95;25} \overset{R}{=} qt(0.95, 25) \approx 1.708141$$

$$\begin{cases} \overline{X}_{26} - t_{0.95;25}\frac{S_{26}}{\sqrt{26}} = 162 \\ \overline{X}_{26} + t_{0.95;25}\frac{S_{26}}{\sqrt{26}} = 178 \end{cases}$$

$$\begin{cases} \overline{X}_{26} = 0.334994 \cdot S_{26} + 162 \\ \overline{X}_{26} = -0.334994 \cdot S_{26} + 178 \end{cases}$$

$$\begin{cases} \overline{X}_{26} = 0.334994 \cdot S_{26} + 162 \\ 0.334994 \cdot S_{26} + 162 = -0.334994 \cdot S_{26} + 178 \end{cases}$$

$$\begin{cases} \overline{X}_{26} = 0.334994 \cdot S_{26} + 162 \\ 0.669988 \cdot S_{26} = 16 \end{cases}$$

$$\begin{cases} \overline{X}_{26} = 0.334994 \cdot S_{26} + 162 \\ S_{26} = 23.881025 \end{cases}$$

$$\begin{cases} \overline{X}_{26} = 170 \\ S_{26} = 23.881025 \end{cases}$$

Próbę losową według rozkładu $N(\overline{x}, s_{26})$ wygenerowano w R i przedstawiona poniżej; wartości zaokrąglone do 5 liczb bo przecinku.

Lp	Vart
1	206.75631
2	150.65441
3	209.90324
4	135.2722
5	164.51461
6	208.97946
7	205.83026
8	133.27125
9	160.10169
10	151.44174
11	141.06375
12	159.56375
13	185.89217
14	173.9433
15	169.70846
16	168.77696
17	180.32007
18	145.65124
19	232.2268
20	163.71309
21	167.86651
22	191.62654
23	180.19727
24	162.64296
25	158.25598
26	186.41663

$\overline{X}_{26}$ i S^2_{26} obliczono zgodnie z odpowiednimi wzorami będą wykorzystane do dalszych obliczeń i wynoszą:

$$\overline{X}_{26} = 172.8688712 \approx 173$$
$$S^2_{26} = 631.9613666$$
$$S_{26} = 25.13884179$$

20.1 a)

Zakładamy że średni wzrost populacji jest wartością m. Wtedy hipoteza że $m > 178$ jest hipotezą alternatywną i, zgodnie z normami statystyki można wyznaczyć hipotezę zerową.

H_0	$m \leqslant 178$
H_1	$m > 178$

Ponieważ znamy rozkład ale nie znamy jego parametrów wyznaczymy sta-

tystykę zgodnie z tabelami.

$$t = \frac{\overline{X}_n - m_0}{\frac{S_n}{\sqrt{n}}}$$

Obliczymy teraz t_0 podstawiając m_0 z hipotezy i wartości obliczone w wygenerowanej próby.

$$t_0 = \frac{173 - 178}{\frac{25.13884179}{\sqrt{26}}} \approx -1.014172$$

Statystyka ta ma rozkład statystyczny zbliżony do rozkładu t-Studenta z $n-1$ stopniami swobody. Można teraz obliczyć przedział krytyczny dla $\alpha = 0.05$.

$$t_{1-0.05;25} \overset{R}{=} qt(0.95, 25) \approx 1.708141$$
$$(1.708141, \infty)$$

Wartość t_0 nie należy do przedziału krytycznego, zatem nie możemy odrzucić hipotezę zerową; zatem nie wiem czy hipoteza alternatywna jest prawdziwa lub nie.

20.2 b)

Hipoteza że odchylenie standardowe jest mniejsze od 24 jest hipoteza alternatywna. Zatem jak poprzednio wyznaczymy hipotezę zerową.

H_0	$\sigma \geqslant 24$
H_1	$\sigma < 24$

Przeprowadzimy natomiast test dla wariancji i z tego testu wywnioskujemy hipotezę dla odchylenia standardowego

Ponieważ znamy rozkład ale nie znamy jego parametrów, zgodnie z tabelami skorzystamy z następującej statystyki:

$$\chi^2 = \frac{(n-1)S_n^2}{\sigma_0^2}$$

Statystyka ta ma w przybliżeniu rozkład statystyki chi kwadrat z $n-1$ stopniami swobody.

Obliczymy teraz χ_0^2 podstawiając odpowiednie wartości.

$$\chi_0^2 = \frac{25 \cdot 631.9613666}{24^2} \approx 27.428879$$

Zgodnie z tabelami wyznaczymy przedział krytyczny.

$$\chi^2_{\alpha;n-1} \overset{R}{=} qchisq(0.05, 25) \approx 14.61141$$
$$(0, 14.61141)$$

Ponownie wartość χ_0^2 nie należy do przedziału krytycznego zatem nie możemy odrzucić hipotezę zerową. Nie wiemy po za tym czy jest ona prawdziwa czy nie.

21 Zadanie 13

Czuły przyrząd pomiarowy powinien mieć niewielką wariancję błędów pomiaru. W próbie 41 błędów pomiaru stwierdzono wariancję 102 [j.m.]^2. Na poziomie istotności $\alpha_1 = 0,05$ i $\alpha_2 = 0,01$ zweryfikować hipotezy:

a) wariancja błędów pomiaru wynosi 120 [j.m.]^2;

b) wariancja błędów pomiaru wynosi poniżej 120 [j.m.]^2.

21.1 a)

Hipoteza że wariancja błędów pomiaru wynosi 120 [j.m.]^2 jest hipotezą zerową, zatem można wyznaczyć hipotezę alternatywną jako jej przeciwieństwo.

H_0	$\sigma^2 = 120$[j.m.]^2
H_1	$\sigma^2 \neq 120$[j.m.]^2

Ponieważ nie znamy rozkład błędów pomiaru ani ich parametrów skorzystamy z następującej statystyki:

$$Z = \frac{S_n^2 - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{n}{2}}$$

Która ma w przybliżeniu rozkład statystyki $N(0,1)$.

Obliczymy teraz Z_0 potrzebne do dalszych rozważań postawiając znaną wariancje z próby i σ_0^2.

$$Z_0 = \frac{102 - 120}{120}\sqrt{\frac{41}{2}} \approx -4.482416$$

Wyznaczymy teraz obszary krytyczne zgodnie z tabelami.
Dla α_1:

$$z_{\frac{0.05}{2}} \overset{R}{=} qnorm(0.05/2, 0, 1) \approx -1.959964$$

$$z_{1-\frac{0.05}{2}} \overset{R}{=} qnorm(1 - 0.05/2, 0, 1) \approx 1.959964$$

$$(-\infty, -1.959964) \cup (1.959964, \infty)$$

Dla α_2:

$$z_{\frac{0.01}{2}} \overset{R}{=} qnorm(0.01/2, 0, 1) \approx -2.575829$$

$$z_{1-\frac{0.01}{2}} \overset{R}{=} qnorm(1 - 0.01/2, 0, 1) \approx 2.575829$$

$$(-\infty, -2.575829) \cup (2.575829, \infty)$$

Wartość Z_0 należy do obszary krytycznego dla oby α, zatem odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną czyli $\sigma^2 = 120$[j.m.]^2.

21.2 b)

Hipotezą że $\sigma^2 > 120[\text{j.m.}]^2$ wariancja błędów jest hipotezą alternatywną, zatem, zgodnie z normami statystki wyznaczamy hipotezę zerową.

H_0	$\sigma^2 \leqslant 120[\text{j.m.}]^2$
H_1	$\sigma^2 > 120[\text{j.m.}]^2$

Statystyka dla tego podpunktu jest taka sama jak w poprzednim podpunkcie, natomiast zmienia się obszar krytyczny.
Zgodnie z tabelami, dla wartości α_1:

$$z_{1-0.05} \overset{R}{=} qnorm(0.95, 0, 1) \approx 1.644854$$

$$(1.281552, \infty)$$

Dla α_2:

$$z_{1-0.01} \overset{R}{=} qnorm(0.99, 0, 1) \approx 2.326348$$

$$(1.281552, \infty)$$

Wartość Z_0 nie należy do obszaru krytycznego, zatem nie możemy odrzucić hipotezy zerowej. Obliczymy zatem wartość *p value* jako:

$$\text{p value} = 1 - \Phi(Z_0) = 1 - \Phi(-4.482416) \overset{R}{=} 1 - pnorm(-4.482416) \approx 0.9999963$$

Ponieważ jest to wartość większa od obu α nie możemy odrzucić hipotezę zerową.

22 Zadanie 16

Dla wylosowanej próby studentów otrzymano następujący rozkład tygodniowego czasu nauki (w godz.):

Czas nauki	[0, 2)	[2, 4)	[4, 6)	[6, 8)	[8, 10)	[10, 12)
Liczba studentów	10	28	42	30	15	7

Na poziomach istotności $\alpha_1 = 0,1$ i $\alpha_2 = 0,01$ sprawdzić hipotezy:

a) średni czas poświęcony tygodniowo na naukę dla badanej populacji studentów wynosi 6 godz.

b) średni czas poświęcony tygodniowo na naukę dla badanej populacji studentów wynosi poniżej 6 godz.;

c) wariancja tego czasu wynosi 4godz.^{2};

d) wariancja tego czasu wynosi ponad 4godz.^{2}.

Jako pierwsze obliczono średnią z podanej próby i wariancję korzystając z następujących wzorów i zakładając środek przedziały jako przedstawiciel przedziału:

$$\overline{X} = \frac{\sum x_i \cdot n_i}{N} = \frac{726}{132} = 5.5$$

$$S_n^2 = \frac{\sum (x_i - \overline{X})^2 \cdot n_i}{N} = \frac{851}{132} \approx 6.496183$$

$$S_n = \sqrt{S_n^2} \approx 2.548761$$

22.1 a)

Ponieważ nie znany jest rozkład zmiennej losowej opisującej czas poświęcony tygodniowo na naukę przez studenta, ani nie są znane jego parametry zastosujemy statystykę następującą:

$$Z = \frac{\overline{X}_n - m_0}{\frac{S_n}{\sqrt{n}}}$$

Która ma rozkład statystyki zbliżony do rozkładu normalnego $N(0, 1)$.
Podane hipotezy są następujące:

	H_0	H_1
α_1	m = 6	$m \neq 6$
α_2	m = 6	$m \neq 6$

Obliczymy teraz Z_0 podstawiając wartości hipotezy do statystyki:

$$Z_0 = \frac{5.5 - 6}{\frac{2.548761}{\sqrt{132}}} \approx -2.253866$$

Wyznaczymy obszar krytyczny zgodnie z tabelami. Dla α_1.

$$z_{\frac{0.1}{2}} \overset{R}{=} qnorm(0.1/2, 0, 1) \approx -1.644854$$

$$z_{1-\frac{0.1}{2}} \overset{R}{=} qnorm(1 - 0.1/2, 0, 1) \approx 1.644854$$

$$(-\infty, -1.644854) \cup (1.644854, \infty)$$

Dla α_2.

$$z_{\frac{0.01}{2}} \overset{R}{=} qnorm(0.01/2, 0, 1) \approx -2.575829$$

$$z_{1-\frac{0.01}{2}} \overset{R}{=} qnorm(1 - 0.01/2, 0, 1) \approx 2.575829$$

$$(-\infty, -2.575829) \cup (2.575829, \infty)$$

Ponieważ wartość Z_0 należy do obszaru krytycznego dla α_1, odrzucamy hipotezę zerową; natomiast dla α_2 wartość Z_0 nie wpada pod obszar krytyczny, zatem nie mamy mocy aby odrzucić hipotezę zerową.

22.2 b)

Obliczenia przechodzą jak w poprzednim podpunkcie ale zmieniają się hipotezy. Hipoteza że $m < 6$ godzin jest hipotezą alternatywną, zatem:

	H_0	H_1
α_1	$m \geqslant 6$	$m < 6$
α_2	$m \geqslant 6$	$m < 6$

Wartość Z_0 pozostaje nie zmieniona, natomiast zmienia się obszar krytyczny który zgodnie z tabelami obliczymy.
Dla α_1:

$$z_{0.1} \overset{R}{=} qnorm(0.1, 0, 1) \approx -1.281552$$

$$(-\infty, -1.644854$$

Dla α_2:

$$z_{0.01} \overset{R}{=} qnorm(0.01, 0, 1) \approx -2.326348$$

$$(-\infty, -2.326348)$$

Jak poprzednio, dla α_1 odrzucamy hipotezę zerową, natomiast dla α_2 nie mamy mocy aby to zrobić.

22.3 c)

Hipoteza że $\sigma^2 = 4$ jest hipotezą zerową, zatem można wyznaczyć hipotezę alternatywną:

H_0	$\sigma^2 = 4$
H_1	$\sigma^2 \neq 4$

Dla sprawdzenia wariancji, ponieważ nie znamy rozkładu ani jego parametrów a mamy wystarczająco dużą próbę skorzystamy ze statystyki:

$$Z = \frac{S_n^2 - \sigma_0^2}{\sigma_0^2}\sqrt{\frac{n}{2}}$$

Która ma rozkład statystyki zbliżony do rozkładu $N(0,1)$.
Obliczymy teraz Z_0 podstawiając $\sigma_0^2 = 4$ i wartości wcześniej obliczone.

$$Z_0 = \frac{6.496183 - 4}{4}\sqrt{\frac{132}{2}} \approx 5.069772$$

Ponieważ obszary krytyczne są takie same jak w poprzednich podpunktach wnioskujemy że, skoro wartość Z_0 dla obu α należy do obszaru krytycznego, odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną $\sigma^2 \neq 4$.

22.4 d)

Hipoteza że $\sigma^2 > 4$ jest hipotezą alternatywną, zatem można wyznaczyć hipotezę zerową korzystając z praw statystyki:

H_0	$\sigma^2 \leqslant 4$
H_1	$\sigma^2 > 4$

Statystyka i wartość Z_0 są takie same jak poprzednim podpunkcie, natomiast zmienia się obszar krytyczny który obliczymy zgodnie z tabelami.
Dla α_1:

$$z_{1-0.1} \overset{R}{=} qnorm(0.9, 0, 1) \approx 1.281552$$

$$(1.281552, \infty)$$

Dla α_2:

$$z_{1-0.01} \overset{R}{=} qnorm(0.99, 0, 1) \approx 2.326348$$

$$(2.326348, \infty)$$

Możemy zatem powiedzieć, skoro wartość Z_0 należy do obszarów krytycznych, że odrzucamy hipotezę zerową i przyjmujemy hipotezę alternatywną że $\sigma^2 > 4$.

Część IV
Laboratoria 10

Celem tych laboratoriów było zapoznanie się z metodami testów dla dwóch lub więcej populacji. Dotyczyło to porównania wartości oczekiwanych, ilorazu dwóch wariancji oraz różnice wskaźników struktury. Dla tych testów nauczyliśmy się także wyznaczać przedziały ufności oraz obliczenie *p-value*. Wzbogaciliśmy także wiedze interpretowania wyników.

23 Zadanie 1 (TG 6.32)

Zbadano wzrost 13 mężczyzn oraz 12 kobiet w pewnym ośrodku sportowym. Dane:
M: 171, 176, 179, 189, 176, 182, 173, 179, 184, 186, 189, 167, 177,
K: 161, 162, 163, 162, 166, 164, 168, 165, 168, 157, 161, 172.
Zakładając, że w obu populacjach rozkład wzrostu jest normalny, czy można powiedzieć, że mężczyźni charakteryzują się większą zmiennością wzrostu? Przyjąć poziom istotności 0,1.

Wyznaczono na początku średnią i wariancję nieobciążoną z podanych prób zgodnie ze wzorami poniżej:

$$\overline{X} = \frac{\sum x_i}{n}$$
$$S^2 = \frac{\sum (x_i - \overline{X})^2}{n-1}$$

Otrzymano następujące wartości:

	$\overline{X}$	S^2	n
M	179.076923	45.74359	13
K	164.083333	16.083333	12

Pytanie czy mężczyźni charakteryzują się większą zmiennością wzrostu można przedstawić jako hipotezę alternatywną wraz z hipotezą zerową następująco:

H_0	$\sigma_M^2 \leqslant \sigma_K^2$
H_1	$\sigma_M^2 > \sigma_K^2$

Ponieważ zakładamy że rozkłady są typu normalnego możemy zastosować test F Snedecora. Poziom istotności $\alpha = 0.1$.

$$F = \frac{max\{S_M^2, S_K^2\}}{min\{S_M^2, S_K^2\}} = \frac{45.74359}{16.083333} \approx 2.844161 = F_0$$

Statystyka ta ma rozkład statystki F Snedeora ze stopniami swobody licznika i mianownika odjąć jeden. W tym przypadku:

$$F(13-1, 12-1) = F(12, 11)$$

Obliczymy teraz *p value* zgodnie ze wzorem:

$$\text{p-value} = 1 - F_{F(12,11)}(F_0) \overset{R}{=} 1 - pf(2.844161, 12, 11) \approx 0.04689163$$

Ponieważ *p-value* jest mniejsze od przyjętego poziomu istotności odrzucamy hipotezę zerową i wnioskujemy że wariancją wzrostu mężczyzn jest znacznie większa niż wariancja wzrostu kobiet.

24 Zadanie 2

Spośród absolwentów pewnej uczelni wylosowano 15 osób z jednego wydziału oraz 12 osób z drugiego wydziału i obliczono średnią ocen ze studiów dla każdego absolwenta. Otrzymano następujące wyniki
dla pierwszego wydziału: 3.71, 4.28, 2.95, 3.20, 3.38, 4.05, 4.07, 4.98, 3.20, 3.43, 3.09, 4.50, 3.12, 3.68, 3.90,
dla drugiego wydziału: 3.10, 3.38, 4.06, 3.60, 3.81, 4.50, 4.00, 3.25, 4.11, 4.85, 2.80, 4.00.
Na poziomie istotności $\alpha = 0,05$ zweryfikować następujące hipotezy:

a) wariancje średnich ocen dla obydwu wydziałów są równe,

b) różnica wartości oczekiwanych ocen uzyskiwanych przez studentów obydwu wydziałów wynosi 0.

Wyznaczono na początku średnią i wariancję nieobciążoną z podanych prób zgodnie ze wzorami poniżej:

$$\overline{X} = \frac{\sum x_i}{n}$$
$$S^2 = \frac{\sum (x_i - \overline{X})^2}{n-1}$$

Otrzymano następujące wartości:

	$\overline{X}$	S^2	n
W1	3.702667	0.347207	15
W2	3.788333	0.349415	12

24.1 a)

Hipoteza że wariancję są sobie równe jest hipotezą zerową, zatem można wyznaczyć hipotezę alternatywną następująco:

H_0	$\sigma^2_{W1} = \sigma^2_{W2}$
H_1	$\sigma^2_{W1} \neq \sigma^2_{W2}$

Do oceny wariancji zastosujemy test F Snedecora wyrażony następująco:

$$F = \frac{max\{S^2_{W1}, S^2_{W2}\}}{min\{S^2_{W1}, S^2_{W2}\}} = \frac{0.349415}{0.347207} \approx 1.006359 = F_0$$

Gdzie statystyka F ma rozkład statystyki F Snedecora ze stopniami swobody licznika i mianownika obniżone o jeden, czyli:

$$F(12-1, 15-1) = F(11, 14)$$

Obliczymy teraz *p-value* zgodnie ze wzorem:

$$\text{p-value} = 2 \cdot min\{F_{F(11,14)}(F_0), 1 - F_{F(11,14)}(F_0)\}$$

$$\boxed{\begin{array}{c} F_{F(11,14)}(F_0) \overset{R}{=} pf(1.006359, 11, 14) \approx 0.513539 \\ 1 - F_{F(11,14)}(F_0) \overset{R}{=} 1 - pf(1.006359, 11, 14) \approx 0.486461 \end{array}}$$

$$= 2 \cdot 0.486461 = 0.972922$$

Ponieważ *p-value* jest większe od przyjętego $\alpha = 0.05$ nie możemy odrzucić hipotezę zerową, zatem jest możliwe że wariancje średnich ocen obydwu wydziałów są sobie równe.

24.2 b)

Podana hipoteza jest hipotezą zerową, zatem można wyznaczyć hipotezę alternatywną:

H_0	$m_{W1} - m_{W2} = 0$
H_1	$m_{W1} - m_{W2} \neq 0$

Ponieważ test przeprowadzony w poprzednim podpunkcie wykazało że wariancje obu populacji są sobie równe, możemy, do tego testu zastosować następującą statystykę:

$$t = \frac{(\overline{X}_{W1} - \overline{X}_{W2}) - m_0}{\sqrt{\frac{(n_{W1}-1)S^2_{W1} + (n_{W2}-1)S^2_{W2}}{n_{W1}+n_{W2}-2} \frac{n_{W1}+n_{W2}}{n_{W1} \cdot n_{W2}}}} \sim t(n_{W1} + n_{W2} - 2)$$

Wtedy t_0 będzie równe:

$$t_0 = \frac{3.702667 - 3.788333}{\sqrt{\frac{14 \cdot 0.347207 + 11 \cdot 0.349415}{25} \frac{27}{15 \cdot 12}}} \approx -0.374854$$

Obliczymy teraz *p-value* zgodnie ze wzorem:

$$\text{p-value} = 2 \cdot min\{F_{t(25)}(t_0), 1 - F_{t(25)}(t_0)\}$$

$$\boxed{\begin{array}{c} F_{t(25)}(t_0) \overset{R}{=} pt(-0.374854, 25) \approx 0.3554651 \\ 1 - F_{t(25)}(t_0) \overset{R}{=} 1 - pt(-0.374854, 25) \approx 0.6445349 \end{array}}$$

$$= 2 \cdot 0.3554651 = 0.7109302$$

Widzimy że *p-value* jest większe od przyjętego $\alpha = 0.05$, zatem nie możemy odrzucić hipotezę zerową. Możemy powiedzieć że wartości oczekiwane średnich ocen z obu wydziałów są sobie równe.

25 Zadanie 3

Badano opony samochodowe typu 11.00-20/14PR/ produkowane przez dwóch producentów, które zostały wycofane z eksploatacji. Spośród zbadanych 1582 opon producenta A, 1250 opon wycofano z powodu zużycia bieżnika, a spośród 589 zbadanych opon producenta B, wycofano z powodu tego defektu 421 sztuk. Na poziomie istotności $\alpha = 0,01$ zweryfikować hipotezę, że frakcje opon wycofanych z eksploatacji na skutek zużycia się bieżnika są jednakowe dla obydwu producentów.

Wyrażona hipoteza jest hipotezą zerową, jako alternatywną wybierzemy że frakcje opon wycofanych z eksploatacji dla producenta A jest większa niż ta z producenta B, czyli:

H_0	$p_A - p_B = 0$
H_1	$p_A - p_B > 0$

Wyznaczymy wskaźnik struktury dla obu prób dzieląc liczebność wycofanych opon przez całkowitą ich ilość:

	$\overline{P}_k$	n
A	0.790139	1582
B	0.714771	589

Dla obu prób sprawdzono warunek:

$$\overline{p}_k \mp 3 \cdot \sqrt{\frac{\overline{p}_k(1-\overline{p}_k)}{n_k}}$$

$$A : 0.759425, 0.820853$$

$$B : 0.658957, 0.770585$$

Wszystkie te liczby należą do przedziału $(0, 1)$, zatem warunek jest spełniony. Ponieważ liczebność obu populacji jest wystarczająco duża można zastosować następującą statystykę:

$$Z = \frac{\overline{P}_A - \overline{P}_B}{\sqrt{\overline{P}(1-\overline{P})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

Gdzie $\overline{P}$ jest iloczynem sumy wyróżnionych elementów oby prób przez sumę całkowitych elementów oby prób:

$$\overline{P} = \frac{k_A + k_B}{n_A + n_B} = \frac{1671}{2171} \approx 0.769691$$

Statystyka ta ma rozkład statystki $\sim N(0, 1)$. Obliczymy teraz Z_0 podstawiając znane wartości:

$$Z_0 = \frac{0.790139 - 0.714771}{\sqrt{0.769691(1-0.769691)(\frac{1}{1582} + \frac{1}{589})}} \approx 3.708551$$

Obliczymy teraz *p-value* zgodnie ze wzorem:

$$\text{p-value} = 1 - \Phi(3.708551) \overset{R}{=} 1 - pnorm(3.708551, 0, 1) \approx 0.0001042243$$

Liczba ta jest zdecydowanie mniejsza niż przyjęty $\alpha = 0.01$, zatem odrzucamy hipotezę zerową i wnioskujemy że frakcja opon wycofanych z eksploatacji na skutek zużycie się biernika producenta A jest większa niż ta producenta B.

26 Zadanie 4

Dział kontroli technicznej uzyskał czasy r1 i r2 palenia się dwu rodzajów świateł ostrzegawczych (w sekundach):

$$r1 = \{15,3; 19,4; 21,5; 17,4; 16,8; 16,6; 20,3; 21,3; 22,5; 23,4; 19,7; 21,0\}$$

,

$$r2 = \{24,7; 16,5; 15,8; 10,2; 13,5; 15,9; 15,7; 14,0; 12,1; 17,4; 15,6; 15,8\}$$

. Na poziomie istotności $\alpha = 0,05$, zweryfikować hipotezy:

a) przeciętne czasy palenia się świateł ostrzegawczych dla obydwu rodzajów różnią się,

b) przeciętny czas palenia się świateł ostrzegawczych pierwszego rodzaju jest dłuższy o 5 sekund niż dla drugie-go rodzaju.

c) wariancje czasów palenia się świateł różnią się.

Obliczymy średnią i wariancję dla obu czasów zgodnie ze wzorami:

$$\overline{X} = \frac{\sum x_i}{n}$$
$$S_n^2 = \frac{\sum (x_i - \overline{X})^2}{n-1}$$

Otrzymano następujące wartości:

	$\overline{X}_{12}$	S_{12}^2
r1	19.6	6.547273
r2	15.6	12.31091

26.1 a)

Podana w treści hipoteza jest hipotezą alternatywną. Zatem, stosując podstawy statystyki wyznaczymy hipotezę zerową:

H_0	$m_{r1} - m_{r2} = 0$
H_1	$m_{r1} - m_{r2} \neq 0$

Załóżmy że $r1$ i $r2$ mają rozkład normalny z nieznanymi parametrami. Wtedy możemy zastosować statystykę Cochrana-Coxa:

$$t = \frac{(\overline{X}_{r1} - \overline{X}_{r2}) - m_0}{\sqrt{\frac{S_{r1}^2}{n_{r1}} + \frac{S_{r2}^2}{n_{r2}}}} \sim t(v)$$

Gdzie v wyznaczany jest następujący wzorem:

$$\begin{aligned} v &= \frac{\left(\frac{S_{r1}^2}{n_{r1}} + \frac{S_{r2}^2}{n_{r2}}\right)^2}{\frac{1}{n_{r1}-1}\left(\frac{S_{r1}^2}{n_{r1}}\right)^2 + \frac{1}{n_{r2}-1}\left(\frac{S_{r2}^2}{n_{r2}}\right)^2} \\ &= \frac{\left(\frac{6.547273}{12} + \frac{12.31091}{12}\right)^2}{\frac{1}{11}\left(\frac{6.547273}{12}\right)^2 + \frac{1}{11}\left(\frac{12.31091}{12}\right)^2} \\ &\approx 23.44327246 \end{aligned}$$

Obliczymy wartość t_0 podstawiając znane wartości:

$$t_0 = \frac{19.6 - 15.6}{\sqrt{\frac{6.547273}{12} + \frac{12.31091}{12}}} \approx 3.190808$$

Wyznaczymy teraz przedziały ufności dla podanego $\alpha = 0.05$:

$$t_{\frac{\alpha}{2};23.44327246} \overset{R}{=} qt(0.25, 23.44327246) \approx -0.685099$$

$$t_{1-\frac{\alpha}{2};23.44327246} \overset{R}{=} qt(0.75, 23.44327246) \approx 0.685099$$

$$R = (-\infty, -0.685099) \cup (0.685099, \infty)$$

Wartość t_0 należy do przedziału krytycznego, zatem odrzucamy hipotezę zerową i wnioskujemy że przeciętne czasy palenia żarówek różnią się.

26.2 b)

Wyznaczymy hipotezy dla tego podpunktu:

H_0	$m_{r1} - m_{r2} \leqslant 5$
H_1	$m_{r1} - m_{r2} > 5$

Rozkład statystyki jest taki sam, ale zmienia się wartość t_0:

$$t_0 = \frac{19.6 - 15.6 - 5}{\sqrt{\frac{6.547273}{12} + \frac{12.31091}{12}}} \approx 0.797702$$

Obliczymy *p-value* dla tego rozkładu:

$$\text{p-value} \overset{R}{=} pt(0.797702, 23.44327246) \approx 0.7834752$$

Jest to wartość znacznie większa od podanego α, zatem nie mamy podstawy aby odrzucić hipotezę zerową.

26.3 c)

Hipoteza że wariancje różnią się jest hipotezą alternatywną, zatem wyznaczymy hipotezę zerową:

H_0	$\sigma_{r1} = \sigma_{r2}$
H_1	$\sigma_{r1} \neq \sigma_{r2}$

Aby zbadać hipotezę zastosujemy statystykę F-Snedecora wyznaczona następująco:

$$F = \frac{max\{S_{r1}^2, S_{r2}^2\}}{min\{S_{r1}^2, S_{r2}^2\}} = \frac{12.31091}{6.547273} \approx 1.880311 = F_0$$

Ponieważ liczebności obu prób są równe, statystyka ta ma rozkład statystyki $\sim F(11, 11)$.
Możemy teraz wyznaczyć przedział krytyczny:

$$F_{1-\frac{\alpha}{2};11;11} \overset{R}{=} qf(0.75, 11, 11) \approx 1.518216$$

$$R = (1.518216, \infty)$$

Wartość F_0 znajduję się w tym przedziale, zatem odrzucamy hipotezę zerową i wnioskujemy że wariancję czasów spalania się żarówek są sobie różne.

27 Zadanie 6

W badaniu granicy plastyczności pewnego gatunku stali otrzymano następujące wyniki dla 15 kawałków tej stali
(wyniki w kg/cm^2): 3520, 3680, 3640, 3840, 3500, 3610, 3720, 3640, 3600, 3650, 3750, 3590, 3600, 3550, 3700.
Natomiast po dodatkowym procesie uszlachetniającym, mającym zwiększyć wytrzymałość tej stali, otrzymano dla tych samych próbek odpowiednio następujące wyniki badania granicy plastyczności:
3580, 3700, 3680, 3800, 3550, 3700, 3730, 3720, 3670, 3710, 3810, 3660, 3700, 3640, 3670.
Na poziomie istotności $\alpha = 0,05$ sprawdzić, czy granica plastyczności stali po dodatkowym procesie uszlachetniającym zwiększyła się.

Niech wyniki przed procesem będą X_1 a wyniki po procesie będą $X2$. Zbadamy $X_1 - X_2$, obliczymy średnią i wariancję zgodnie ze wzorami podanymi w poprzednich zadaniach.

Lp	$X_1 - X_2$
1	-60
2	-20
3	-40
4	40
5	-50
6	-90
7	-10
8	-80
9	-70
10	-60
11	-60
12	-70
13	-100
14	-90
15	30
SUM	-730
$\overline{X_1 - X_2}$	-48.666667
S^2	1769.52381
S	42.065708

Przedstawioną hipotezę można wyrazić następująco wraz z hipotezą zerową:

H_0	$m_{X_1} - m_{X_2} \geqslant 0$
H_1	$m_{X_1} - m_{X_2} < 0$

Zakładając że wyniki te mają rozkład normalny można zastosować następu-

jącą statystykę dla rozkładów sparowanych:

$$t = \frac{\overline{X_1 - X_2} - m_0}{\frac{S_{\overline{X_1 - X_2}}}{\sqrt{n}}} = \frac{-48.666667}{\frac{42.065708}{\sqrt{15}}} \approx -4.480733 = t_0$$

Statystyka ta ma rozkład statystyki $\sim t(n-1) = t(14)$. Obliczymy teraz *p-value* zgodnie ze wzorami:

$$\text{p-value} = F_{t(14)}(t_0) \stackrel{R}{=} pt(-4.480733, 14) \approx 0.0002589772$$

Wartość ta jest mniejsza od przyjętego $\alpha = 0.05$, zatem odrzucamy hipotezę zerową i wnioskujemy że granica plastyczności stali po dodatkowym procesie uszlachetniającym zwiększyła się.

28 Zadanie 7

Zmierzono czasy (w godzinach) usuwania awarii dla dwóch brygad remontowych. Dla pierwszej otrzymano czasy: 12, 13, 18, 25, 42, 19, 22, 35 a dla drugiej brygady: 23, 30, 27, 17, 21, 33, 31. Na poziomie istotności 0,05 zweryfikować hipotezy:

a) przeciętne czasy usuwania awarii dla obydwu brygad są równe,

b) wariancje czasów usuwania awarii dla obydwu brygad są równe.

Obliczono średnią i wariancję dla obu brygad i otrzymano następujące wyniki:

	$\overline{X}$	S^2	n
B1	23.25	110.214286	8
B2	26	34.333333	7

28.1 a)

Hipotezę tą można wyznaczyć wraz z hipotezą alternatywną w następujący sposób:

H_0	$m_{B_1} - m_{B_2} = 0$
H_1	$m_{B_1} - m_{B_2} \neq 0$

Jeżeli dane mają rozkład normalny możemy zastosować następującą statystykę:

$$t = \frac{(\overline{X}_{B_1} - \overline{X}_{B_2}) - m_0}{\sqrt{\frac{S^2_{B_1}}{n_{B_1}} + \frac{S^2_{B_2}}{n_{B_2}}}} \sim t(v)$$

Gdzie v jest wyrażony następującym wzorem:

$$\begin{aligned} v &= \frac{\left(\frac{S^2_{B_1}}{n_{B_1}} + \frac{S^2_{B_2}}{n_{B_2}}\right)^2}{\frac{1}{n_{B_1}-1}\left(\frac{S^2_{B_1}}{n_{B_1}}\right)^2 + \frac{1}{n_{B_2}-1}\left(\frac{S^2_{B_2}}{n_{B_2}}\right)^2} \\ &= \frac{\left(\frac{110.214286}{8} + \frac{34.333333}{7}\right)^2}{\frac{1}{7}\left(\frac{110.214286}{8}\right)^2 + \frac{1}{6}\left(\frac{34.333333}{7}\right)^2} \\ &\approx 11.213324 \end{aligned}$$

Obliczymy wartość t_0 podstawiając znane wartości:

$$t_0 = \frac{23.25 - 26}{\sqrt{\frac{110.214286}{8} + \frac{34.333333}{7}}} \approx -0.636248$$

Wyznaczymy teraz *p-value* zgodnie ze wzorem:

$$\text{p-value} = 2 \cdot min\{F_{t(11.213324)}(t_0), 1 - F_{t(11.213324)}(t_0)\}$$

$$\boxed{\begin{array}{c} F_{t(11.213324)}(t_0) \overset{R}{=} pt(-0.636248, 11.213324) \approx 0.2686926 \\ 1 - F_{t(11.213324)}(t_0) \overset{R}{=} 1 - pt(-0.636248, 11.213324) \approx 0.7313074 \end{array}}$$

$$= 2 \cdot 0.2686926 = 0.5373852$$

Wartość ta jest większa niż przyjęty $\alpha = 0.05$ zatem nie możemy odrzucić hipotezę zerową. Wnioskujemy że przeciętne czasy usuwania awarii dla obydwu brygad są równe.

28.2 b)

Hipotezę tą można wyznaczyć w następujący sposób wraz z hipotezą alternatywną:

H_0	$\sigma^2_{B_1} = \sigma^2_{B_2}$
H_1	$\sigma^2_{B_1} \neq \sigma^2_{B_2}$

Ponieważ badana jest wariancja zastosujemy statystykę F Snedecora wyrażona następująco:

$$F = \frac{max\{S^2_{B_1}, S^2_{B_2}\}}{min\{S^2_{B_1}, S^2_{B_2}\}} = \frac{110.214286}{34.333333} \approx 3.210125 = F_0$$

Gdzie statystyka F ma rozkład statystyki F Snedecora ze stopniami swobody licznika i mianownika obniżone o jeden, czyli:

$$F(8-1, 7-1) = F(7, 6)$$

Obliczymy teraz *p-value* zgodnie ze wzorem:

$$\text{p-value} = 2 \cdot min\{F_{F(7,6)}(F_0), 1 - F_{F(7,6)}(F_0)\}$$

$$\boxed{\begin{array}{c} F_{F(7,6)}(F_0) \overset{R}{=} pf(3.210125, 7, 6) \approx 0.9117178 \\ 1 - F_{F(7,6)}(F_0) \overset{R}{=} 1 - pf(3.210125, 7, 6) \approx 0.08828215 \end{array}}$$

$$= 2 \cdot 0.08828215 = 0.1765643$$

Liczba ta jest większa od przyjętego $\alpha = 0.05$ zatem nie możemy odrzucić hipotezę zerową. Wnioskujemy że wariancje czasów usuwania awarii dla obydwu brygad są sobie równe.

29 Zadanie 9

Wygenerować próby o liczebnościach 80 i 60 według rozkładów N(900;50) i N(1000;60) i na ich podstawie przeprowadzić test, że wartości oczekiwane różnią się o 50.

Przyjmijmy $\alpha = 0.05$. Podaną hipotezę można przedstawić następująco wraz z hipotezą zerową:

H_0	$m_1 - m_2 = 50$
H_1	$m_1 - m_2 \neq 50$

Gdzie m_1 jest nieznaną wartością oczekiwaną z próby $N(900; 50)$ a m_2 jest nieznaną wartością oczekiwaną z próby $N(1000; 60)$. Tabele wygenerowanej próby znajdują się na końcu pliku.
Wariancję i średnią z wygenerowanej próby obliczono w R za pomocą funkcji *mean()* dla średniej, i *var()* dla wariancji nie obciążonej. Utrzymano następujące wyniki:

	N1	N2
$\overline{X}$	899.4644	978.8235
S^2	2094.456	3204.845

Zakładając znane są wariancję z populacji możemy zastosować następującą statystykę:

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - m_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{899.4644 - 978.8235 - 50}{\sqrt{\frac{50}{80} + \frac{60}{60}}} \approx -101.477627 = Z_0$$

Wyznaczymy teraz przedział krytyczny wiedząc że statystyka ta ma w przybliżeniu rozkład statystyki $N(0, 1)$:

$$z_{\frac{\alpha}{2}} \overset{R}{=} qnorm(0.25, 0, 1) \approx -0.6744898$$

$$z_{1-\frac{\alpha}{2}} \overset{R}{=} qnorm(0.75, 0, 1) \approx 0.6744898$$

$$R = (-\infty, -0.6744898) \cup (0.6744898, \infty)$$

Wyznaczona wartość Z_0 znajduję się w obszarze krytycznym, zatem odrzucamy hipotezę zerową i wnioskujemy że wartości oczekiwane różnią się o 50.

30 Dane

Lp	N1
1	837.71
2	1014
3	903.5
4	927.4
5	822.95
6	896.64
7	887.16
8	885.61
9	819.63
10	793.28
11	871.83
12	932.02
13	894.1
14	872.22
15	934.94
16	943.1
17	902.11
18	815.1
19	920.56
20	860.9
21	830.93
22	926.59
23	937.53
24	900.73
25	934.26
26	799.76
27	878.97
28	939.1
29	897.3
30	899
31	922.75
32	742.5
33	961.97
34	897.39
35	915.18
36	846.9
37	943.2
38	888.99
39	931
40	899.54
41	957.55
42	956.11
43	956.54
44	906.14
45	888.97
46	919.11
47	809.89
48	901.7
49	927.8
50	846.31
51	887.97
52	877.64
53	903.52
54	859.87
55	901.88
56	873.51
57	854.52
58	887.07
59	898.03
60	868.11
61	929.89
62	961.89
63	901.24
64	856.2
65	943.55
66	897.22
67	977.59
68	884.56
69	913.55
70	933.02
71	920.08
72	944.09
73	960.17
74	929.28
75	920.1
76	933.35
77	927.47
78	905.34
79	896.03
80	911.94

Lp	N2
1	1009.3
2	921.27
3	989.81
4	988.38
5	1068.31
6	943.57
7	1037.98
8	1063.53
9	923.83
10	962.06
11	1012.57
12	1110.48
13	991.84
14	959.25
15	904.62
16	921.11
17	949.67
18	913.92
19	1001.19
20	958.45
21	1074.65
22	1027.91
23	971.89
24	1059.78
25	994.2
26	986.01
27	940.05
28	916.34
29	961.72
30	945.86
31	959.87
32	938.51
33	860.71
34	849.43
35	1057.9
36	1031.3
37	901.43
38	1005.31
39	975.97
40	984
41	966.68
42	999.01
43	1019.56
44	1050.89
45	878.05
46	969.09
47	1022.94
48	960.51
49	929.41
50	960.81
51	1024.3
52	1039.01
53	915.8
54	1067.31
55	979.72
56	992.32
57	1061.34
58	914.67
59	928.15
60	975.86

Część V
Laboratoria 11

Celem tych laboratoriów było zapoznanie się z testami nieparametrycznymi. Testy nieparametryczne to są testy do sprawdzenia czy dane mają znany rozkład statystyki, jak na przykład rozkład normalny. Nauczyliśmy się zatem zastosować tego typu testów za pomocą wspomagania komputerowego a także ręcznie. Uzyskaliśmy także wiedze o oszacowaniu rozkładu danych nie znając metodę generowania tych danych.

31 Zadanie 5

Do każdej z 20 tarcz oddano po 5 niezależnych strzałów i zanotowano liczbę trafień. Wyniki strzelania podane są w tabeli:

Liczba trafień	0	1	2	3	4	5
i)	0	2	8	6	3	1
Liczba tarcz ii)	1	2	3	10	3	1

Na poziomie istotności 0,1 zweryfikować hipotezę orzekającą, że wyniki strzelania mają rozkład dwumianowy.

Hipoteza zerowa jest taka że dane mają rozkład dwumianowy. Ponieważ nie podany został parametr p estymujemy go korzystając z następującego wzoru:

$$\hat{\mathrm{p}} = \frac{\overline{X}}{n}$$

Gdzie średnia jest obliczona ze wzoru na dane punktowe:

$$\overline{X} = \frac{\sum x_i \cdot n_i}{n}$$

Wyniki zapisano dla obu liczb tarcz, i dla sumy:

	i)	ii)	i+ii)
$\overline{X}$	2.65	2.75	2.7
$\hat{\mathrm{p}}$	0.53	0.55	0.54

Następnie, aby zastosować statystykę χ^2 skorzystaliśmy ze wzoru poniżej:

$$\chi_0^2 = \sum_{k=0}^{n} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$$

gdzie $p_i = P(X = k_i) \sim binom(6, \hat{\mathrm{p}})$, n_i to liczebność danego punktu, a n to całkowita liczebność próby. Poniżej przykładowe obliczenie p_i dla **i)**:

$$p_0 \overset{R}{=} dbinom(0, 5, 0.53) \approx 0.0229345$$

p_i			$n \cdot p_i$			χ^2		
i	ii	i + ii	i	ii	i + ii	i	ii	i + ii
0.022935	0.018453	0.020596	0.458690	0.369056	0.823852	0.458690	1.078670	0.037662
0.129312	0.112767	0.120891	2.586231	2.255344	4.835652	0.132883	0.028909	0.144410
0.291639	0.275653	0.283832	5.832776	5.513063	11.353271	0.805253	1.145549	0.010992
0.328869	0.336909	0.333194	6.577386	6.738188	13.327753	0.050685	1.578974	0.535792
0.185426	0.205889	0.195570	3.708526	4.117781	7.822812	0.135366	0.303424	0.424738
0.041820	0.050328	0.045917	0.836391	1.006569	1.836660	0.032004	0.000043	0.014526
SUM (χ_0^2)						1.614881	4.135569	1.168120

Aby obliczyć *p-value* potrzebujemy stopnie swobody które się wyznacza następująco:

$$\text{deg of freedom} = n - k - 1$$

Gdzie n to liczba przedziałów, k to liczba estymowanych parametrów, w tym przypadku 1. Zatem deg of freedom $= 6 - 1 - 1 = 4$. Możemy teraz obliczyć *p-value* następująco:

$$\text{p-value}_i = 1 - F_{\chi_4^2}(\chi_0^2) \overset{R}{=} pchisq(1.164881, 4, lower.tail = FALSE) \approx 0.8838462$$

$$\text{p-value}_{ii} = 1 - F_{\chi_4^2}(\chi_0^2) \overset{R}{=} pchisq(4.135569, 4, lower.tail = FALSE) \approx 0.3879694$$

$$\text{p-value}_{i+ii} = 1 - F_{\chi_4^2}(\chi_0^2) \overset{R}{=} pchisq(1.168120, 4, lower.tail = FALSE) \approx 0.883319$$

Widzimy że we wszystkich trzech przypadkach α jest mniejsze od *p-value*, zatem nie możemy odrzucić hipotezę zerową, co oznacza że dane mają rozkład dwumianowy.

32 Zadanie 8

przeprowadzono badanie wytrzymałości betonu na ściskanie. Uzyskane wyniki pomiarów (w N/cm^2) są podane w tabeli:

Wytrzymałość	Liczba próbek	
	i)	ii)
(1900 - 2000]	14	10
(2000 - 2100]	26	26
(2100 - 2200]	52	56
(2200 - 2300]	58	64
(2300 - 2400]	33	30
(2400 - 2500]	17	14

Na poziomie istotności 0,05 sprawdzić, czy wytrzymałość betonu na ściskanie

a) ma rozkład normalny;

b) ma rozkład $N(2200; \sigma)$;

c) ma rozkład $N(2200; 100)$.

32.1 a)

Aby sprawdzić czy dany układ ma rozkład normalny musimy najpierw estymować parametry m i σ jako:

$$m = \overline{X}$$
$$\sigma = \sqrt{\frac{1}{n-1}\sum(x_i - \overline{X})^2} = S$$

Jako wskaźnik przedziału (x_i) wzięto środek przedziałów i średnią obliczono ze wzoru:

$$\overline{X} = \frac{1}{n}\sum(n_i \cdot x_i)$$

Wariancję natomiast obliczono z następującego wzoru:

$$S^2 = \frac{\sum(x_i - \overline{X})^2 \cdot n_i}{n-1}$$

Uzyskano następujące wartości:

	i	ii
$\overline{X}$	2210.5	2210
S^2	17678.140704	15276.381910
S	132.959169	123.597661

Na podstawie tych parametrów możemy wyznaczyć statystykę χ^2 korzystając ze wzoru następującego:

$$\chi_0^2 = \sum_{k=0}^{n} \frac{(n_i - n \cdot p_i)^2}{n \cdot p_i}$$

Gdzie kolejne p_i wyznacza się za następująco:

$p_1 = F(2000) - F(1900)$

$\stackrel{R}{=} pnorm(2000, 2210.5, 132.959169) - pnorm(1900, 2210.5, 132.959169) \approx 0.046925$

$p_2 = F(2100) - F(2000)$

$\stackrel{R}{=} pnorm(2100, 2210.5, 132.959169) - pnorm(2000, 2210.5, 132.959169) \approx 0.146275$

...

p_i		$n \cdot p_i$		χ^2	
i	ii	i	ii	i	ii
0.046925	0.038585	9.384995	7.717072	2.269396	0.675355
0.146275	0.142083	29.254966	28.416658	0.362154	0.205522
0.265564	0.281021	53.112801	56.204115	0.023315	0.000741
0.281043	0.298987	56.208592	59.797456	0.057093	0.295353
0.173387	0.171138	34.677381	34.227697	0.081137	0.522192
0.062316	0.052637	12.463136	10.527342	1.651521	1.145527
SUM (χ_0^2)				4.444616	2.844690

Obszar krytyczny dla oby prób wyznaczymy następująco, ponieważ szukane są dwa parametry $k = 2$, natomiast $n = 6$.

$$R_{0.05} = (\chi^2_{1-0.05,6-2-1}, \infty)$$

$$\chi^2_{1-0.05,6-2-1} \stackrel{R}{=} qchisq(0.95, 3) \approx 7.814728$$

Otrzymane wartości zatem są dostatecznie małe że można stwierdzić że dane mają rozkład normalny.

32.2 b)

Obliczenia dokonują się analogicznie, natomiast zmienia się wartość k przy stopniach swobody rozkładu χ^2 i zamiast estymowana wartość oczekiwana podstawia się znaną już wartość oczekiwaną.

p_i		$n \cdot p_i$		χ^2	
i	ii	i	ii	i	ii
0.054237	0.045207	10.847456	9.041491	0.916209	0.101614
0.159730	0.156421	31.946013	31.284147	1.106713	0.892536
0.274008	0.290765	54.801545	58.152903	0.143220	0.079703
0.274008	0.290765	54.801545	58.152903	0.186676	0.587908
0.159730	0.156421	31.946013	31.284147	0.034774	0.052711
0.054237	0.045207	10.847456	9.041491	3.489647	2.719331
SUM (χ_0^2)				5.877238	4.433804

Obszar krytyczny wynosi natomiast:

$$\chi^2_{1-0.05,6-1-1} \overset{R}{=} qchisq(0.95, 4) \approx 9.487729$$

Zatem, jak poprzednio, wnioskujemy że dane mają rozkład normalny z $m = 2200$ ponieważ wartości χ_0^2 nie należą do obszaru krytycznego.

32.3 c)

Analogicznie do podpunktu b) podstawiamy znane wartości $m = 2200$ i $\sigma = 100$ aby obliczyć $n \cdot p_i$ do statystyki. Stopnie swobody statystyki zmieniają się ponownie, gdzie tym razem $k = 0$.

p_i		$n \cdot p_i$		χ^2	
i	ii	i	ii	i	ii
0.021400	0.021400	4.280047	4.280047	22.073939	7.644277
0.135905	0.135905	27.181024	27.181024	0.051316	0.051316
0.341345	0.341345	68.268949	68.268949	3.877000	2.204913
0.341345	0.341345	68.268949	68.268949	1.544645	0.266943
0.135905	0.135905	27.181024	27.181024	1.245740	0.292359
0.021400	0.021400	4.280047	4.280047	37.802673	22.073939
SUM (χ_0^2)				66.595313	32.533748

Obszar krytyczny wynosi natomiast:

$$\chi^2_{1-0.05,6-0-1} \overset{R}{=} qchisq(0.95, 5) \approx 11.0705$$

W tym przypadku wartości χ_0^2 należą do obszaru krytycznego, zatem musimy odrzucić hipotezę że dane mają rozkład $\sim N(2200, 100)$.

Zatem, z wniosków otrzymanych z poprzednich podpunktów, wnioskujemy że dane mają rozkład normalny z parametrem $m = 2200$, natomiast nie wiemy jakie jest odchylenie standardowe odpowiednie.

33 Zadanie 9

Dane z próby zostały pogrupowane w tabeli:

Przedział	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 9]	(9, 10]
liczba i) wyników ii)	52 33	38 36	20 19	12 14	7 9	6 8	5 4	5 6	4 5	1 4

Na poziomie istotności 0,02 zweryfikować hipotezę, że dane te pochodzą z rozkładu o gęstości $f(x)$ określonej wzorem (wyznaczyć a):

$$f(x) = \begin{cases} a(10-x) & \text{dla} \quad x \in [0,10] \\ 0 & \text{dla} \quad \text{w p. p.} \end{cases}$$

Wyznaczymy a z własności funkcji gęstości:

$$\begin{aligned} \int_{\mathbb{R}} f(x)dx &= 1 \\ &= \int_{\mathbb{R}} a(10-x)\mathbb{I}_{[0,10]}(x)dx \\ &= \int_0^{10} a(10-x)dx = a(10x - \frac{x^2}{2})\Big|_0^{10} \\ &= a(100-50) = 50a = 1 \\ a &= \frac{1}{50} = 0.02 \end{aligned}$$

Znając a możemy wyznaczyć dystrybuantę:

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} 0.02(10-t)\mathbb{I}_{[0,10]}(t)dt \\ &= \int_0^x 0.02(10-t)dt \\ &= 0.02(10t - \frac{t^2}{2})\Big|_0^x \\ &= 0.02(10x - \frac{x^2}{2}) \end{aligned}$$

$$F(x) = \begin{cases} 0 & , \quad x < 0 \\ 0.02(10x - \frac{x^2}{2}) & , \quad x \in [0,10] \\ 1 & , \quad x > 10 \end{cases}$$

Wyznaczymy dystrybuantę z próby korzystając ze wzoru poniżej i biorąc lewe końce każdego przedziału:

$$F_n(x|X) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{I}_{(-\infty,x]}(X_k), x = 1, 2, \ldots, 10$$

Poniżej przykładowe jedno obliczenie, wystarczy wziąć liczebność danego przedziału i poprzednich i podzielić przez całkowitą liczebność.

$$F(1|X) = \frac{52}{150} \approx 0.346667$$

$$F(2|X) = \frac{52+38}{150} = \frac{90}{150} = 0.6$$

$$\dots$$

Tak wyznaczono dystrybuantę dla i i ii. Wartości zapisano poniżej wraz z wartosciami dystrybuanty wyznaczonej na początku

x	i)	ii)	F(x)
1	0.346667	0.239130	0.19
2	0.600000	0.500000	0.36
3	0.733333	0.637681	0.51
4	0.813333	0.739130	0.64
5	0.860000	0.804348	0.75
6	0.900000	0.862319	0.84
7	0.933333	0.891304	0.91
8	0.966667	0.934783	0.96
9	0.993333	0.971014	0.99
10	1.000000	1.000000	1

	SUM n
i)	150
ii)	138

Zastosujemy test Kołmogorowa, zatem musimy wyznaczyć wartość $D_n = max\{|F_n(x|X) - F(x)|\}$, zatem obliczone zostały szukane różnice i wyznaczono wartości maksymalne wynoszące:

i) 0.24

ii) 0.14

Korzystając z tablic wyznaczono wartości krytyczne:

$$\text{i)} : \frac{1.51743}{\sqrt{150}} \approx 0.123898$$
$$\text{ii)} : \frac{1.51743}{\sqrt{138}} \approx 0.129172$$

Widzimy zatem że wartości te są większe od wartości krytycznych, zatem odrzucamy hipotezę zerową mówiąca że wartości z próby mają podana dystrybuantę

34 Zadanie 13

Wygenerować próby o liczebności 100 obserwacji według rozkładów:

i) $N(900; 50)$,

ii) $TR(725; 1075)$,

Następnie

a) obliczyć podstawowe statystyki,

b) sporządzić wykresy histfit, normplot, Q-Q,

c) przeprowadzić testy losowości,

d) przeprowadzić testy normalności,

e) przeprowadzić testy zgodności z innymi rozkładami,

f) przeprowadzić test zgodności dla wygenerowanych prób.

Dane losowe wygenerowane w R za pomocą funkcji poniżej. Zaokrąglono wartości do dwóch liczb po przecinku.

- **rnorm(100, 900, 50)**
- **rtri(725, 1075, (1075-725)/2 + 725)** (pakiet "EnvStats")

Wartości prób losowych podane pod zakładką **Dane**.

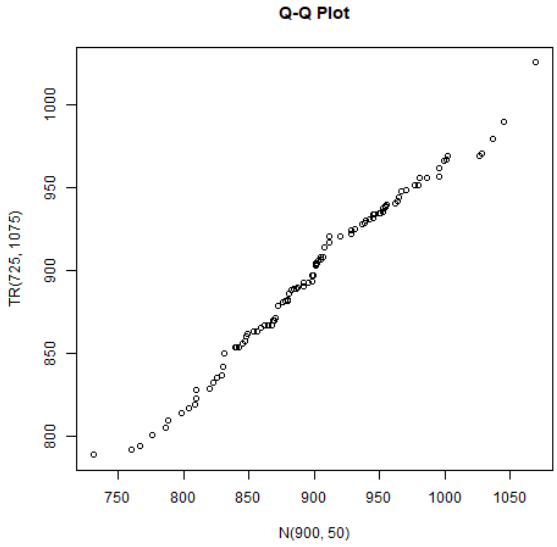
34.1 a)

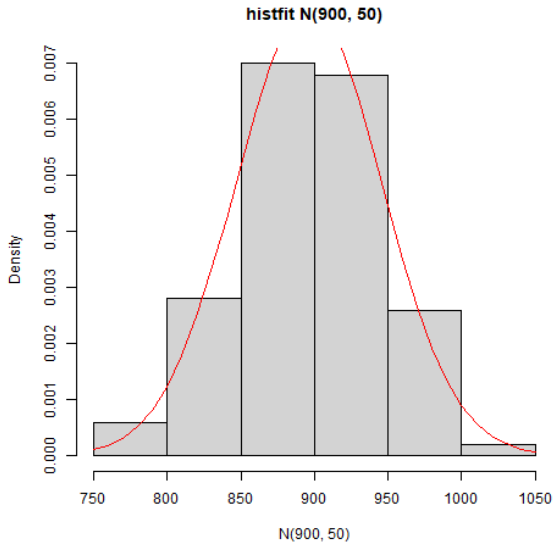
Korzystając z gotowych funkcji w R, obliczono średnią, wariancje i odchylenie standardowe. Wyniki zapisano poniżej.

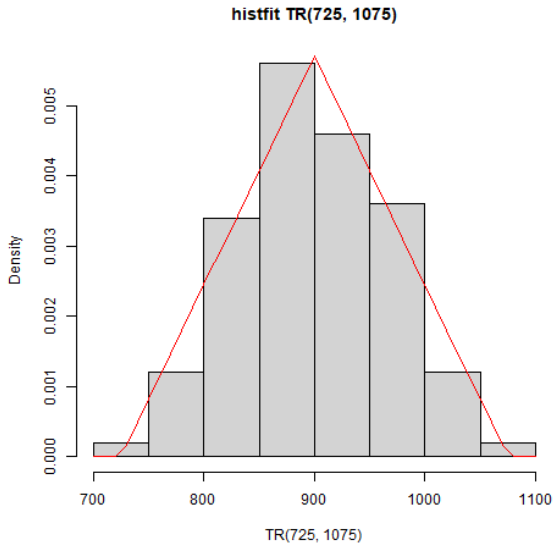
- **mean()** - średnia
- **var()** - wariancja
- **sqrt(var())** - odchylenie standardowe

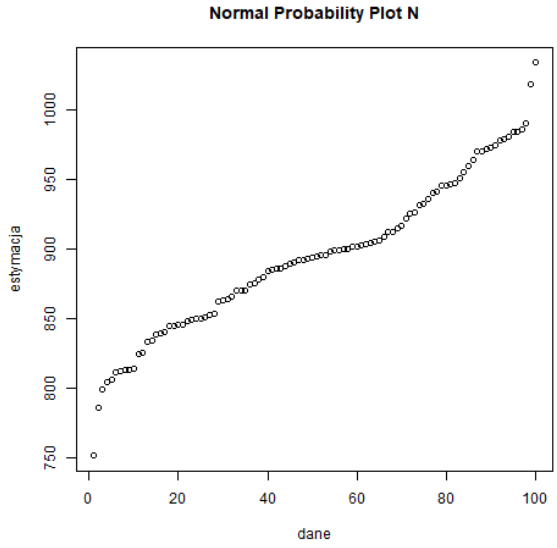
	i)	ii)
$\overline{X}$	896.18	899.5525
S^2	2485.676	4611.553

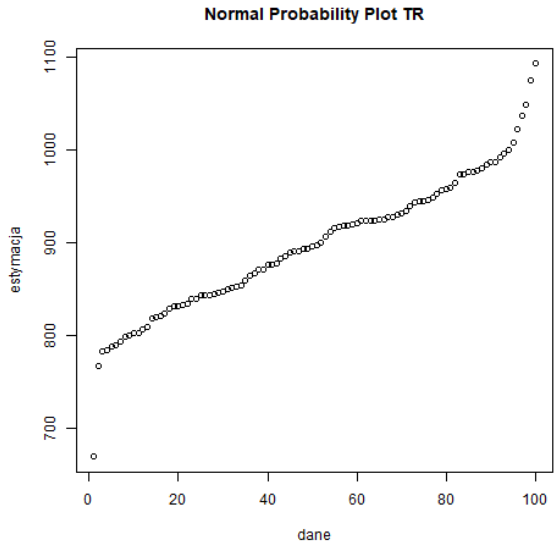
34.2 b)


Q-Q Plot
TR(725, 1075)
N(900, 50)
800
850
900
950
1000
750
800
850
900
950
1000
1050


histfit N(900, 50)
Density
N(900, 50)
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
750
800
850
900
950
1000
1050


histfit TR(725, 1075)
Density
0.000
0.001
0.002
0.003
0.004
0.005
700
800
900
1000
1100
TR(725, 1075)


Normal Probability Plot N
estymacja
750
800
850
900
950
1000
0
20
40
60
80
100
dane



34.3 c)

Aby zbadać losowość próby zastosujemy test Walda Wolfowitza. Wyznaczymy statystykę następująco:

$$Z = \frac{R - \mu_R}{\sigma_R}$$

Statystyka ta ma rozkład statystyki $\sim N(0,1)$. R jest liczbą serii która wyznaczamy jako ilość liczb mniejszych od mediany.

$$\begin{aligned}\mu_R &= \frac{2 \cdot n_1 \cdot n_2}{n_1 + n_2} + 1 \\ &= \frac{2 \cdot 50 \cdot 50}{50 + 50} + 1 \approx 51\end{aligned}$$

$$\begin{aligned}\sigma_R^2 &= \frac{2 \cdot n_1 \cdot n_2 \cdot (2 \cdot n_1 \cdot n_2 - n_1 - n_2)}{(n_1 + n_2)^2 (n_1 + n_2 - 1)} \\ &= \frac{2 \cdot 50 \cdot 50 (2 \cdot 50 \cdot 50 - 50 = 50)}{(50 + 50)^2 (50 + 50 - 1)} \approx 24.747475\end{aligned}$$

$$\sigma_R = \sqrt{\sigma_R^2} \approx 4.974683$$

Wtedy, ponieważ R jest równe dla oby danych i wynosi 50:

$$Z = \frac{50 - 51}{4.974683} \approx 0.201018$$

Obliczymy *p-value* dla prawostronnej i lewostronnej hipotezy o losowości:

$$\text{p-value}_1 \overset{R}{=} pnorm(0.201018, 0, 1) \approx 0.5796578$$

$$\text{p-value}_2 \overset{R}{=} 1 - pnorm(0.201018, 0, 1) \approx 0.4203422$$

W oby przypadkach nie możemy odrzucić hipotezę że wartości pochodzą z próby losowej.

34.4 d)

Test normalności został przeprowadzony za pomocą funkcji w R **ks.test(x,test)**, gdzie **test** jest dystrybuantą:

- $F_{N(900,50)}$ dla i)
- $F_{N(\overline{X},s)}$ dla ii)

Test jest dwustronny i oddaje wartości *p-value*, odpowiednio, 0.7774 dla **i)**, 0.8638 dla **ii)**. Zatem, przyjmując $\alpha = 0.05$ nie możemy odrzucić hipotezę o normalności. Wnioskujemy że oba rozkłady są normalne.
Ponieważ drugi rozkład pochodzi od rozkładu trójkątnego, możemy powiedzieć że tego typu rozkład jest zbliżony do normalnego.

34.5 e)

Nie przeprowadzono testu.

34.6 f)

Jak w podpunkcie **d**, zastosujemy test Kołmogorowa w R następująco: **ks.testx,y**. Gdzie x są dane pierwszej próby a y są dane drugiej próby. Orzymano następujący wynik:
Two-sample Kolmogorov-Smirnov test
data: x and y
D = 0.13, p-value = 0.3667
alternative hypothesis: two-sided

p-value jest większe od α, zatem wnioskujemy że rozkłady są do siebie podobne. Natomiast, z tabeli w **Tabele** weźmiemy wartość D_n dla n = 100, uzyskamy:

$$D_n = \frac{1.35810}{\sqrt{100}} = 0.13581$$

Zatem widzimy że rozkłady są bardzo blisko bycia rożnych, ponieważ gdy $D_n < D$ to odrzucamy hipotezę równania się rozkładów.

35 Zadanie 18

Wygenerować dużą próbę według jednego z rozkładów: beta, gamma, Weibulla lub logarytmiczno-normalnego i przekazać uzyskane dane drugiej osobie do identyfikacji rozkładu – nie informując o mechanizmie generowania. Dokonać oceny jakości dokonanej identyfikacji.

Uzyskane dane załadowano w R i, za pomocą funkcji w pakiecie "moments" obliczono współczynnik asymetrii.

$$\tilde{\mu}_3 = \frac{\mu_3}{\sigma^3} \stackrel{R}{=} skewness(data) \approx 3.454167$$

Jest to wartość dodatnia, zatem rozkład danych może być logarytmiczno-normalny, Weibulla lub gamma.

Aby przeprowadzić testy skorzystano z następującego skryptu w R, który oblicza dystrybuantę dla podanych danych, dokonuje "fitdistr" dla danego rozkładu testowego, oblicza wartość $D_{n,\alpha}$ na poziomie $\alpha = 0.05$ według tablicy w **Tablice** i oblicza $D_n = max\{|F_n(x) - F(x)|\}$.

```
library("moments")
library("MASS")
data = read.csv("data.csv")
x = sort(data[[1]])
D = 1.3581/sqrt(length(x))
cat("Skewness = ", skewness(x), "\ n")
# Dystrybuanta empiryczna
X = c()
s = 0
for(i in x) { s = length(which(x <= i))/length(x)
   X = c(X, s)
}

# Weibull
estimate = fitdistr(x,"weibull")
k = estimate[[1]][1]
lambda = estimate[[1]][2]
Y = pweibull(x, k, lambda)
d0 = max(abs(X-Y))
cat("D0 = ", d0, " D = ", D, "\n")

# gamma
estimate = fitdistr(x,"gamma")
alpha = estimate[[1]][1]
sigma = estimate[[1]][2]
Y = pgamma(x, alpha, rate = sigma)
d0 = max(abs(X-Y))
```

```
cat("D0 = ", d0, " D = ", D, "\ n")

#lognormal
estimate = fitdistr(x,"lognormal")
meanlog = estimate[[1]][1]
sdlog = estimate[[1]][2]
Y = plnorm(x, meanlog, sdlog)
d0 = max(abs(X-Y))
cat("D0 = ", d0, " D = ", D, "\n")
```

Wynik tego skryptu jest następujący:

- Skewness = 3.454167
- (Weibull) D0 = 0.06315938 D = 0.03036804
- (gamma) D0 = 0.07382558 D = 0.03036804
- (lognormal) D0 = 0.008872974 D = 0.03036804

Aby wiedzieć czy dane mają podaną rozkład porównujemy D z $D0$, jeżeli D jest większe od $D0$ to przyjmujemy że dane mają podany rozkład, w przeciwnym wypadku nie mają danego rozkładu. Widzimy zatem że dla Rozkładu Weibulla i gamma $D < D0$ zatem dane nie mają żadne z tych danych; natomiast widzimy że dla rozkładu logarytmiczno normalnego $D > D0$, zatem wnioskujemy że dane mają rozkład logarytmiczno normalny.

36 Tablice

Tablica wartości $D_{n,\alpha}$ testu Kołmogorowa.

n\α	0.001	0.01	0.02	0.05	0.1	0.15	0.2
1		0.99500	0.99000	0.97500	0.95000	0.92500	0.90000
2	0.97764	0.92930	0.90000	0.84189	0.77639	0.72614	0.68377
3	0.92063	0.82900	0.78456	0.70760	0.63604	0.59582	0.56481
4	0.85046	0.73421	0.68887	0.62394	0.56522	0.52476	0.49265
5	0.78137	0.66855	0.62718	0.56327	0.50945	0.47439	0.44697
6	0.72479	0.61660	0.57741	0.51926	0.46799	0.43526	0.41035
7	0.67930	0.57580	0.53844	0.48343	0.43607	0.40497	0.38145
8	0.64098	0.54180	0.50654	0.45427	0.40962	0.38062	0.35828
9	0.60846	0.51330	0.47960	0.43001	0.38746	0.36006	0.33907
10	0.58042	0.48895	0.45662	0.40925	0.36866	0.34250	0.32257
11	0.55588	0.46770	0.43670	0.39122	0.35242	0.32734	0.30826
12	0.53422	0.44905	0.41918	0.37543	0.33815	0.31408	0.29573
13	0.51490	0.43246	0.40362	0.36143	0.32548	0.30233	0.28466
14	0.49753	0.41760	0.38970	0.34890	0.31417	0.29181	0.27477
15	0.48182	0.40420	0.37713	0.33760	0.30397	0.28233	0.26585
16	0.46750	0.39200	0.36571	0.32733	0.29471	0.27372	0.25774
17	0.45440	0.38085	0.35528	0.31796	0.28627	0.26587	0.25035
18	0.44234	0.37063	0.34569	0.30936	0.27851	0.25867	0.24356
19	0.43119	0.36116	0.33685	0.30142	0.27135	0.25202	0.23731
20	0.42085	0.35240	0.32866	0.29407	0.26473	0.24587	0.23152
25	0.37843	0.31656	0.30349	0.26404	0.23767	0.22074	0.20786
30	0.34672	0.28988	0.27704	0.24170	0.21756	0.20207	0.19029
35	0.32187	0.26898	0.25649	0.22424	0.20184	0.18748	0.17655
40	0.30169	0.25188	0.23993	0.21017	0.18939	0.17610	0.16601
45	0.28482	0.23780	0.22621	0.19842	0.17881	0.16626	0.15673
50	0.27051	0.22585	0.21460	0.18845	0.16982	0.15790	0.14886
OVER 50	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$

37 Dane

Lp	x
1	788.93
2	935.01
3	835.42
4	860.39
5	878.83
6	849.94
7	908.35
8	968.99
9	870.11
10	892.63
11	897.37
12	920.72
13	966.1
14	866.96
15	832.1
16	855.79
17	804.91
18	853.83
19	947.65
20	955.98
21	800.77
22	934.35
23	931.69
24	870.11
25	888.78
26	871.46
27	935.5
28	904.21
29	842.15
30	866.94
31	885.97
32	897.12
33	882.85
34	979.81
35	941.73
36	828.26
37	853.63
38	865.62
39	937.56
40	890.47
41	794.4
42	810.03
43	881.28
44	814.12
45	892.84
46	938.28
47	1025.62
48	822.67
49	881.54
50	889.49
51	917.36
52	951.67
53	931.21
54	865.7
55	951.29
56	922.04
57	930.19
58	904.3
59	961.71
60	791.94
61	938.9
62	867.11
63	925.02
64	948.97
65	889.95
66	904.09
67	970.95
68	828.7
69	908.15
70	881.51
71	888.93
72	920.37
73	933.91
74	863.51
75	888.06
76	967.33
77	957
78	955.78
79	989.64
80	905.95
81	940.8
82	924.3
83	854.09
84	863.63
85	869.84
86	939.85
87	862.09
88	836.87
89	903.25
90	928.97
91	928.12
92	969.11
93	893.51
94	944.17
95	913.88
96	906.93
97	857.38
98	933.96
99	819.42
100	817.33

Lp	x
1	842.23
2	859.28
3	951.14
4	839.27
5	964.35
6	901.59
7	840.32
8	901.09
9	878.21
10	899.32
11	1069.14
12	809.81
13	882.93
14	892.01
15	919.61
16	939.53
17	804.38
18	731.17
19	998.84
20	979.77
21	831.48
22	901.58
23	954.37
24	879.23
25	856.04
26	942.74
27	946.08
28	898.02
29	995.29
30	819.69
31	868.23
32	867.35
33	766.67
34	787.84
35	945.2
36	1036.35
37	825.6
38	760.42
39	963.32
40	896.01
41	809.51
42	938.51
43	829.88
44	955.69
45	887.25
46	928.08
47	808.55
48	898.84
49	879.96
50	798.1
51	936.57
52	928.21
53	904.59
54	870.65
55	945.23
56	911.54
57	952.39
58	786.45
59	980.29
60	859.18
61	853.97
62	906.85
63	905.28
64	911.85
65	1000.95
66	864.62
67	823.08
68	1026.75
69	872.34
70	869.66
71	930.93
72	1028.37
73	1001.82
74	875.69
75	884.66
76	776.26
77	952.72
78	949.97
79	848.93
80	861.74
81	886.01
82	986.02
83	869.18
84	887.4
85	954.19
86	847.53
87	846.72
88	902.28
89	903.08
90	961.64
91	977.29
92	995.46
93	880.4
94	844.93
95	891.68
96	870.76
97	829.35
98	966.82
99	908.08
100	1045.02

Część VI
Laboratoria 12

Celem tych laboratoriów było zapoznanie się z metodami analizy regresji i korelacji. Uzyskaliśmy wiedze o wyznaczaniu prostej regresji dla danych oraz sprawdzenie czy dwa zestawy danych są skorelowane dodatnio lub ujemnie oraz w jakim stopniu. Nauczyliśmy się także wyznaczać błędy zastosowanego modelu.

38 Zadanie 1

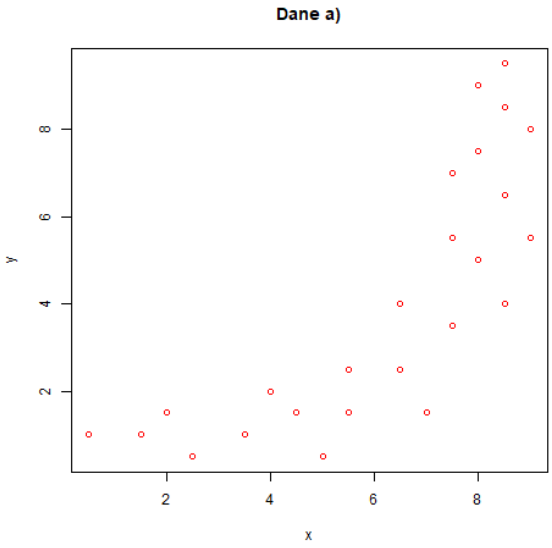
Sporządzić diagram rozrzutu, wyznaczyć oceny współczynników korelacji i determinacji, wyznaczyć równania prostych regresji (Y względem X, X względem Y), błędy standardowe estymacji oraz wykreślić równanie regresji dla podanych prób:

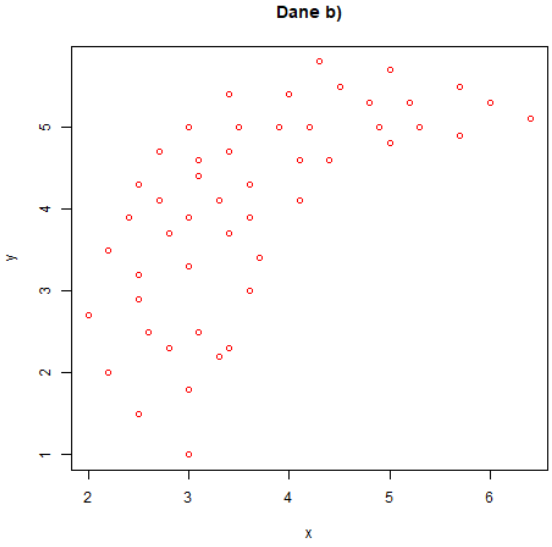
a) $[x; y] =$
$\{[5.5, 1.5], [8.5, 4.0], [4.0, 2.0], [8.0, 7.5], [2.5, 0.5], [8.0, 5.0], [8.5, 8.5], [3.5, 1.0],$
$[6.5, 2.5], [9.0, 8.0], [0.5, 1.0], [8.5, 6.5], [7.5, 3.5], [1.5, 1.0], [8.5, 9.5],$
$[2.0, 1.5], [8.0, 9.0], [7.5, 5.5], [9.0, 5.5], [7.0, 1.5], [7.5, 7.0], [5.0, 0.5],$
$[4.5, 1.5], [5.5, 2.5], [6.5, 4.0]\}$.

b) $[x; y] =$
$\{[3.4, 3.7], [2.7, 4.7], [4.4, 4.6], [2.6, 2.5], [5.2, 5.3], [3.1, 4.6], [2.2, 3.5], [3.3, 4.1],$
$[6.0, 5.3], [4.0, 5.4], [2.0, 2.7], [3.9, 5.0], [2.5, 1.5], [2.5, 4.3], [3.6, 3.0], [6.4, 5.1],$
$[2.8, 3.7], [4.3, 5.8], [5.7, 5.5], [2.5, 3.2], [4.9, 5.0], [3.0, 1.8], [3.6, 4.3], [5.7, 4.9],$
$[3.0, 1.0], [4.1, 4.1], [5.0, 4.8], [2.2, 2.0], [3.7, 3.4], [5.0, 5.7], [3.1, 4.4], [3.4, 5.4],$
$[3.4, 2.3], [2.5, 2.9], [5.3, 5.0], [4.1, 4.6], [3.0, 5.0], [2.8, 2.3], [3.0, 3.9], [2.4, 3.9],$
$[4.5, 5.5], [3.5, 5.0], [4.8, 5.3], [3.1, 2.5], [2.7, 4.1], [3.0, 3.3], [4.2, 5.0], [3.3, 2.2],$
$[3.6, 3.9], [3.4, 4.7]\}$.

38.1 Diagram rozrzutu

Jako pierwsze sporządzono diagram rozrzutu gdzie na osi X są wartości x, a na osi Y wartości y. Wykres sporządzono w R.


Dane a)
y
x
2
4
6
8


Dane b)
y
x
1
2
3
4
5
6

38.2 Współczynnik korelacji

Obliczymy teraz współczynnik korelacji zgodnie ze wzorem:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

Gdzie:

$$SS_{xx} = \sum(x_i^2) - \frac{(\sum x_i)^2}{n} \stackrel{R}{=} sum(x^2) - sum(x)^2/length(x)$$

$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n} \stackrel{R}{=} sum(x * y) - sum(x) * sum(y)/length(x)$$

Otrzymano następujące wartości:

	SS_{xx}	SS_{yy}	SS_{xy}	r
a)	155.64	209.74	142.44	0.7883711
b)	57.4248	74.1122	42.3684	0.6494526

Ponieważ oba r są dodatnie możemy sformułować hipotezę że współczynnik korelacji pomiędzy x i y jest dodatni.

H_0	$\rho \leqslant 0$
H_1	$\rho > 0$

Aby sprawdzić tę hipotezę zastosujemy następującą statystykę:

$$Z = (U - u_0) \cdot \sqrt{n - 3}$$

Która, dla $n > 7$ ma w przybliżeniu standardowy rozkład normalny. Poniżej przedstawiono obliczenia dla:

a)

$$U = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln(\frac{1.7883711}{0.2116289} \approx 1.067113$$

$$u_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2} = 0$$

$$Z_0 = 2.51587 \cdot \sqrt{10-3} \approx 5.005205$$

b)

$$U = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln(\frac{1.6494526}{0.3505474} \approx 0.7743514$$

$$u_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2} = 0$$

$$Z_0 = 2.51587 \cdot \sqrt{10-3} \approx 5.308686$$

Wtedy można obliczyć *p-value* dla oby prób, zgodnie ze wzorem:

$$\text{p-value}_a = 1 - \Phi(5.005205) \overset{R}{=} 1 - pnorm(5.005205, 0, 1) \approx 2.790134e - 07$$
$$\text{p-value}_b = 1 - \Phi(5.308686) \overset{R}{=} 1 - pnorm(5.308686, 0, 1) \approx 5.520921e - 08$$

Przyjmując $\alpha = 0.05$ oba *p-value* są mniejsze od α; zatem odrzucamy hipotezę zerową i wnioskujemy że korelacja pomiędzy x i y jest typu dodatniego, co widać na wykresach i z obliczonych wartości r.

38.3 Współczynnik determinacji i równania regresji

Aby wyznaczyć współczynnik determinacji potrzebne jest wyliczenie SSE, które wyraża się następująco:

$$\text{SSE} = \sum (y_i - \hat{\text{y}}_i)^2$$

lub

$$\text{SSE} = \sum (x_i - \hat{\text{x}}_i)^2$$

Zatem potrzebujemy najpierw wyznaczyć równanie regresji. Do tego równania potrzebujemy dwa współczynniki:

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \overset{R}{=} mean(y) - b1 * mean(x)$$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = b1$$

Natomiast dla X zależne od Y parametry są następujące:

$$\beta_0 = \overline{x} - \beta_1 \overline{y} \overset{R}{=} mean(x) - b1 * mean(y)$$

$$\beta_1 = \frac{SS_{xy}}{SS_{yy}} = b1$$

Rozważymy najpierw Y zależne od X. Równania wyglądają następująco:

$$y = \beta_0 + \beta_1 \cdot x$$
$$y_a = -1.580956 + 0.9151889 \cdot x$$
$$y_b = 1.342481 + 0.7378067 \cdot x$$

Wtedy parametry SSE wynoszą:

	SSE
a)	79.38049
b)	42.85251

Możemy teraz wyznaczyć współczynnik determinacji, który wynosi:

$$r^2 = 1 - \frac{SSE}{SS_{yy}}$$

	r^2
a)	0.6215291
b)	0.4217887

Następnie rozważymy dla X zależnego od Y. Wtedy, analogicznie do wcześniej:

$$x = \beta_0 + \beta_1 \cdot y$$
$$x_a = 3.389911 + 0.6791265 \cdot y$$
$$x_b = 1.341846 + 0.5716792 \cdot y$$

Wtedy parametry SSE wynoszą:

	SSE
a)	58.90522
b)	33.20367

Możemy teraz wyznaczyć współczynnik determinacji, który wynosi:

$$r^2 = 1 - \frac{SSE}{SS_{yy}}$$

	r^2
a)	0.6215291
b)	0.4217887

Wartości r^2 nie zmieniają się, zatem, równanie regresji zmniejsza całkowitą sumę kwadratów o 62% dla próby a) i 42% dla próby b) od średniej arytmetycznej.

38.4 Błąd modelu

Następnie obliczymy błędy modelu zgodnie ze wzorem:

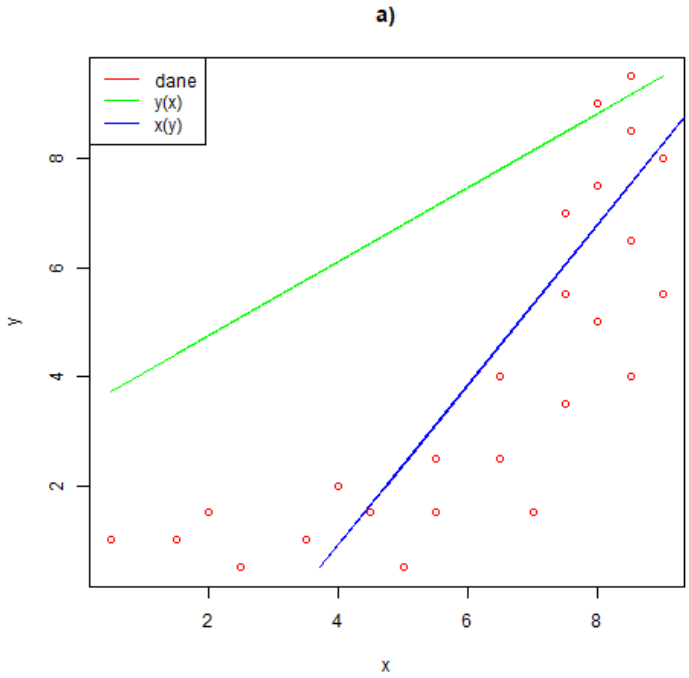
$$S^2 = \frac{SSE}{n-2}$$

Liczebności prób są, dla a) $n = 25$, dla b) $n = 50$. Wtedy błędy modelu są następujące:

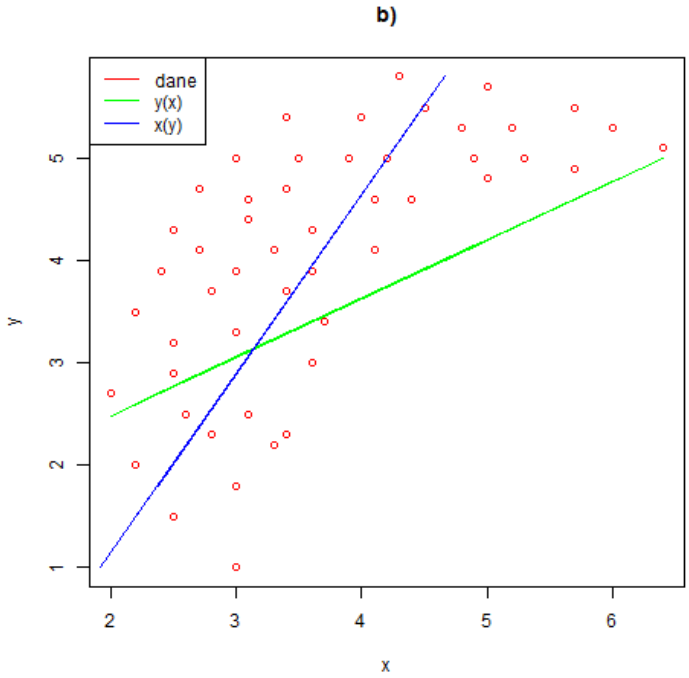
	Y od X	X od Y
a)	3.451326	2.561096
b)	0.8927607	0.6917431

38.5 Wykresy regresji

Jako ostatnie przedstawiono wykresy prostych regresji na wykresach z danymi.



Dla danych a) prosta x(y) lepiej obrazuje przebieg danych, natomiast dla danych b) nie ma znacznej różnicy pomiędzy prostymi.



39 Zadanie 2

Odnotowano miesięczne dochody przypadające na jednego członka rodziny (w zł) - cecha X oraz wyrażoną w procentach część budżetu rodzinnego przeznaczoną na zakup artykułów żywnościowych i utrzymanie mieszkania - cecha Y.

X	200	300	150	225	175	350	150	250	325	250
Y	70	80	95	75	90	60	60	65	85	90

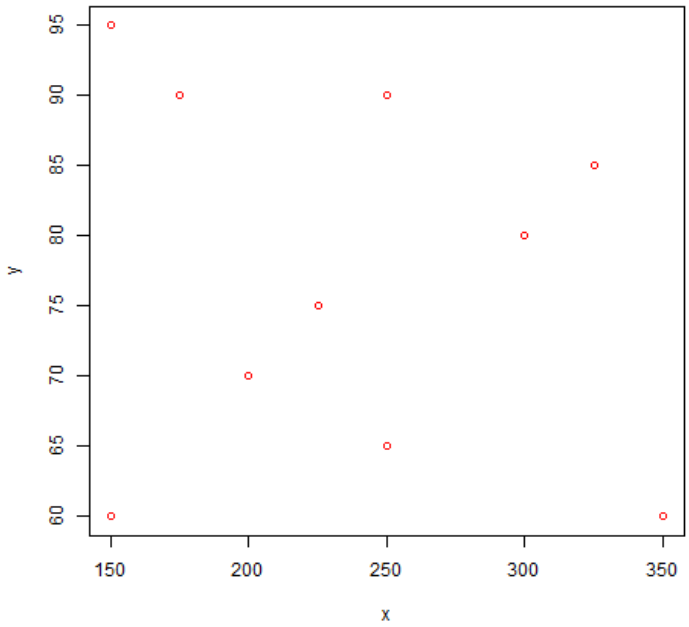
Sporządzić diagram rozrzutu, wyznaczyć oceny współczynników korelacji i determinacji między dochodem przypadającym na jednego członka rodziny a wydatkami na artykuły żywnościowe i utrzymanie mieszkania.

Wykres sporządzono w R wygląda następująco:
Obliczymy teraz współczynnik korelacji zgodnie ze wzorem:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx} \cdot SS_{yy}}}$$

Gdzie:

$$SS_{xx} = \sum(x_i^2) - \frac{(\sum x_i)^2}{n} \overset{R}{=} sum(x^2) - sum(x)^2/length(x)$$



$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \cdot \sum y_i}{n} \overset{R}{=} sum(x * y) - sum(x) * sum(y)/length(x)$$

Otrzymano następujące wartości:

SS_{xx}	SS_{yy}	SS_{xy}	r
45312.5	1510	-1625	-0.1964517

Aby sprawdzić ten współczynnik to wyznaczymy hipotezę zerową orzekającą, że istnieje dodatnia korelacja:

H_0	$\rho \geqslant 0$
H_1	$\rho < 0$

Skorzystamy ze statystki testowej:

$$Z = (U - u_0) \cdot \sqrt{n-3}$$

Która ma w przybliżeniu standardowy rozkład normalny.

$$U = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln(\frac{0.8035483}{1.1964517} \approx -0.199039$$

$$u_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2} = 0$$

$$Z_0 = -0.199039 \cdot \sqrt{10-3} \approx -0.526608$$

Obliczymy teraz *p-value* która wynosi:

$$\text{p-value} = \Phi(-0.526608) \overset{R}{=} pnorm(-0.526608, 0, 1) \approx 0.2992329$$

Wartość ta jest większa niż większość standardowo przyjętych α zatem nie możemy odrzucić hipotezę zerową więc nie istnieje ujemna korelacja między cechą X i Y.

Aby wyznaczyć współczynnik determinacji potrzebne jest wyliczenie SSE, które wyraża się następująco:

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

Zatem potrzebujemy najpierw wyznaczyć równanie regresji. Do tego równania potrzebujemy dwa współczynniki:

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \overset{R}{=} mean(y) - b1 * mean(x) \approx 85.51724$$

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}} = b1 \approx -0.03586207$$

Wtedy podstawiając kolejne wartosci x_i do równania regresji możemy obliczyć SSE:

$$\hat{y}_i = \beta_0 + \beta_1 \cdot x_i = 85.51724 - 0.03586207 \cdot x_i$$

Wtedy $SSE = 1451.724$ i można obliczyć współczynnik determinacji:

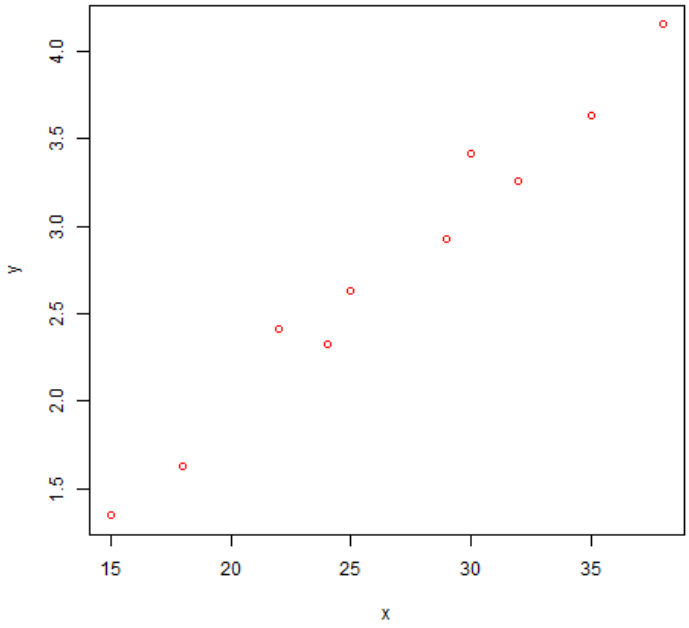
$$r^2 = 1 - \frac{SSE}{SS_{yy}} = 1 - \frac{1451.724}{1510} \approx 0.03859329$$

Zatem równanie regresji zmniejsza całkowitą sumę kwadratów próby o 3% od średniej arytmetycznej.

40 Zadanie 3

Naturalne jest przekonanie, że powinna być silna korelacja pomiędzy miesięcznymi obrotami firmy a jej liczebnością personelu handlowego. Dla pewnej firmy zostały zebrane dane dotyczące liczby sprzedawców w ostatnich 10 kwartałach oraz osiągane średniomiesięczne obroty (w mln zł) w tym czasie.
Wynoszą one: [15, 1.35], [18, 1.63], [24, 2.33], [22, 2.41], [25, 2.63], [29, 2.93], [30, 3.41], [32, 3.26], [35, 3.63], [38, 4.15].
Sprawdzić, czy to przekonanie potwierdziło się dla badanej firmy.

Sporządzono wykres rozrzutu w celu wstępnego sprawdzenia danych; wykres przygotowany w R.



Z wykresu widać że może istnieć dodatnia korelacja pomiędzy danymi. x oznacza liczbę pracowników, natomiast y oznacza średnio-miesięczne obroty.
Następnie został obliczony współczynnik korelacji zgodnie ze wzorami podanymi w poprzednim zadaniu. Uzyskano następujące wartości:

SS_{xx}	SS_{yy}	SS_{xy}	r
485.6	6.97801	57.456	0.9870298

Wartość ta jest dodatnia, zatem jest możliwe że korelacja jest typu dodatniego. W celu sprawdzenia tego sporządzona została teza alternatywna i zerowa

wyznaczona poniżej. ρ jest rzeczywistym współczynnikiem korelacji.

H_0	$\rho \leqslant 0$
H_1	$\rho > 0$

Ponieważ liczba obserwacji $n = 10 > 7$ możemy zastosować statystykę jak w poprzednim zadaniu.

$$Z = (U - u_0) \cdot \sqrt{n-3}$$

$$U = \frac{1}{2} \ln \frac{1+r}{1-r} = \frac{1}{2} \ln(\frac{1.9870298}{0.0129702} \approx 2.51587$$

$$u_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0} + \frac{\rho_0}{2n-2} = 0$$

$$Z_0 = 2.51587 \cdot \sqrt{10-3} \approx 6.656366$$

Obliczymy teraz *p-value* zgodnie ze wzorem:

$$\text{p-value} = 1 - \Phi(6.656366) \overset{R}{=} 1 - pnorm(6.656366, 0, 1) \approx 1.4034e-11$$

Przyjmując $\alpha = 0.05$ widzimy że *p-value* jest od tej wartości mniejsze; zatem możemy odrzucić hipotezę zerową i wnioskować że istnieje dodatnia korelacja między liczbą personelu i miesięcznymi obrotami.

Część VII
Laboratoria 13

Celem tych laboratoriów było zapoznanie się z metodami analizy wariancji czyli testów ANOVA dla zależności od jednego parametru dla danych normalnych i w postaci blokowej. Testy tego typu zostały przeprowadzone za pomocą programów komputerowych, zatem uzyskaliśmy wiedzę jak te testy przeprowadzać ze wspomaganiem oraz jak dane przygotować żeby program mógł je prawidłowo odczytać.

41 Zadanie 1

(Krysicki 5.2). Zmierzono długości czasów świecenia trzech typów żarówek, otrzymując (w h):

dla typu 1: 1802, 1992, 1854, 1880, 1761, 1900;

dla typu 2: 1664, 1755, 1823, 1862;

dla typu 3: 1877, 1710, 1882, 1720, 1950.

Na poziomie istotności $\alpha = 0,05$ zweryfikować hipotezę, że wartości przeciętne czasów świecenia żarówek tych typów są jednakowe.

Dane przygotowano w postaci pliku csv w celu obliczenia testu ANOVA w R.

data	**type**
1802	"1"
1992	"1"
1854	"1"
1880	"1"
1761	"1"
1900	"1"
1664	"2"
1755	"2"
1823	"2"
1862	"2"
1877	"3"
1710	"3"
1882	"3"
1720	"3"
1950	"3"

Dane te wgrano w R w następujący sposób: *data = read.csv("w12zad1.csv", colClasses = c("numeric", "factor").*

Obliczenie testu ANOVA w R przeprowadzono w następujący sposób:
model = aov(data∼type, data)
summary(model)
Otrzymano następujący tablicowy wynik:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	2	18947	9473	1.127	0.356
Residuals	12	100864	8405		

Funkcja oblicza *p-value* zapisane w ostatniej kolumnie, które jest większe od przyjętego poziomu istotności $\alpha = 0.05$. Zatem nie mamy podstaw aby odrzucić hipotezę zerowa o równaniu się wartości przeciętnych czasów świecenia żarówek.

42 Zadane 2

(Krysicki 5.3). Spośród trzech odmian ziemniaków każdą uprawiano na 12 działkach tej samej wielkości i rodzaju. Działki te podzielono na 4 grupy po 3 działki i dla każdej grupy zastosowano różny rodzaj nawozu. Plony w q zestawione w tabeli:

Odmiana	Nawóz											
	1			2			3			4		
1	5,6	6,1	5,9	6,6	6,7	6,6	7,7	7,3	7,4	6,3	6,4	6,3
2	5,7	4,9	5,1	6,5	6,7	6,6	6,9	7,1	6,5	6,6	6,7	6,7
3	6,3	6,1	6,3	6,5	6,4	6,2	6,6	6,6	6,8	6,3	6,1	6,0

Na poziomie istotności $\alpha = 0,05$ zweryfikować następujące hipotezy:

a) wartości przeciętne plonów dla różnych odmian nie różnią się istotnie niezależnie od stosowanego nawozu,

b) wartości przeciętne plonów dla różnych nawozów nie różnią się istotnie niezależnie od odmiany,

c) interakcja między odmianami i nawozami jest równa 0.

Dane przygotowano w postaci pliku csv w celu obliczenia testu ANOVA w R.

data	odmiana	nawoz
5.6	"1"	"1"
6.1	"1"	"1"
5.9	"1"	"1"
5.7	"1"	"2"
4.9	"1"	"2"
5.1	"1"	"2"
6.3	"1"	"3"
6.1	"1"	"3"
6.3	"1"	"3"
6.6	"1"	"4"
6.7	"1"	"4"
6.6	"1"	"4"
6.5	"2"	"1"
6.7	"2"	"1"
6.6	"2"	"1"
6.5	"2"	"2"
6.4	"2"	"2"
6.2	"2"	"2"
7.7	"2"	"3"
7.3	"2"	"3"
7.4	"2"	"3"
6.9	"2"	"4"
7.1	"2"	"4"
6.5	"2"	"4"
6.6	"3"	"1"
6.6	"3"	"1"
6.8	"3"	"1"
6.3	"3"	"2"
6.4	"3"	"2"
6.3	"3"	"2"
6.6	"3"	"3"
6.7	"3"	"3"
6.7	"3"	"3"
6.3	"3"	"4"
6.1	"3"	"4"
6.0	"3"	"4"

42.1 a)

Dane te wgrano w R w następujący sposób: *data = read.csv("w12zad2.csv", colClasses = c("numeric", "factor", "factor").*
Obliczenie testu ANOVA w R przeprowadzono w następujący sposób:
model = aov(data∼odmiana + nawoz, data)
summary(model)
Otrzymano następujący tablicowy wynik:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
odmiana	2	4.101	2.0503	16.900	1.21e-05
nawoz	3	3.116	1.0388	8.563	0.000294
Residuals	30	3.639	0.1213		

Funkcja oblicza *p-value* w ostatniej kolumnie. Dla typu nawozu widzimy że wartość 0.000294 jest mniejsza od przyjętego poziomu istotności $\alpha = 0.05$, zatem wartości przeciętne plonów różnią się istotnie zależnie od stosowanego nawozu.

42.2 b)

Korzystając z tabeli poprzedniego podpunktu widzimy że *p-value = 1.12e-05* dla typu odmiany jest mniejsze od przyjętego poziomu istotności $\alpha = 0.05$. Zatem wartości przeciętne plonów dla rożnych nawozów różnią się istotnie zależnie od odmiany.

42.3 c)

Aby sprawdzić interakcje między nawozami i odmianami możemy zastosować test Tukeya w R następująco: *TukeyHSD(model, conf.level = 0.95)*. Funkcja ta oddaje następujące tablice:

$odmiana				
	diff	lwr	upr	p adj
2-1	0.8250000	0.4744532	1.17554680	0.0000071
3-1	0.4583333	0.1077865	0.80888014	0.0083242
3-2	-0.3666667	-0.7172135	-0.01611986	0.0388934
$nawoz				
2-1	-0.4000000	-0.84645464	0.04645464	0.0917316
3-1	0.4111111	-0.03534353	0.85756575	0.0796887
4-1	0.1555556	-0.29089908	0.60201019	0.7797038
3-2	0.8111111	0.36465647	1.25756575	0.0001551
4-2	0.5555556	0.10910092	1.00201019	0.0102547
4-3	-0.2555556	-0.70201019	0.19089908	0.4179349

Obliczone są *p-value* zatem możemy dokonać wnioski. Dla różnicy odmian widzimy że wszystkie *p-value* są mniejsze od $\alpha = 0.05$, zatem każda wartość przeciętna dla odmian różni się, więc różnice nie są równe 0.
Dla nawozów natomiast istnieją wartości *p-value* większe od $\alpha = 0.05$ są to różnice: 1-3, 1-4, 3-4. Oznacza to że dla tych nawozów różnica wartości przeciętnych jest z dużym prawdopodobieństwem 0. Natomiast dla reszty nawozów różnice nie są równe 0 bo przeciętne wartości różnią się za dużo.

43 Zadanie 3

Rozwiązać zadanie 5.7 z Krysickiego.

Z trzech różnych wydziałów pewnej uczelni wylosowano po pięciu studentów z każdego roku studiów i obliczono średnią ocen uzyskaną przez każdego studenta w ostatnim semestrze. Uzyskano rezultaty

Rok studiów	Wydział											
	A				B				C			
I	2.6	4.1	3.1	2.4	3.1	2.5	3.3	3.8	2.7	4.2	2.9	3.7
II	2.8	4.3	3.8	3.0	3.9	2.6	3.2	3.3	3.0	4.4	3.9	3.1
III	3.2	4.1	4.8	4.0	3.4	2.9	4.1	2.8	4.0	3.3	3.4	3.0
IV	3.2	3.9	4.2	3.6	3.6	4.4	2.8	3.9	3.7	5.0	2.6	3.4
V	4.0	4.0	3.5	3.8	4.0	3.0	4.5	3.7	3.0	3.8	4.8	3.5

Zakładając, że średnie uzyskiwanych ocen mają rozkłady normalne o tej samej wariancji na poziomie $\alpha = 0.05$, zweryfikować następujące hipotezy:

a) wartości przeciętne średnich ocen dla studentów różnych wydziałów są jednakowe;

b) wartości przeciętne średnich ocen dla różnych lat studiów są jednakowe;

c) wartości przeciętne ocen średnich dla pierwszych dwóch lat są jednakowe;

Wartości z tabeli przepisano do pliku csv w celu wgrania go do programu R.

data	wydzial	rok
2.6	"A"	"I"
4.1	"A"	"I"
3.1	"A"	"I"
2.4	"A"	"I"
3.1	"B"	"I"
2.5	"B"	"I"
3.3	"B"	"I"
3.8	"B"	"I"
2.7	"C"	"I"
4.2	"C"	"I"
2.9	"C"	"I"
3.7	"C"	"I"
2.8	"A"	"II"
4.3	"A"	"II"
3.8	"A"	"II"
3.0	"A"	"II"
3.9	"B"	"II"
2.6	"B"	"II"
3.2	"B"	"II"
3.3	"B"	"II"
3.0	"C"	"II"
4.4	"C"	"II"
3.9	"C"	"II"
3.1	"C"	"II"
3.2	"A"	"III"
4.1	"A"	"III"
4.8	"A"	"III"
4.0	"A"	"III"
3.4	"B"	"III"
2.9	"B"	"III"
4.1	"B"	"III"
2.8	"B"	"III"
4.0	"C"	"III"
3.3	"C"	"III"
3.4	"C"	"III"
3.0	"C"	"III"
3.2	"A"	"IV"
3.9	"A"	"IV"
4.2	"A"	"IV"
3.6	"A"	"IV"
3.6	"B"	"IV"
4.4	"B"	"IV"
2.8	"B"	"IV"
3.9	"B"	"IV"
3.7	"C"	"IV"
5.0	"C"	"IV"
2.6	"C"	"IV"
3.4	"C"	"IV"
4.0	"A"	"V"
4.0	"A"	"V"
3.5	"A"	"V"
3.8	"A"	"V"
4.0	"B"	"V"
3.0	"B"	"V"
4.5	"B"	"V"
3.7	"B"	"V"
3.0	"C"	"V"
3.8	"C"	"V"
4.8	"C"	"V"
3.5	"C"	"V"

43.1 a)

Dane z pliku csv wgrano pod zmienną "data" w następujący sposób: *data = read.csv("w12zad3.csv", colClasses = c("numeric", "factor", "factor").*

Test ANOVA natomiast wykonano w następujący sposób: *model = aov(data∼wydzial, data)*. Otrzymano następujący wynik *summary(model) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wydzial	2	0.345	0.1727	0.437	0.648
Residuals	57	22.502	0.3948		

Ponieważ *p-value*, obliczone w ostatniej kolumnie, jest większe niż $\alpha = 0.05$ wnioskujemy że wartości przeciętne średnich ocen dla studentów różnych wydziałów są sobie równe.

43.2 b)

Test ANOVA przeprowadzono podobnie jak w poprzednim podpunkcie, tj: *model = aov(data∼rok, data)*. Otrzymano następujące wyniki *summary(model) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rok	4	2.612	0.6531	1.775	0.147
Residuals	55	20.235	0.3679		

Ponieważ *p-value*, obliczone w ostatniej kolumnie, jest większe niż $\alpha = 0.05$ wnioskujemy że wartości przeciętne średnich ocen dla studentów różnych lat studiów są sobie równe.

43.3 c)

Ponieważ zakładamy że dane mają rozkład normalny możemy zastosować test Tukeya w następujący sposób: *tukey = TukeyHSD(model, conf.level = 0.95)* Otrzymano następujący wynik

	diff	lwr	upr	p adj
II-I	0.2416667	-0.45671717	0.9400505	0.8648729
III-I	0.3833333	-0.31505051	1.0817172	0.5364514
IV-I	0.4916667	-0.20671717	1.1900505	0.2865813
V-I	0.6000000	-0.09838384	1.2983838	0.1244862
III-II	0.1416667	-0.55671717	0.8400505	0.9785925
IV-II	0.2500000	-0.44838384	0.9483838	0.8498736
V-II	0.3583333	-0.34005051	1.0567172	0.6004999
IV-III	0.1083333	-0.59005051	0.8067172	0.9921775
V-III	0.2166667	-0.48171717	0.9150505	0.9048593
V-IV	0.1083333	-0.59005051	0.8067172	0.9921775

Dla tego typu testu dostajemy także *p-value* i widzimy że każda ta wartość jest większa od $\alpha = 0.05$. Zatem wnioskujemy że wartości średnie dla pierwszych dwóch lat są jednakowe.

44 Zadanie 4

Korzystając ze wspomagania komputerowego rozwiązać przykład 3 z wykładu.

Jednym z aspektów jakości samochodów osobowych jest koszt naprawy uszkodzeń spowodowanych drobnymi ulicznymi stłuczkami. Decydujące znaczenie mają tu zderzaki. Producent rozważa wprowadzenie nowego typu zderzaków spośród czterech zaprojektowanych typów. Zainstalowano po siedem zderzaków każdego typu na pojazdach popularnej klasy i poddano je próbom zderzania ze ścianą z prędkością 30 km/h. Następnie oszacowano koszty napraw powstałych uszkodzeń (w j.m.). Wyniki są przedstawione w tablicy.

Typ zderzaka			
1	2	3	4
315	285	269	255
288	292	277	287
293	263	273	265
306	249	252	279
299	275	263	241
310	266	251	312
282	252	272	310

a) Przyjmując 5-procentowy poziom istotności zbadać, czy są istotne różnice w kosztach usuwania uszkodzeń dla badanych czterech typów zderzaków.

b) W przypadku występowania różnic ustalić typy zderzaków różniących się ze względu na koszty usuwania awarii.

44.1 a)

Wartości z tablicy wybrano do pliku csv aby wczytać je w R. Wartości zapisano w kolumnie nazwaną "data" a odpowiadające wartościom typy zderzaka zapisano pod kolumną "type" i uwzględniono w R że to ma być kolumna typu "factor". Dane wgrano korzystając z funkcji *data = read.csv("w12zad4.csv").*
Aby sprawdzić czy są istotne różnice w kosztach usuwania uszkodzeń wykorzystano następującą funkcje: *model = aov(data∼type, data=data)* która wykonuje test ANOVA jedno kierunkowy na zależność kosztu od typu zderzaka. Poniżej uzyskane tą funkcją wyniki (*summary(model)*).

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
type	3	4805	1601.6	5.197	0.00658
Residuals	24	7396	308.2		

Funkcja ta oddaje *p-value* zapisane pod ostatnią kolumną, zatem, przyjmując $\alpha = 0.05$ stwierdzamy że typ zderzaka ma wpływ na koszt usuwania uszkodzeń.

44.2 b)

Ponieważ w podpunkcie **a)** okazało się że występują różnice w kosztach usuwania uszkodzeń więc możemy poszukać które typy różnią się od siebie. W R istnieje funkcja obliczająca test Tukeya na podstawie wcześniej wyznaczonej analizy wariancji (*aov()*). Ten test został wywołany następująco: *TukeyHSD(model, conf.level = 0.95)*.
Otrzymano następującą tabele:

	diff	lwr	upr	p adj
2-1	-30.142857	-56.02790	-4.257812	0.0182410
3-1	-33.714286	-59.59933	-7.829241	0.0074533
4-1	-20.571429	-46.45647	5.313616	0.1540625
3-2	-3.571429	-29.45647	22.313616	0.9807839
4-2	9.571429	-16.31362	35.456474	0.7394841
4-3	13.142857	-12.74219	39.027902	0.5111022

Funkcja ta także zwraca *p-value* zatem, porównując z $\alpha = 0.05$ widzimy że jedynie typ 2 i typ 1 się różnią a także typ 1 i typ 3.
Obliczenia programem potwierdzają wyniki obliczone w przykładzie z wykładu.

45 Zadanie 6

Korzystając ze wspomagania komputerowego rozwiązać przykład 6 z wykładu.

Wycena prywatyzowanego przedsiębiorstwa państwowego poprzedzona jest szczegółową analizą wartości majątku, potencjału produkcyjnego, możliwości przestawienia produkcji, sposobów zabezpieczenia socjalnego pracowników, itp. Szacowanie wartości majątku przeprowadzają specjalistyczne firmy zajmujące się wyceną. Przeszacowanie wartości przedsiębiorstwa zmniejsza szanse prywatyzacji firmy, natomiast zaniżenie wartości zmniejsza przychód z prywatyzacji.

W celu zmniejszenia ryzyka popełnienia błędu przedstawiciel odpowiedniego ministerstwa zamierza porównać średnie oszacowania wartości trzech niezależnych firm wyceniających majątek zanim zleci jednej z nich dokonanie oszacowania wartości rynkowej prywatyzowanego przedsiębiorstwa.

Przedstawiciel zebrał informacje o wycenie majątku tych samych czterech przedsiębiorstw przez każdą z rozważanych trzech firm wyceniających. Uzyskane dane o wycenach (w mln zł) są podane w tabeli.

Firma	Wycena przedsiębiorstwa			
wyceniająca	1	2	3	4
A	4.6	6.2	5.0	6.6
B	4.9	6.3	5.4	6.8
C	4.4	5.9	5.4	6.3

a) Przeprowadzić analizę wariancji dla przeprowadzonych wycen. Na poziomie istotności $\alpha = 0,05$ sprawdzić, czy są istotne różnice między oczekiwanymi wycenami dla zabiegów i bloków.

b) Wyznaczyć 90-procentowy przedział ufności dla różnic między oczekiwanymi wycenami dla firm wyceniających A i B.

Wartości z tabeli zapisano w pliku csv następującej postaci tak aby można było dokonać obliczenia w R.

data	**firma**	**wycena**
4.6	"A"	"1"
6.2	"A"	"2"
5.0	"A"	"3"
6.6	"A"	"4"
4.9	"B"	"1"
6.3	"B"	"2"
5.4	"B"	"3"
6.8	"B"	"4"
4.4	"C"	"1"
5.9	"C"	"2"
5.4	"C"	"3"
6.3	"C"	"4"

45.1 a)

Jak w poprzednim zadaniu wykorzystamy funkcję R-owską *model = aov(data∼firma+wycena)*. Wynik tej funkcji można odczytać za pomocą *summary(model)*.

	Df	Sum Sq	Mean Sq	F value	Pr(¿F)	
firma	2	0.260	0.1300	4.179	0.073	.
wycena	3	6.763	2.2544	72.464	4.2e-05	***
Residuals	6	0.187	0.0311			

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Obliczone zostały przez funkcje *p-value* zatem, porównując z $\alpha = 0.05$ widzimy że nie ma różnicy wyceniania pomiędzy firmami, natomiast istnienie różnica wyceny między przedsiębiorstwami. Potwierdzone jest zatem to co zostało obliczone na wykładzie.

45.2 b)

Aby wyznaczyć przedział ufności wykorzystano funkcje *TukeyHSD(model, conf.level = 0.9)* która oddaje następującą tablice:

$firma				
	diff	lwr	upr	p adj
B-A	0.25	-0.06381888	0.56381888	0.1918699
C-A	-0.10	-0.41381888	0.21381888	0.7156978
C-B	-0.35	-0.66381888	-0.03618112	0.0692699
$wycena				
2-1	1.5000000	1.0860287	1.9139713	0.0001924
3-1	0.6333333	0.2193620	1.0473047	0.0178616
4-1	1.9333333	1.5193620	2.3473047	0.0000445
3-2	-0.8666667	-1.2806380	-0.4526953	0.0038512
4-2	0.4333333	0.0193620	0.8473047	0.0851260
4-3	1.3000000	0.8860287	1.7139713	0.0004312

Przedział ufności ma granice zaznaczone pod "lwr" i "upr". Zatem dla różnicy B-A przedział ufności jest następujący:

$$(-0.06381888; 0.56381888)$$

Wynik ten nie zgadza się z wartościami obliczonymi na wykładzie, może to wynikać z tego że funkcja została źle użyta lub że wynik z wykładu jest nie prawidłowy.

Część VIII
Laboratoria 14

Celem tych laboratoriów była kontynuacja pracy nad testami ANOVA. Wzbogaciliśmy zatem wiedzę w tej dziedzinie.

46 Zadanie 2

W celu sprawdzenia wpływu trzech typów maszyn M1, M2 i M3 na wydajność pracy robotników, przeprowadzono eksperyment, w którym jako wydajność pracy mierzono liczbę detali wyprodukowanych w ciągu godziny dla pięciu robotników R1, R2, R3, R4 i R5 pracujących na poszczególnych typach obrabiarek.

Wyniki doświadczenia

Robotnik / maszyna	M1	M2	M3
R1	28	30	26
R2	24	21	27
R3	20	22	18
R4	25	25	25
R5	32	28	30

Przyjmując, że wydajność pracy ma rozkład normalny, sprawdzić na poziomie istotności $\alpha = 0,05$ wpływ typu maszyn oraz indywidualnych cech robotników na ich wydajność pracy.

Jako pierwsze przygotowano dane w pliku csv w celu wykorzystania ich w obliczeniach za pomocą języka programowania R. Poniżej przedstawione tego typu dane:

czas	**robotnik**	**maszyna**
28	"R1"	"M1"
30	"R1"	"M2"
26	"R1"	"M3"
24	"R2"	"M1"
21	"R2"	"M2"
27	"R2"	"M3"
20	"R3"	"M1"
22	"R3"	"M2"
18	"R3"	"M3"
25	"R4"	"M1"
25	"R4"	"M2"
25	"R4"	"M3"
32	"R5"	"M1"
28	"R5"	"M2"
30	"R5"	"M3"

Wyznaczymy hipotezę zerową dla każdego badania:

$$H_0 : \mu_R = \mu_M$$

To znaczy że przeciętne czasy są sobie równe, albo inaczej, że badana cecha nie ma wpływu na czas przeciętny pracy pracownika. Najpierw zbadamy czy czas jest zależny od robotnika następująca funkcją w R: *sModel1 = aov(czas* $\sim$ *robotnik, data)*. Wynik tego jest następujący *summary(sModel1)*=

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
robotnik	4	177.6	44.4	10.57	0.00129	**
Residuals	10	42.0	4.2			

Obliczone *p-value* w przedostatniej kolumnie porównany z przyjętym poziomem istotności $\alpha = 0.05$ jest mniejsze, co oznacza że odrzucamy hipotezę zerową i wnioskujemy że czas pracy może być zależny od robotnika; natomiast musimy dalej sprawdzić inne cechy.

Następnie sprawdzimy czy czas pracy jest zależny od maszyny. Wykorzystamy tą samą funkcje co wygląda następująco: *sModel2 = aov(czas ∼ maszyna, data)*. Wynik tego jest następujący *summary(sModel2)=*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
maszyna	2	1.2	0.6	0.033	0.968
Residuals	12	218.4	18.2		

Obliczone *p-value*, znajdujące się w ostatniej kolumnie, jest większe od przyjętego poziomu istotności $\alpha = 0.05$; zatem nie możemy odrzucić hipotezę zerową mówiącą ze typ maszyny nie ma wpływu na czas pracy.

Ponieważ badamy dwie cechy dokonamy teraz testu korzystając z modelu addytywnego w celu lepszego sprawdzenia czy dana cecha wpływa na czas pracy. Funkcja wygląda podobnie jak poprzednio z lekką zmianą: *model1 = aov(czas ∼ maszyna+robotnik, data)*. Wynik tego jest następujący *summary(model1)=*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
robotnik	4	177.6	44.4	8.706	0.00518	**
maszyna	2	1.2	0.6	0.118	0.89052	
Residuals	8	40.8	5.1			

Widzimy że poprzednio wyznaczone wnioski są poprawne. Można także zauważyć że wartości *p-value* się zwiększyła dla zależności od czasu od robotnika, a zmniejszyła się dla zależności czasy od maszyny.

Pod koniec, aby sprawdzić czy typ maszyny nie wpływa na danego pracownika zastosujemy model multiplikatywny i dokonamy tego samego badania. Funkcja wykorzystana będzie wyglądać następująco: *model2 = aov(czas ∼ maszyna*robotnik, data)*. Wynik tego jest następujący *summary(model2)=*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
maszyna	2	1.2	0.6			
robotnik	4	177.6	44.4			
maszyna:robotnik	8	40.8	5.1			

Widzimy że tym razem nie zostało obliczone *p-value* i że wartości nie różnią się od zastosowanej poprzednio funkcji. Zatem możemy wywnioskować że nie istnieje interakcja pomiędzy tymi dwoma cechami.

Podsumowując, typ maszyny nie ma wpływu na wydajność pracy, co znaczy że możemy zastosować dowolną maszynę i utrzymać tę samą wydajność; natomiast pracownik ma wpływ na wydajność, co mogło się wydawać oczywiste ponieważ nie każda osoba pracuje tak samo.

47 Zadanie 3

W celu zbadania wpływu zestawu zadaniowego A, B C i D na ocenę zaliczeniową ze statystyki przeprowadzono eksperyment na czterech studentach z kierunku technicznego T1, T2, T3, T4 dając im do rozwiązania wszystkie zestawy zadaniowe.

Uzyskane wyniki punktowe

student / zestaw	A	B	C	D
T1	60	54	50	56
T2	48	40	36	42
T3	56	50	50	52
T4	82	74	70	80

Na poziomie istotności $\alpha = 0,05$ zbadać wpływ osobowości studenta oraz zestawu zadaniowego na oceny zaliczeniowe.

Jak w poprzednim zadaniu dane przygotowano w pliku csv w celu wgrania je do języka programowania R. Poniżej został ten plik csv przedstawiony.

ocena	**student**	**zestaw**
60	"T1"	"A"
54	"T1"	"B"
50	"T1"	"C"
56	"T1"	"D"
48	"T2"	"A"
40	"T2"	"B"
36	"T2"	"C"
42	"T2"	"D"
56	"T3"	"A"
50	"T3"	"B"
50	"T3"	"C"
52	"T3"	"D"
82	"T4"	"A"
74	"T4"	"B"
70	"T4"	"C"
80	"T4"	"D"

Jako hipotezę zerową przyjmiemy że na ocenę nie wpływają badane cechy, to znaczy że przeciętne wartości się nie różnią. W celu badania tej hipotezy najpierw zbadamy zależność oceny od studenta następującą funkcją: *sModel1 = aov(ocena∼student, data)*. Wynik tej funkcji jest następujący: *summary(sModel1) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
student	3	2589	863.0	42.79	1.1e-06	***
Residuals	12	242	20.2			

Obliczona funkcją *p-value* znajdujące się w przedostatniej kolumnie jest mniejsze od przyjętego poziomu istotności $\alpha = 0.05$, zatem odrzucamy hipotezę zerową i wnioskujemy że prawdopodobnie ocena zależy od studenta. Potwierdzimy tę hipotezę po zbadaniu zależności od zestawu zadaniowego.

Podobnie jak dla studentów, zbadamy zależność oceny od zestawu zadaniowego podobną funkcją: *sModel2 = aov(ocena∼zestaw, data)*. Wynik tej funkcji jest następujący *summary(sModel2) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
zestaw	3	219	73.0	0.335	0.8	
Residuals	12	2612	217.7			

Obliczona wartość *p-value* jest większa od przyjętego poziomu istotności $\alpha = 0.05$. Zatem nie możemy odrzucić hipotezę zerową co oznacza że z dużym prawdopodobieństwem ocena nie zależy od zestawu zadaniowego.

Ponieważ badamy dwie cechy spróbujemy zastosować model addytywny, gdye możemy sprawdzić bardziej dokładnie czy ocena jest zależna od danej cechy. Model ten potwierdzi uzyskane już wyniki lub je poprawi. Funkcja do przeprowadzenia testu jest następująca: *model1 = aov(ocena∼student+zestaw, data)*. Wynik tej funcji jest następujący *summary(model1) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
student	3	2589	863.0	337.70	1.45e-09	***
zestaw	3	219	73.0	28.57	6.25e-05	***
Residuals	9	23	2.6			

Widzimy że w tym przypadku oba *p-value* są mniejsze od przyjętego poziomu istotności $\alpha = 0.05$, zatem odrzucamy hipotezę zerową. Oznacza to, że student jak zestaw zadaniowy wpływają na ocenę studenta.

Zbadamy teraz czy badane cechy nie wpływają jedna na drugą stosując model multiplikatywny; funkcja wygląda następująco: *model2 = aov(ocena∼student*zestaw, data)*. Wynik tej funkcji podany poniżej *summary(model2) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
student	3	2589	863.0			
zestaw	3	219	73.0			
student:zestaw	9	23	2.6			

Możemy zauważyć że nie uzyskaliśmy *p-value* ani statystykę F, a także, że wartości są takie same jak dla poprzedniego modelu. Zatem woskujemy że student nie wpływa na zestaw zadań, co jest logicznie prawdą.

Podsumowując, ocena jest zależna i od danego studenta, i od danego zestawu zadań ponieważ nie każdy student jest nauczony tak samo, nie każdy student rozumie zestaw zadań w ten sam sposób i nie każdy zestaw zadań jest w stanie sprawdzić dokładnie wiedzę studenta.

48 Zadanie 4

W celu zbadania wpływu różnych receptur sporządzania betonu i różnego surowca na jego wytrzymałość, przeprowadzono eksperyment dla trzech typów betonu B1, B2 i B3 oraz dla czterech receptur R1, R2, R3 i R4.

Wyniki wytrzymałości na ściskanie betonu (w kG/cm^2)

Beton/receptura	R1,	R2,	R3,	R4
B1	210	200	230	204
B2	202	196	220	200
B3	200	190	210	198

Na poziomie istotności $\alpha = 0,05$ zbadać wpływ typu betonu oraz receptury na wytrzymałość uzyskiwanego betonu na ściskanie.

Jak w poprzednich zadaniach dane przegotowano w pliku csv w celu odczytu przez język programowania R. Plik przedstawiony poniżej.

wytrzymalosc	**beton**	**receptura**
210	"B1"	"R1"
200	"B1"	"R2"
230	"B1"	"R3"
204	"B1"	"R4"
202	"B2"	"R1"
196	"B2"	"R2"
220	"B2"	"R3"
200	"B2"	"R4"
200	"B3"	"R1"
190	"B3"	"R2"
210	"B3"	"R3"
198	"B3"	"R4"

Hipoteza zerowa jest że dane nie są zależne od cechy "beton" ani od cechy "receptura" to znaczy że wartości przeciętne od tej cechy są sobie równe. Najpierw sprawdzimy osobno cechy. Dla cechy "beton" wykorzystana została następująca funkcja: *sModel1 = aov(wytrzymalosc∼beton, data)*. Wynik tej funkcji jest następujący *summary(sModel1)* =

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
beton	2	266	133.0	1.115	0.369	
Residuals	9	1074	119.3			

Funkcja oblicza *p-value*, które znajduje się w przedostatniej kolumnie. Widzimy że ta wartość jest większa od przyjętego poziomu istotności $\alpha = 0.05$; oznacza to że nie mamy podstaw do odrzucenia hipotezy zerowej, czyli ze typ betonu nie wpływa na wytrzymałość. Natomiast wynik końcowy zostanie potwierdzony lub obalony gdy sprawdzimy model addytywny.

Dla cechy "receptura" postępujemy w taki sam sposób. Funkcja wykorzystana do tego jest następująca: *sModel2 = aov(wytrzymalosc∼receptura, data).* Wynik tej funkcji jest następujący *summary(sModel2) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
receptura	3	1014.7	338.2	8.317	0.00767	**
Residuals	8	325.3	40.7			

Obliczona dla tej cechy *p-value* jest mniejsze od przyjętego poziomu istotności $\alpha = 0.05$, zatem odrzucamy hipotezę zerowa i wnioskujemy że typ receptury ma wpływ na wytrzymałość betonu. Natomiast, jak dla poprzedniej cechy, potwierdzimy to badając model addytywny w kolejmyn korku.

Zbadamy teraz model addytywny w celu potwierdzenia uzyskanych wyników; funkcja wygląda następująco : *model1 = aov(wytrzymalosc∼beton+receptura, data).* Wynik tej funkcji jest tabela podana poniżej *summary(model1) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
beton	2	266.0	133.0	13.45	0.006066	**
receptura	3	1014.7	338.2	34.20	0.000361	***
Residuals	6	59.3	9.9			

Możemy zauważyć że w przypadku tego modelu oba *p-value* są mniejsze od przyjętego poziomu istotności $\alpha = 0.05$. Oznacza to że typ betonu jak i receptura wpływają na jego wytrzymałościowy. Oznacza to także że obaliliśmy obliczony wynik dla cechy "beton" obliczona na początku zadania.

Zbadamy teraz czy typ betonu wpływa na recepturę lub odwrotnie; wykorzystamy do tego model multiplikatywny wyglądający w następujący sposób : *model2 = aov(wytrzymalosc∼beton*receptura, data).* Wynik tej funkcji jest następujący *summary(model2) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
beton	2	266.0	133.0			
receptura	3	1014.7	338.2			
beton:receptura	6	59.3	9.9			

Widzimy że wartości uzyskane są identyczne jak w poprzednim modelu co oznacza że cechy nie wpływają na siebie.

Podsumowując, na wytrzymałość betonu wpływa typ betonu jak i receptura ponieważ nie każdy beton jest taki sam i nie każda receptura oddaje te same charakterystyki danemu betonowi.

49 Zadanie 5

(B-M.Ł s.230). Aby przekonać się, czy wielkość frakcji proszku grafitowego (FPG) i ciśnienie (C) wpływają istotnie na kurczenie się sproszkowanego żelaza (duże kurczenie się jest niepożądane), przeprowadzono badania dla sześciu różnych wielkości frakcji proszku grafitowego, ściskając żelazo pod dwoma różnymi ciśnieniami.

Wyniki badań

C	FPG												suma
	1		2		3		4		5		6		
25	1.20	1.15	1.14	1.22	1.15	1.21	1.22	1.14	1.16	1.14	1.22	1.14	14.09
50	1.12	1.09	1.16	1.10	1.14	1.10	1.11	1.17	1.02	1.10	1.08	1.00	13.19
suma	2.32	2.24	2.30	2.32	2.29	2.31	2.33	2.31	2.18	2.24	2.30	2.14	27.28

Wyciągnąć wnioski na podstawie przeprowadzonej analizy wariancji.

Dane przygotowano w pliku csv w celu wgrania je do języka programowania R. Plik przedstawiony poniżej.

kurczenie	**C**	**FPG**
1.20	"25"	"1"
1.15	"25"	"1"
1.14	"25"	"2"
1.22	"25"	"2"
1.15	"25"	"3"
1.21	"25"	"3"
1.22	"25"	"4"
1.14	"25"	"4"
1.16	"25"	"5"
1.14	"25"	"5"
1.22	"25"	"6"
1.14	"25"	"6"
1.12	"50"	"1"
1.09	"50"	"1"
1.16	"50"	"2"
1.10	"50"	"2"
1.14	"50"	"3"
1.10	"50"	"3"
1.11	"50"	"4"
1.17	"50"	"4"
1.02	"50"	"5"
1.10	"50"	"5"
1.08	"50"	"6"
1.00	"50"	"6"

Hipoteza zerowa jest taka że badane cechy FPG i C nie wpływają na kurczenie żelaza, czyli że wartości przeciętne są sobie równe. Ponieważ mamy dużo danych skorzystamy od razu z modelu addytywnego w funkcji tak zebranej :

model1 = aov(kurczenie∼C+FPG, data). Wynik tej funkcji jest następujący *summary(model1) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
C	1	0.03375	0.03375	18.719	0.000458	***
FPG	5	0.01113	0.00223	1.235	0.335902	
Residuals	17	0.03065	0.00180			

Obliczone zostały przez funkcje *p-value*. Przyjmując poziom istotności $\alpha = 0.05$ widzimy że *p-value* dotyczące "C" jest jedynie mniejsze od przyjętego poziomu. Zatem dla tej cechy odrzucamy hipotezę zerową, natomiast dla cechy "FPG" jej nie odrzucamy. Oznacza to że kurczenie żelaza jest zależne od ciśnienia ale nie od frakcji proszku grafitowego.

Ponieważ cechy mogą na siebie wpływać zbadamy jeszcze model multiplikatywny który nam potwierdzi lub nie czy cechy na siebie wpływają; funkcja wygląda następująco : *model2 = aov(kurczenie∼C*FPG, data).* Wynik tej funkcji jest podany w tabeli poniżej *summary(model2) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
C	1	0.03375	0.03375	16.805	0.00147	**
FPG	5	0.01113	0.00223	1.109	0.40567	
C:FPG	5	0.00655	0.00131	0.652	0.66568	
Residuals	12	0.02410	0.00201			

Widzimy że pojawiają się wartości statystyki F jak i *p-value* dla tego modelu, zatem cechy mają na siebie wpływ. Patrząc na *p-value* z wiersza "C:FPG", która mówi nam o stopniu wpływania jednej cechy na drugą, widzimy że jest większa od przyjętego poziomu istotności $\alpha = 0.05$. Zatem, pomimo tego że badane cechy na siebie wpływają, nie mają tak mocny na siebie wpływ żeby nam to przeszkadzało. Potwierdzone zostało także, ze jedynie ciśnienie wpływa na skurczenie żelaza.

Część IX
Bibliografia

50 Bibliografia - Lab 9

(1) https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test

(2) https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/wilcox.test

(3) https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/var.test

(4) https://www.rdocumentation.org/packages/dgof/versions/1.2/topics/ks.test

51 Bibliografia - Lab 11

(1) https://en.wikipedia.org/wiki/Kolmogorov%E2%80%93Smirnov_test

(2) https://www.real-statistics.com/statistics-tables/kolmogorov-smirnov-table/

(3) https://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/kolmogorov-smirnov-test/

(4) https://kindsonthegenius.com/blog/how-to-perform-wald-wolfowitz-test-testing-for-homogeneity-with-run-test/

(5) http://www.jbstatistics.com/chi-square-tests-goodness-of-fit-for-the-binomial-distribution/

(6) https://www.brainkart.com/article/Fitting-of-Binomial,-Poisson-and-Normal-distributions_35137/