Kolokwium zaliczeniowe nr. 2

Krystian Baran 145000

9 czerwca 2021

Spis treści

1 Zadanie 1

Zadanie 1 (15). Stosowane w konstrukcjach budowlanych liny stalowe produkowane są trzema technologiami. Przedmiotem badania są wytrzymałości na rozciąganie W_1, W_2, W_3 tych lin mierzone w kG/cm^2. W tym celu przeprowadzone zostały dla nich badania wytrzymałościowe i uzyskano następujące wyniki:

Dla lin produkowanych pierwszą technologią: 999, 1025, 982, 959, 996, 946, 1041, 1138, 1034, 958, 1039, 1038, 995, 1099, 886, 1064, 1038, 1005, 1010, 1020, 1126, 992, 1029, 1060, 957, 980, 1057, 1037, 891, 923, 1064, 965, 910, 906, 1013, 829, 1066, 1006, 955, 992.

Dla lin produkowanych drugą technologią: 944, 982, 919, 886, 940, 867, 1005, 1148, 995, 883, 1003, 1001, 938, 1090, 779, 1039, 1001, 953, 960, 975.

Dla lin produkowanych trzecią technologią: 967, 861, 980, 1034, 1045, 925, 991, 920, 958, 1004, 949, 975, 932, 909, 946, 896, 991, 1088, 984, 908, 989, 988, 945, 1049, 836, 1014, 988, 955, 960, 970, 1076, 942, 979, 1010, 907, 930, 1007, 987, 841, 873, 779, 1016, 956, 905, 942, 1014, 915, 860, 856, 963.

Polecenia:

a) Przeprowadzić testy losowości danych pomiarowych.

b) Oceniając skośność oraz kurtozę dokonać wstępnej oceny rozkładów wytrzymałości i na tej podstawie wskazać rodziny rozkładów, do których mogą należeć badane wytrzymałości lin.

c) Przeprowadzając odpowiednie testy dokonać identyfikacji rozkładów wytrzymałości lin dla poszczególnych technologii. Uzasadnić wybory zastosowanych testów.

d) Przyjmując poziom ufności 0,9 wyznaczyć minimalną wielkość próby potrzebną do oszacowania oczekiwanej wytrzymałości lin produkowanych drugą technologią z dopuszczalnym błędem maksymalnym 20 kG/cm^2.

e) Wyznaczyć 95-procentowe prawostronne przedziały ufności dla przeciętnych wytrzymałości lin produkowanych poszczególnymi technologiami. Uzasadnić wybór konstrukcji przedziału.

f) Wyznaczyć 95-procentowe dwustronne przedziały ufności dla odchyleń standardowych wytrzymałości lin produkowanych poszczególnymi technologiami.

g) W pewnych zastosowaniach budowlanych norma wytrzymałości lin wynosi 865 kG/cm^2. Dla lin produkowanych trzecią technologią wyznaczyć

95-procentowy prawostronny przedział ufności dla wskaźnika lin nadających się do tego zastosowania. W przypadku niespełniania odpowiedniego założenia zdublować wszystkie dane.

h) Dla wytrzymałości lin produkowanych pierwszą i trzecią technologią sprawdzić, czy odchylenia standardowe σ_1 i σ_3 istotnie różnią się.

i) Dla wytrzymałości lin produkowanych pierwszą i drugą technologią sprawdzić hipotezę $\sigma_1^2 = 0,5\sigma_2^2$.

j) Przyjmując, że hipoteza $\sigma_1^2 = 0,5\sigma_2^2$ jest prawdziwa sprawdzić hipotezę $m_1 - m_2 = 50$, gdzie m_1, m_2 są oczekiwanymi wytrzymałościami lin produkowanych pierwszą i drugą technologią.

k) Na poziomie istotności 0,05 sprawdzić hipotezę, że wytrzymałość lin produkowanych pierwszą technologią jest istotnie większa od wytrzymałości lin produkowanych trzecią technologią.

l) Przeprowadzić testy równości wariancji wytrzymałości lin produkowanych trzema technologiami.

m) Wygenerować próbę o liczebności $n_4 = 80$ wytrzymałości $W_4 \sim N(1000; 65)$ z kodem złożonym z pięciu lub czterech cyfr numeru legitymacji studenckiej. Następnie na podstawie danych dotyczących W_1, W_3, W_4 przeprowadzić analizę wariancji.

1.1 a)

Jako test losowości próby zastosujemy test Walda Wolfowitza. W pakiecie "randtests" w R istnieje funkcja wykonująca taki test, zatem nie musimy obliczać wartości pośrednie. Jest to funkcja *runs.test()*. Funkcja ta oddała następujące wartości które zostały zebrane w tabeli.

Lina	statistic	runs	n1	n2	n	p-value
W_1	0.64072	23	20	20	40	0.5217
W_2	0.45947	12	10	10	20	0.6459
W_3	0.85732	29	25	25	50	0.3913

"statistic" to wartość statystyki testowej, "runs" to ile razy zmienił się znak ponieważ z ciągu liczb przekształcamy na ciąg "+" i "-", "n1" to liczba znaków "+" lub "-", "n2" to liczba znaków przeciwna do znaku liczb z "n1", "n" to całkowita liczebność próby, "p-value" to interesowana nas wartość do porównania z poziomem istotności który przyjmiemy $\alpha = 0.05$. Widzimy że każde *p-value* jest od poziomu istotności większe, a zatem nie możemy odrzucić hipotezę zerową mówiącą że dane pochodzą z losowej próby.

1.2 b)

Za pomocą funkcji w R obliczono skośność *skewness* i kurtozę *kurtosis*; uzyskano następujące dane dla każdego zestawu:

Lina	skośność	kurtoza
W_1	-0.3325828	3.229922
W_2	0.002122275	3.694011
W_3	-0.385574	3.210685

W_1 i W_3 mają ujemną skośność co oznacza że rozkład może być typu Weibulla z odpowiednio dobranym parametrem lub rozkład minimum Gumbela. Natomiast W_2 ponieważ ma skośność w aproksymacji równą 0 i dodatnią kurtozę to może to być rozkład normalny, lub, nie aproksymując skośność do 0, rozkład logarytmiczno-normalny, maksymalnego Gumbela, gamma lub Weibulla; najbardziej prawdopodobne jest rozkład normalny dal tych danych.

1.3 c)

Po przeprowadzeniu wstępnej analizy skośności i kurtozy możemy przeprowadzić testy zgodności z danym rozkładem aby sprawdzić czy dane mają jeden z wymienionych wcześniej rozkładów. Zastosujemy test Kołmogorowa Smirnova dostępny w R pod *ks.test()*. Aby sprawdzić zgodność z danym rozkładem parametry tego rozkładu estymowano funkcją *fitdistr()*.

Dla W_1 testowano zgodność z rozkładem Weibulla i normalnym. Wartosci oddane przez *fitdistr()* i podane poniżej:

	fitdistr		ks.test	
Rozklad	shape	scale	D	p-value
Weibull	17.276191	1030.197109	0.10195	0.8
	mean	sd		
Normal	1000.750000	64.078760	0.095693	0.8573

Ponieważ *p-value* jest większe niż poziom istotności $\alpha = 0.05$ nie możemy odrzucić hipotezę o równaniu się temu rozkładowi. Zatem dane mogą pochodzić z rozkładu Weibulla lub z rozkładu normalnego.

Dla W_2 testowano zgodność z rozkładem normalnym, Weibulla, logarytmiczno-normalny i gamma. Użyto te same funkcje co poprzednio i otrzymano wartości podane poniżej.
Normalny : mean = 965.40000, sd = 78.34437
Weibulla : shape = 12.630921, scale = 1001.631919
Gamma : shape = 144.2522661, rate = 0.1494225
Lognormal : meanlog = 6.86920904, sdlog = 0.08200440

Rozkład	D	p-value
Normalny	0.15662	0.7105
Weibulla	0.20229	0.3863
Gamma	0.15371	0.7321
Lognormal	0.14775	0.7751

Porównując *p-value* z przyjętym poziomem istotności $\alpha = 0.05$ widzimy że nie możemy odrzucić żadną hipotezę zerową równania się rozkładów. Natomiast, porównując otrzymane wartości, widzimy że najmniejsza jest dla rozkładu Weibulla, co oznacza że dane te najmniej prawdopodobnie pochodzą z rozkładu Weibulla.

Dla W_3 także przeprowadzono test dla rozkładu Weibulla i normalnym jak dla poprzednich danych. Otrzymane wartości zapisano poniżej:

	fitdistr		ks.test	
Rozklad	shape	scale	D	p-value
Weibull	17.177821	982.722097	0.095061	0.7569
	mean	sd		
Normal	954.300000	62.404888	0.081875	0.8909

Obliczone *p-value* jest większe od przyjętego poziomu istotności $\alpha = 0.05$ zatem nie możemy odrzucić hipotezę zerową że dane pochodzą z rozkłady Weibulla tak jak i z rozkładu normalnego.

1.4 d)

Ponieważ dane z drugiej metody są z dużym prawdopodobieństwem normalnie rozłożone a nie znamy wartości oczekiwanej ani wariancji tego rozkładu, możemy zastosować poniższy wzór na obliczenie minimalnej liczebności próby.

$$n = \left\lceil \frac{t^2_{1-\frac{\alpha}{2};n_0-1} S^2_{n_0}}{d^2} \right\rceil$$

Gdzie $d = 20$ jest maksymalny błąd, $n_0 = 20$ to liczebność próby, $1 - \alpha = 0.9 \rightarrow \alpha = 0.1$

$$S^2_{n_0} \stackrel{R}{=} var(w2) = 6460.884$$

$$t_{1-\frac{\alpha}{2};n_0-1} \stackrel{R}{=} qt(0.95, 19) = 1.729133$$

Podstawiając do wzoru otrzymujemy:

$$n = \left\lceil \frac{1.729133^2 \cdot 6460.884}{20^2} \right\rceil = \lceil 48.292507 \rceil = 49$$

Zatem, minimalna próba żeby na poziomie ufności 0.9 i z błędem maksymalnym 20 kG/cm^2 wyznaczyć wartość oczekiwaną jest 49.

1.5 e)

Aby wyznaczyć przedział ufności na poziome ufności $1 - \alpha = 0.95$ potrzebujemy rozważyć dane. Dla W_1 i W_3 nie przyjmujemy że to rozkład normalny, natomiast liczebność jest na tyle wielka że można zastosować poniższy wzór na prawostronny przedział ufności.

$$(\overline{W} - z_{1-\alpha}\frac{S_n}{\sqrt{n}}, \infty)$$

Gdzie:

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

Lina	$\overline{W}$	S_n^2	S_n
W_1	1000.75	4211.372	64.89508
W_2	965.4	6460.884	80.37963
W_3	954.3	3973.847	63.03846

Zatem dla W_1:

$$1000.75 - 1.644854\frac{64.89508}{\sqrt{40}} \approx 983.872461$$

$$(983.872461; \infty)$$

Dla W_3:

$$954.3 - 1.644854\frac{63.03846}{\sqrt{50}} \approx 939.636152$$

$$(939.636152; \infty)$$

Dla W_2, ponieważ nie odrzuciliśmy hipotezę o równości z rozkładem normalnym, możemy zastosować następujący wzór na przedział ufności:

$$(\overline{W} - t_{1-\alpha;n-1}\frac{S_n}{\sqrt{n}}, \infty)$$

$$t_{1-\alpha;n-1} \overset{R}{=} qt(0.95, 19) \approx 1.729133$$

Zatem

$$965.4 - 1.729133\frac{80.37963}{\sqrt{20}} \approx 934.321546$$

$$(934.321546; \infty)$$

1.6 f)

Ponieważ przeprowadziliśmy test równości z rozkładem normalny i nie uzyskaliśmy sprzecznego wyniku możemy dla każdego zestawu danych wyznaczyć dwustronny przedział ufności. Natomiast, dla W_2, sokoro mamy mniej niż 20 elementów to zastosujemy inna konstrukcje jak dla W_1 i W_3.

Dla W_1 i W_3 wzór na przedział będzie następujący, gdzie badane jest odchylenie standardowe:

$$\left(\frac{\sqrt{\frac{n-1}{n}}S_n}{1+\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{2n}}};\frac{\sqrt{\frac{n-1}{n}}S_n}{1-\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{2n}}}\right)$$

Gdzie

$$z_{1-\frac{\alpha}{2}} \stackrel{R}{=} qnorm(0.975, 0, 1) \approx 1.959964$$

Zatem dla W_1 przedział wygląda następująco:

$$\left(\frac{\sqrt{\frac{39}{40}}\cdot 64.89508}{1+\frac{1.959964}{\sqrt{80}}};\frac{\sqrt{\frac{39}{40}}\cdot 64.89508}{1-\frac{1.959964}{\sqrt{80}}}\right)$$

$$(52.561026; 82.06079)$$

Dla W_3 przedział będzie wyglądać następująco:

$$\left(\frac{\sqrt{\frac{49}{50}}\cdot 63.03846}{1+\frac{1.959964}{\sqrt{100}}};\frac{\sqrt{\frac{49}{50}}\cdot 63.03846}{1-\frac{1.959964}{\sqrt{100}}}\right)$$

$$(53.714921; 79.903686)$$

Dla W_2 wzór będzie następujący gdzie badana jest wariancja zamiast odchylenia standardowego

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

Gdzie

$$\chi^2_{1-\frac{\alpha}{2};n-1} \stackrel{R}{=} qchisq(0.975, 19) \approx 32.85233$$

$$\chi^2_{\frac{\alpha}{2};n-1} \stackrel{R}{=} qchisq(0.025, 19) \approx 8.906516$$

Zatem przedział ufności będzie wyglądać następująco

$$\left(\frac{(19)\cdot 6460.884}{32.85233};\frac{(19)\cdot 6460.884}{8.906516}\right)$$

$$(3736.623734; 13782.80755)$$

Chcąc przejść na odchylenie standardowe wystarczy pierwiastkować końce przedziałów:

$$(61.12793; 117.400203)$$

1.7 g)

Aby wyznaczyć przedział ufności wskaźnika struktury dla liczb nadając się do zastosowań budowlanych musimy obliczyć które liczby spełniają warunek. Utworzymy wtedy ciąg 0 i jedynek który będzie miał rozkład Bernoullego i będziemy mogli zbadać wskaźnik struktury. Przypiszemy do ciągu liczbę 1 gdy dana liczba z próby jest większa lub równa $865kG/cm^2$ a w przeciwnym wypadku liczbę 0. Ilość jedynek podzielono przez liczebność i otrzymano $\overline{P}_n$

Lina	"1"	"0"	$\overline{P}_n$
W_3	44	6	0.88

Sprawdzimy warunek konieczny istnienia przedziału ufności:

$$\begin{array}{lcr} 0 & < \overline{P}_n \pm 3\sqrt{\frac{\overline{P}_n(1-\overline{P}_n)}{n}} & < 1 \\ 0 & < 0.88 \pm 3\sqrt{\frac{0.88 \cot 0.12}{50}} & < 1 \\ \hline 0 & < 0.742130 & < 1 \\ 0 & < 1.017870 & < 1 \end{array}$$

Drugi warunek nie jest spełniony, zatem dublujemy dane, wtedy liczebność $n = 100$.

Lina	"1"	"0"	$\overline{P}_n$
W_3	88	12	0.88

Warunek wygląda następująco:

$$\begin{array}{lcr} 0 & < 0.782511 & < 1 \\ 0 & < 0.977488 & < 1 \end{array}$$

Warunek jest teraz spełniony, zatem możemy wyznaczyć przedział ufności prawostronny z następującego wzoru:

$$\left(\overline{P}_n - z_{1-\alpha}\sqrt{\frac{\overline{P}_n(1-\overline{P}_n)}{n}}; \infty\right)$$

Gdzie :

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

Wtedy :

$$\left(0.88 - 1.644854\sqrt{\frac{0.88 \cdot 0.12}{100}}; \infty\right)$$

$$(0.826549; \infty)$$

Podany powyżej jest 95-procentowy przedział ufności wskaźnika struktury lin spełniających normę.

1.8 h)

Ponieważ nie odrzuciliśmy hipotezę o normalności danych pierwsza i trzecią technologią możemy zastosować test F aby sprawdzić czy odchylenia standardowe różnią się. Wyznaczymy hipotezę zerową oraz hipotezę alternatywną:

H_0	$\frac{\sigma_1^2}{\sigma_3^2} = 1$
H_0	$\frac{\sigma_1^2}{\sigma_3^2} \neq 1$

Zastosujemy do tego statystykę :

$$F = \frac{max\{S_1^2, S_3^2\}}{min\{S_1^2, S_3^2\}} = \frac{4211.372}{3973.847} \approx 1.059772 = F_0$$

Statystyka ta ma rozkład statystyki $\sim F(n_1 - 1, n_3 - 1)$ gdzie n_1 to liczebność licznika a n_3 to liczebność mianownika.
Obszar krytyczny, dla poziomu istotności $\alpha = 0.05$ jest następujący:

$$\begin{aligned} R &= (F_{1-\alpha/2;n_1-1;n_3-1}, \infty) \\ &\stackrel{R}{=} (qf(0.975, 39, 49), \infty) \\ &= (1.808208, \infty) \end{aligned}$$

Obliczona wcześniej wartość F_0 nie należy do obszaru krytycznego, zatem nie możemy odrzucić hipotezę zerową że wariancje są sobie równe. Zatem odchylenia standardowe technologi pierwszej i trzeciej nie różnią się istotnie.

1.9 i)

Przedstawiona hipoteza jest hipotezą zerową, zatem dopełnimy o hipotezę alternatywną:

H_0	$\frac{\sigma_1^2}{\sigma_2^2} = 0.5$
H_0	$\frac{\sigma_1^2}{\sigma_2^2} \neq 0.5$

Zastosujemy statystkę F która w tym przypadku będzie wyznaczona następująco:

$$F = \frac{\frac{s_1^2}{\sigma_1^2}}{\frac{s_2^2}{0.5\sigma_2^2}} = 0.5\frac{s_1^2}{s_2^2} = 0.5\frac{max\{S_1^2, S_3^2\}}{min\{S_1^2, S_3^2\}}$$

Statystka ta ma jak taki sam rozkład jak w poprzednim podpunkcie. Obliczymy F_0 i obszar krytyczny:

$$F_0 = 0.5\frac{6460.884}{4211.372} \approx 0.767076$$

$$\begin{aligned} R &= (F_{1-\alpha/2;n_1-1;n_3-1}, \infty) \\ &\stackrel{R}{=} (qf(0.975, 19, 39), \infty) \\ &= (2.095977, \infty) \end{aligned}$$

Wartość F_0 nie należy do obszaru krytycznego zatem nie możemy odrzucić hipotezę zerową. Oznacza to że prawdopodobnie $\sigma_1^2 = 0.5\sigma_2^2$

1.10 j)

Ponieważ zakładamy że wariancje są od siebie różne, możemy zastosować test Cochrana Coxa. Wyznaczymy hipotezę zerową oraz alternatywną:

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

Zastosujemy zastępującą statystykę:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - m_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Statystyka ta ma rozkład t-Studenta z v stopniami swobody wyznaczone poniżej:

$$v = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_1^2}{n_1}\right)^2} = \frac{\left(\frac{4211.372}{40} + \frac{6460.884}{20}\right)^2}{\frac{1}{39}\left(\frac{4211.372}{40}\right)^2 + \frac{1}{19}\left(\frac{6460.884}{20}\right)^2} \approx 31.75938$$

Wyznaczymy t_0 podstawiając do statystyki znane wartości:

$$t_0 = \frac{(1000.75 - 965.4) - 50}{\sqrt{\frac{4211.372}{40} + \frac{6460.884}{20}}} \approx -0.707863$$

Obszar krytyczny wyznacza się jak poniżej:

$$\begin{aligned} R &= (-\infty, t_{\frac{\alpha}{2};v}) \cup (t_{1-\frac{\alpha}{2};v}, \infty) \\ &\overset{R}{=} (-\infty, qt(0.025, 31.75928)) \cup (qt(0.975, 31.75928), \infty) \\ &= (-\infty, -2.037539) \cup (2.037539, \infty) \end{aligned}$$

Obliczona wartość t_0 nie należy do obszaru krytycznego, zatem nie możemy odrzucić hipotezę zerową że różnica wartości oczekiwanych pierwszej i drugiej techniki jest równa 50.

1.11 k)

Hipoteza że wytrzymałość lin produkowanych pierwsza technologią jest istotnie większa od tych produkowanych trzecią technologią jest hipotezą alternatywną; zatem dopełnimy hipotezę alternatywną:

H_0	$m_1 - m_3 \leqslant 0$
H_0	$m_1 - m_3 > 0$

Zakładając że rozkład nie znamy, i wiemy że liczebności próby są większe od 30 możemy zastosować następującą statystykę:

$$z = \frac{(\overline{X}_1 - \overline{X}_2) - m_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Która ma rozkład statystyki zbliżony asymptotycznie do $\sim N(0,1)$.
Wyznaczymy najpierw wartość z_0 podstawiając znane już wartości:

$$z_0 = \frac{1000.75 - 954.3}{\sqrt{\frac{4211.372}{40} + \frac{3973.847}{50}}} \approx 3.417278$$

Obliczymy tym razem *p-value* zgodnie ze wzorem:

$$\text{p-value} = 1 - F_{N(0,1)}(z_0) \overset{R}{=} 1 - pnorm(3.417278, 0, 1) \approx 0.0003162533$$

Wartość ta jest mniejsza od przyjętego poziomu istotności $\alpha = 0.05$, zatem odrzucamy hipotezę zerową i wnioskujemy że wytrzymałość lin produkowanych pierwszą technologią jest istotnie większa od lin produkowanych trzecią technologią.

1.12 l)

Ponieważ nie odrzuciliśmy hipotezę równości danych z rozkładem normalnym to do badania równości wariancji wszystkich trzech technologi produkcji możemy zastosować test Bartletta. Wyznaczymy hipotezę zerową oraz alternatywną:

H_0	$\sigma_1^2 = \sigma_2^2 = \sigma_3^2$
H_1	$\sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$

Dla tego testy stosuję się następującą statystykę:

$$M = \frac{1}{C}\left[(N-k)\ln S^2 - \sum_{i=1}^{k}(n_i - 1)\ln S_i^2\right]$$

Statystyka ta ma rozkład statystyki asymptotycznie zbliżony do rozkładu $\sim \chi^2(k-1)$. Wyznaczymy najpierw parametry zgodnie ze wzorami:

$$N = \sum_{i=1}^{k} n_i = 40 + 20 + 50 = 110$$

$$\begin{aligned} C &= 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right) \\ &= 1 + \frac{1}{6}\left(\frac{1}{39} + \frac{1}{19} + \frac{1}{49} - \frac{1}{107}\right) \\ &\approx 1.014889 \end{aligned}$$

$$\begin{aligned} S^2 &= \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)S_i^2 \\ &= \frac{1}{107}(39 \cdot 4211.372 + 19 \cdot 6460.884 + 49 \cdot 3973.847) \\ &\approx 4502.044925 \end{aligned}$$

Możemy teraz obliczyć M_0 podstawiając znane wartości:

$$\begin{aligned} M_0 = & \frac{1}{1.014889}\left[107 \ln 4502.044925 - (39 \cdot \ln 4211.372 + \right. \\ & \left. +19 \cdot \ln 6460.884 + 49 \cdot \ln 3973.847)\right] \\ & \approx 1.827381 \end{aligned}$$

Wyznaczymy teraz obszar krytyczny zgodnie ze wzorem:

$$R = (\chi^2_{1-\alpha;k-1}, \infty) \stackrel{R}{=} (qchisq(0.95, 2), \infty) = (5.991465, \infty)$$

Wartość M_0 nie należy do obszaru krytycznego, zatem nie możemy odrzucić hipotezę o równości wariancji dla wszystkich trzech technologi produkcji.

1.13 m)

Aby przeprowadzić test równości wariancji przygotowano dane w pliku csv i wgrano do R gdzie wykonano test funkcją *model = aov(pomiar∼technologia, data)*. Do tego pliku dodano dane wygenerowane za pomocą funkcji w R *round(rnorm(80, 1000,65))*. Jako seed wzięto liczbę 1450 (*set.seed(1450)*) Wynik testu jest następujący *summary(model) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
technologia	3	60950	20317	4.882	0.00273	**
Residuals	186	774053	4162			

Widzimy że obliczone *p-value* w przedostatniej kolumnie jest mniejsze niż przyjęty poziom istotności $\alpha = 0.05$ co oznacza że wartości przeciętne są zależne od technologi produkcji linii stalowej.

Przeprowadzimy teraz test Tukeya aby sprawdzić które technologie najbardziej się różnią. Funkcja jest następująca : *TukeyHSD(model)*. Oddana jest następująca tabela:

	diff	lwr	upr	p adj
2-1	-35.350	-81.149711	10.44971	0.1913526
3-1	-46.450	-81.926304	-10.97370	0.0046313
4-1	-11.775	-44.160286	20.61029	0.7819284
3-2	-11.100	-55.346724	33.14672	0.9153216
4-2	23.575	-18.234225	65.38422	0.4627272
4-3	34.675	4.525939	64.82406	0.0169986

Porównując obliczone w ostatniej kolumnie *p-value* z poziomem istotności $\alpha = 0.05$ widzimy że metoda W_1 różni się zasadniczo od metody i że metoda W_3 różni się od metody W_4. Inne kombinacje metod natomiast nie różnią się zasadniczo gzie najmniej różniące między sobą się technologie są W_3 i W_2, natomiast najmniej nie różniące się technologie są W_1 i W_2.

2 Zadanie 2

Wygenerować 100 elementowe próby według rozkładów Weibulla *wbl(2; 2000)*, *wbl(3; 2000)*, *wbl(4; 2000)*

a) Dla otrzymanych prób obliczyć podstawowe statystyki w tym skośność i kurtozę.

b) Sporządzić histogramy na tle wykresów teoretycznych gęstości.

c) Przyjmując, że mechanizm generowania danych jest nieznany dokonać identyfikacji dystrybuant $\hat{F}_i$.

d) Sporządzić wykresy funkcji $\hat{R}_i(t) = 1 - \hat{F}_i(t)$.

e) Przeprowadzić test równości wariancji

f) Przeprowadzić test równości wartości oczekiwanych.

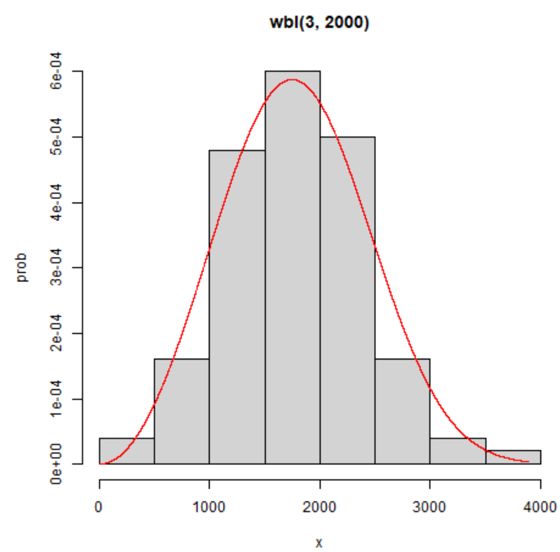
2.1 a)

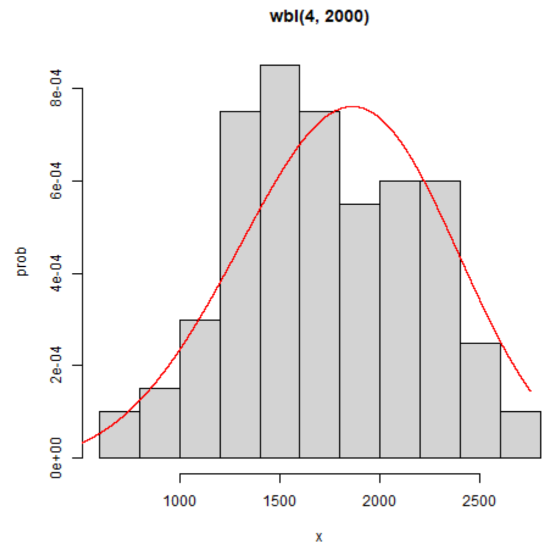
Dane wygenerowano w R za pomocą funkcji *rweibull(2, 2000)* zmieniając odpowiednio parametry kal podano w zadaniu. Uzyskane dane zaokrąglono do liczby całkowitej i zapisano w pliku csv. Następnie, odpowiednimi funkcjami obliczono: średnia, wariancje, odchylenie standardowe, skośność i kurtozę. Wyniki przedstawiono w tabeli poniżej. Dane znajdują się pod zakładką dane.

Rozkład	$\overline{X}$	S_n^2	S_n	skewness(data)	kurtosis(data)
wbl(2, 2000)	1706.28	696814.7	834.7543	0.7214066	3.22036
wbl(3, 2000)	1790.01	392755.3	626.7019	0.2040186	3.523997
wbl(4, 2000)	1726.02	211770.1	460.1849	-0.02536129	2.369829

2.2 b)

Wykresy także sporządzono w R i przedstawiono poniżej.





2.3 c)

Obliczone wartości skośności są różne od 0. Dla pierwszych dwóch rozkładów są one większe od 0, zatem może to być rozkład Logarytmiczno-normalny, Gamma lub Weibulla. Natomiast Dla ostatniego rozkład jest to wartość ujemna zatem może to być rozkład Weibulla. Ponieważ skośność tej ostatniej jest bliska zero i większa od -1 to może to także być rozkład normalny, natomiast dla każdego zestawu danych dokonamy odpowiednie testy.
Na podstawie wyznaczonych możliwych rozkładów dla każdego z danych wykonano test Kołmogorowa Smirnowa wyznaczając dystrybuantę empiryczną, obliczając maksimum z wartości bezwzględnej różnicy dystrybuanty empirycznej i estymowanej (D0) i porównując te wartość z D wzięte z tabeli pod **Tabele** przyjmując poziom istotności $\alpha = 0.05$ estymując parametry rozkładu za pomocą funkcji *fitdistr()* i porównując dane z próby z taką dystrybuantą.
Dla pierwszych danych:

Rozkład	fitdistr()		Test	
	shape	scale	D0	D
Weibull	2.1896307	1934.5700201	0.04786828	0.13581
	shape	rate		
Gamma	4.0557889946	0.0023769782	0.04386573	0.13581
	meanlog	sdlog		
Lognormal	7.31375382	0.52995926	0.05646625	0.13581
	mean	sd		
Normal	1706.28000	830.57001	0.08132326	0.13581

Widzimy że wszystkie obliczone D0 są mniejsze od wartości D, oznacza to że nie odrzucimy hipotezę zerową o równaniu się danych z każdym rozkładem. Natomiast Widzimy że największe obliczone D0 jest przy rozkładzie normalnym, co oznacza że rozkład Normalny jest najmniej prawdopodobny dla tych danych. Najbardziej prawdopodobny rozkład to rozkład Gamma a w drugiej kolejności rozkład Weibulla.

Dla drugich danych:

Rozkład	fitdistr()		Test	
	shape	scale	D0	D
Weibull	3.0780526	1996.3882743	0.05232416	0.13581
	shape	rate		
Gamma	6.6467323587	0.0037132413	0.09520625	0.13581
	meanlog	sdlog		
Lognormal	7.41286911	0.43681013	0.120455	0.13581
	mean	sd		
Normal	1790.01000	623.56054	0.04106466	0.13581

Jak poprzednio nie odrzucimy hipotezę równości z każdym z rozkładów, natomiast w tym przypadku rozkład Logarytmiczno normalny jest najmniej prawdopodobny. Najbardziej prawdopodobny rozkład jest rozkład normalny a w drugiej kolejności rozkład Weibulla.

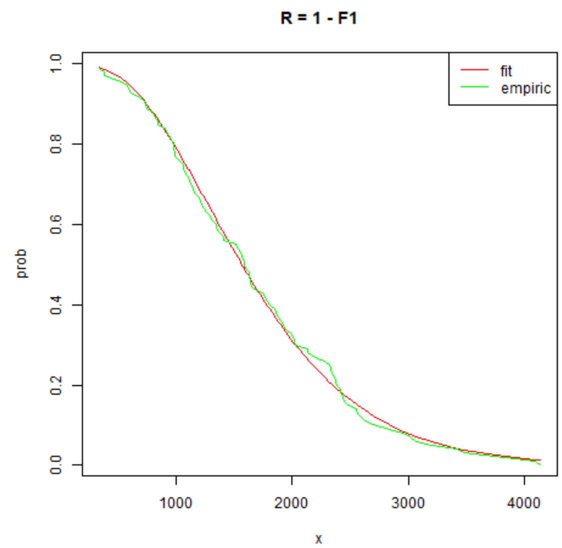
Dla trzecich danych:

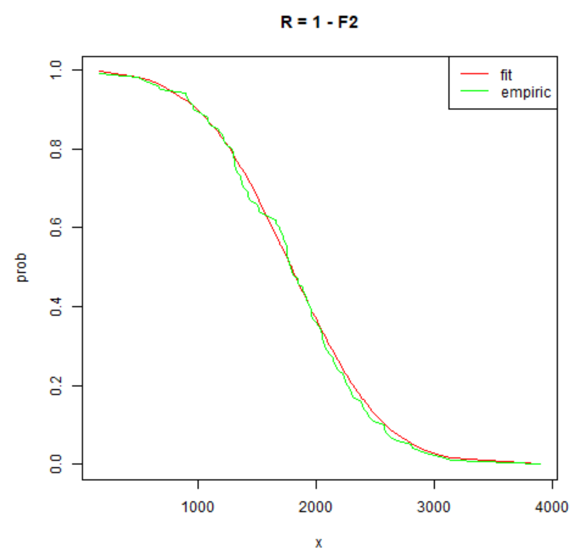
Rozkład	fitdistr()		Test	
	shape	scale	D0	D
Weibull	4.2490242	1899.4982792	0.05586626	0.13581
	mean	sd		
Normal	1726.02000	457.87819	0.05934267	0.13581

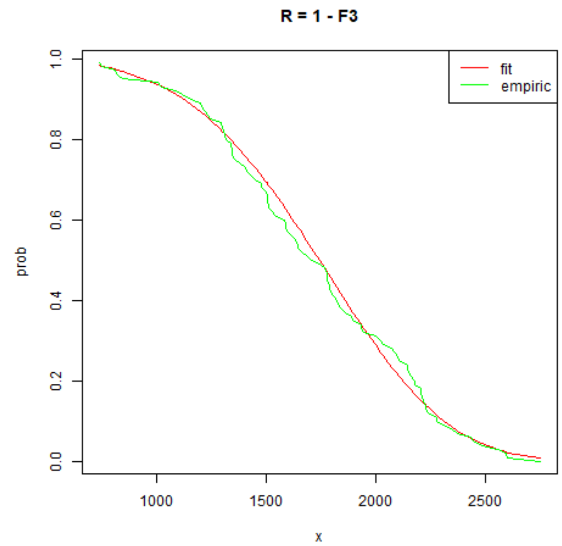
W tym przypadku także nie odrzucimy hipotezę równości z rozkładem normalnym i Weibulla, natomiast rozkład Weibulla wydaję się najbardziej prawdopodobny

2.4 d)

Ponieważ nie został znaleziony rozkład w sposób jednoznaczny dla danych wykorzystamy ten najbardziej prawdopodobny, oraz dokonamy wykres wykorzystując dystrybuantę empiryczną. Wykresy w kolejności pojawienia się danych.


R = 1 - F1
fit
empiric
prob
x
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000


R = 1 - F2
fit
empiric
prob
x
0.0
0.2
0.4
0.6
0.8
1.0
1000
2000
3000
4000



2.5 e)

Ponieważ nie odrzuciliśmy hipotezę równości danych z rozkładem normalnym to do badania równości wariancji wszystkich trzech technologi produkcji możemy zastosować test Bartletta. Wyznaczymy hipotezę zerową oraz alternatywną:

H_0	$\sigma_1^2 = \sigma_2^2 = \sigma_3^2$
H_1	$\sigma_1^2 \neq \sigma_2^2 \neq \sigma_3^2$

Dla tego testy stosuję się następującą statystykę:

$$M = \frac{1}{C}\left[(N-k)\ln S^2 - \sum_{i=1}^{k}(n_i - 1)\ln S_i^2\right]$$

Statystyka ta ma rozkład statystyki asymptotycznie zbliżony do rozkładu $\sim \chi^2(k-1)$. Wyznaczymy najpierw parametry zgodnie ze wzorami:

$$N = \sum_{i=1}^{k} n_i = 100 + 100 + 100 = 300$$

$$\begin{aligned} C &= 1 + \frac{1}{3(k-1)}\left(\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{N-k}\right) \\ &= 1 + \frac{1}{6}\left(\frac{1}{99} + \frac{1}{99} + \frac{1}{99} - \frac{1}{297}\right) \\ &\approx 1.004493 \end{aligned}$$

$$S^2 = \frac{1}{N-k}\sum_{i=1}^{k}(n_i-1)S_i^2$$
$$= \frac{99}{297}(696814.7 + 392755.3 + 211770.1)$$
$$\approx 433780$$

Możemy teraz obliczyć M_0 podstawiając znane wartości:

$$M_0 = \frac{1}{1.004493}\left[297 \ln 433780 - 99(\ln 696814.7 + \ln 392755.3 + \ln 211770.1)\right]$$
$$\approx 33.746466$$

Wyznaczymy teraz obszar krytyczny zgodnie ze wzorem:

$$R = (\chi^2_{1-\alpha;k-1}, \infty) \stackrel{R}{=} (qchisq(0.95, 2), \infty) = (5.991465, \infty)$$

Wartość M_0 należy do obszaru krytycznego, zatem możemy odrzucić hipotezę o równości wariancji dla wszystkich trzech technologi produkcji i wnioskować że chociaż jedna z wariancji jest różna od pozostałych.

2.6 f)

Aby przeprowadzić test równości wartości oczekiwanych wykonamy test ANOVA z hipotezą zerową i alternatywną jak zapisane poniżej.

H_0	$\mu_1 = \mu_2 = \mu_3$
H_1	$\mu_1 \neq \mu_2 \neq \mu_3$

Hipotezą alternatywa, pomimo tego że jest zapisana tak że żadna wartość oczekiwana jest do siebie równa tak na prawdę jest że chociaż jedna para wartosci oczekiwanych nie jest do siebie równa. Test wykonujemy w R za pomocą następujące funkcji : *model = aov(dane∼zestaw, data)*. Wynik tej funkcji jest następujący *summary(model) =*

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
zestaw	2	383170	191585	0.442	0.643	
Residuals	297	128832673	433780			

Obliczona wartość *p-value* jest większa od przyjętego poziomu istotności $\alpha = 0.05$, zatem nie możemy odrzucić hipotezę zerową o równaniu się wartości oczekiwanych.

3 Tablice

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

$n \backslash \alpha$	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	$\frac{1.94947}{\sqrt{n}}$	$\frac{1.62762}{\sqrt{n}}$	$\frac{1.51743}{\sqrt{n}}$	$\frac{1.35810}{\sqrt{n}}$	$\frac{1.22385}{\sqrt{n}}$	$\frac{1.13795}{\sqrt{n}}$	$\frac{1.07275}{\sqrt{n}}$