Zadania

1. Firma rozważa pięć projektów nazw swojego nowego produktu. Przed wybraniem jednej z nich firma postanowiła sprawdzić, czy wszystkie pięć nazw równie silnie przyciąga klientów. Wybrano próbę losową 100 osób i każdą z nich poproszono o wskazanie najlepszej spośród pięciu nazw. Liczby osób, które wybrały kolejne nazwy są podane poniżej:

Nazwa		*A*	*B*	*C*	*D*	*E*
Liczba osób	a)	8	16	30	34	12
	b)	4	12	34	40	10
	c)	12	30	15	15	28

Przeprowadź test. **Odp.: b)** $\chi_0^2 = 50{,}8$, odrzucamy hipotezę zerową.

Utworzonym przedziałom klasowym odpowiadają liczebności teoretyczne np_i, gdzie $p_i = \mathrm{P}(X \in \Delta_i) = F(g_i) - F(g_{i-1})$, $i = 1, 2, \ldots, k$ oraz $F(g_0) = 0$, $F(g_k) = 1$.

Statystykę

$$\chi^2 = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i}$$

Przyjmujemy, że prawdopodobieństwo każdej z nazwy jest takie samo, stąd $p_i = 0.2$.

Hipoteza zerowa – sprawdzenie losowości próby – próba była losowa.

Korzystając z Excel:

	Nazwa	A	B	C	D	E	Suma
	a)	8	16	30	34	12	100
Liczba osó	b)	4	12	34	40	10	100
	c)	12	30	15	15	28	100
	a) p	0,2	0,2	0,2	0,2	0,2	1
	b) p	0,2	0,2	0,2	0,2	0,2	1
	c) p	0,2	0,2	0,2	0,2	0,2	1
	a) np.	20	20	20	20	20	
	b) np.	20	20	20	20	20	
	c) np.	20	20	20	20	20	100
	a) chi^2	7,2	0,8	5	9,8	3,2	26
	b) chi^2	12,8	3,2	9,8	20	5	50,8
	c) chi^2	3,2	5	1,25	1,25	3,2	13,9

Obszar krytyczny: $R_\alpha = \left(\chi^2_{1-\alpha,k-r-1}, \infty\right)$, gdzie $\chi^2_{1-\alpha,k-r-1}$ jest kwantylem rzędu $1 - \alpha$ rozkładu chis$(k - r - 1)$.

```
> qchisq(0.95,5)
[1] 11.0705
```

Obszar krytyczny: (11.0707; ∞). Wniosek, odrzucamy hipotezę zerowa dla wszystkich przypadków. Próba nie była próbą losową.

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 $\mathcal{N}(2200;\ \sigma)$;
c) ma rozkład $\mathcal{N}(2200;\ 100)$.

Odp.: ii) a) Szacujemy dwa parametry, stąd 3 stopnie swobody, $\chi_0^2 = 2{,}91 < \chi^2_{0{,}95;3}$=7,8147.

a) Musimy estymować parametry:

$$m = \bar{X}$$

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

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

Dokonałam obliczeń korzystając z Excel:

Wytrzymałość			Liczba próbek		
lewy	srodek	prawy	i)	srodek*licz	srodek-srednia
1900	1950	2000	14	27300	950043,5
2000	2050	2100	26	53300	669766,5
2100	2150	2200	52	111800	190333
2200	2250	2300	58	130500	90494,5
2300	2350	2400	33	77550	642188,3
2400	2450	2500	17	41650	975124,3
		suma	200	2210,5	3517950
		sr=m	2210,5		
		S	17678,14		
		S^2	132,9592		

Wyniki: $m = \bar{X} = 2210{,}5$; $\sigma = S = 17678{,}14$; $S^2 = 132{,}9592$

Nie znamy parametrów dystrybuanty więc możemy skorzystać z statystyki χ^2.

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

Gdzie $p_1 = F(2000) - F(1900), p_2 = F(2100) - F(2000)$

Dokonuje obliczeń z wykorzystaniem R:

```
> pnorm(2000,2210.5,132.9592)- pnorm(1900,2210.5,132.9592)
[1] 0.046925
> pnorm(2100,2210.5,132.9592)- pnorm(2000,2210.5,132.9592)
[1] 0.1462748
> pnorm(2200,2210.5,132.9592)- pnorm(2100,2210.5,132.9592)
[1] 0.265564
> pnorm(2300,2210.5,132.9592)- pnorm(2200,2210.5,132.9592)
[1] 0.2810429
> pnorm(2400,2210.5,132.9592)- pnorm(2300,2210.5,132.9592)
[1] 0.1733869
> pnorm(2500,2210.5,132.9592)- pnorm(2400,2210.5,132.9592)
[1] 0.06231571
```

Wytrzymałość			Liczba prób	p	n*p	chi^2
lewy	srodek	prawy	i)			
1900	1950	2000	14	0,046925	9,385	2,26939
2000	2050	2100	26	0,146275	29,255	0,362161
2100	2150	2200	52	0,265564	53,1128	0,023315
2200	2250	2300	58	0,281043	56,2086	0,057093
2300	2350	2400	33	0,173387	34,6774	0,081138
2400	2450	2500	17	0,062316	12,4632	1,651466
					Suma	4,444564

$$\chi^2 = 4{,}444564$$

Następnie sprawdzamy obszar krytyczny:

Obszar krytyczny: $R_\alpha = (\chi^2_{1-\alpha,k-r-1}, \infty)$, gdzie $\chi^2_{1-\alpha,k-r-1}$ jest kwantylem rzędu $1-\alpha$ rozkładu chis$(k-r-1)$.

$$R_\alpha = (\chi^2_{1-\alpha,k-r-1}, \infty)$$

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

$$\chi^2_{1-0.05,6-2-1} = 7{,}814728$$

```
> qchisq(0.95,3)
[1] 7.814728
```

Stąd nasz obszar krytyczny to: $(7.814728;\ \infty)$

Zatem nie odrzucamy hipotezy, dane mają rozkład normalny na podanym poziomie istotności.

b) $N(2200, \sigma)$

Obliczenie odchylenie standardowego już dokonaliśmy w podpunkcie a), wyniosło:

$$\sigma = 17678{,}14$$
$$\sigma^2 = 132{,}9592$$

Natomiast m = 2200 mamy podane.

Dokonujemy analogicznych pomiarów:

```
> pnorm(2000,2200,132.9592)- pnorm(1900,2200,132.9592)
[1] 0.05423731
> pnorm(2100,2200,132.9592)- pnorm(2000,2200,132.9592)
[1] 0.1597301
> pnorm(2200,2200,132.9592)- pnorm(2100,2200,132.9592)
[1] 0.2740077
> pnorm(2300,2200,132.9592)- pnorm(2200,2200,132.9592)
[1] 0.2740077
> pnorm(2400,2200,132.9592)- pnorm(2300,2200,132.9592)
[1] 0.1597301
> pnorm(2500,2200,132.9592)- pnorm(2400,2200,132.9592)
[1] 0.05423731
```

Wytrzymałość			Liczba pró	p	n*p	chi^2
lewy	srodek	prawy	i)			
1900	1950	2000	14	0,0542373	10,84746	0,916206
2000	2050	2100	26	0,1597301	31,94602	1,106715
2100	2150	2200	52	0,2740077	54,80154	0,143219
2200	2250	2300	58	0,2470077	49,40154	1,496583
2300	2350	2400	33	0,1597301	31,94602	0,034773
2400	2450	2500	17	0,0542373	10,84746	3,489642
					SUMA	7,187139

$$\chi^2 = 7{,}187139$$

Następnie sprawdzamy obszar krytyczny:

$$R_\alpha = (\chi^2_{1-\alpha,k-r-1}, \infty)$$

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

$$\chi^2_{1-0.05,6-1-1} =^R 9.487729$$

```
> qchisq(0.95,4)
[1] 9.487729
```

Stąd nasz obszar krytyczny to: $(9.487729;\ \infty)$

Podobnie jak w poprzednim podpunkcie nie odrzucamy hipotezy, dane mają rozkład normalny z m=2200.

c) Tym razem mamy podane oba parametry naszego rozkładu:

$$m = 2200, \sigma = 100$$

Obliczenia wykonujemy analogicznie do poprzednich podpunktów:

```
> pnorm(2000,2200,100)- pnorm(1900,2200,100)
[1] 0.02140023
> pnorm(2100,2200,100)- pnorm(2000,2200,100)
[1] 0.1359051
> pnorm(2200,2200,100)- pnorm(2100,2200,100)
[1] 0.3413447
> pnorm(2300,2200,100)- pnorm(2200,2200,100)
[1] 0.3413447
> pnorm(2400,2200,100)- pnorm(2300,2200,100)
[1] 0.1359051
> pnorm(2500,2200,100)- pnorm(2400,2200,100)
[1] 0.02140023
```

Wytrzymałość			Liczba pró	p	n*p	chi^2
lewy	srodek	prawy	i)			
1900	1950	2000	14	0,0214002	4,28004	22,074
2000	2050	2100	26	0,1359051	27,18102	0,051316
2100	2150	2200	52	0,3413447	68,26894	3,876996
2200	2250	2300	58	0,3413447	68,26894	1,544643
2300	2350	2400	33	0,1359051	27,18102	1,245742
2400	2450	2500	17	0,0214002	4,28004	37,80277
					SUMA	66,59547

$$\chi^2 = 7{,}187139$$

Następnie sprawdzamy obszar krytyczny:

$$R_\alpha = (\chi^2_{1-\alpha,k-r-1}, \infty)$$

$$R_{0.05} = \left(\chi^2_{1-0.05,6-0-1}, \infty\right)$$

$$\chi^2_{1-0.05,6-0-1} =^R 11.0705$$

```
> qchisq(0.95,5)
[1] 11.0705
```

Stąd nasz obszar krytyczny to: $(11.0705;\ \infty)$.

W tym przypadku odrzucamy hipotezę zerową, ponieważ wartość należy do obszaru krytycznego.

Możemy zatem wnioskować, że dane mają rozkład normalny z parametrem m=2200, jednak nie znamy odchylenia standardowego.

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.

```
> skewness(dane)
[1] 0.1146696
```

Współczynnik asymetrii wyszedł dodatni więc mamy szanse, że jest to rozkład Gamma, Weilbull albo logarytmiczno-normalny.

Gamma:
Gdzie X-dystrybuanta empiryczna

```
> X<-c()
> for (i in dane){ s<-length(which(dane<=i))/length(dane)
+ X<-c(X,s)}
> estimate<-fitdistr(dane,"gamma")
> estimate[[1]][2]
     rate
10.22638
> estimate[[1]][1]
    shape
101.4927
> Y<-pgamma(dane, estimate[[1]][1],estimate[[1]][2])
> d0=max(abs(X-Y))
> D<-1.3581/sqrt(length(dane))
> D
[1] 0.04294689
> d0
[1] 0.02701685
```

Log-norm:

```
> estimate<-fitdistr(dane,"lognormal")
> estimate[[1]][1]
 meanlog
2.290059
> estimate[[1]][2]
      sdlog
0.09986524
> Y<-plnorm(dane, estimate[[1]][1],estimate[[1]][2])
> d0=max(abs(X-Y))
> D<-1.3581/sqrt(length(dane))
> d0
[1] 0.03198707
> D
[1] 0.04294689
```

Weilbull:

```
> estimate<-fitdistr(dane,'weibull')
Komunikaty ostrzegawcze:
1: W poleceniu 'densfun(x, parm[1], parm[2], ...)': wypro
2: W poleceniu 'densfun(x, parm[1], parm[2], ...)': wypro
> estimate[[1]][1]
   shape
10.38606
> estimate[[1]][2]
   scale
10.37202
> Y<-pweilbull(dane, estimate[[1]][1],estimate[[1]][2])
Błąd w poleceniu 'pweilbull(dane, estimate[[1]][1], estim
  nie udało się znaleźć funkcji 'pweilbull'
> Y<-pweibull(dane, estimate[[1]][1],estimate[[1]][2])
> d0=max(abs(X-Y))
> d0
[1] 0.04884945
> D<-1.3581/sqrt(length(dane))
> d
BŁĄD: nie znaleziono obiektu 'd'
> D
[1] 0.04294689
```

Rozkład	D	D0
Gamma	0,04294689	0,02701685
Log-norm	0,04294689	0,03198707
Weilbull	0,04294689	0,04884945

D>D0 – dane mają podany rozkład

Wnioski:

Dane nie mają rozkładu Weilbull.

Dane mają rozkład Gamma, bądź log-norm.

17. Zbadać, czy udziały w rynku firm A, B, C, D, E wynajmujących samochody zmieniły się, jeśli dane z dwóch lat dla prób losowych są następujące:

Rok \ Firma	A	B	C	D	E
I	39	26	18	14	3
II	29	25	16	19	11

```
> x<-c(39,26,18,14,3)
> y<-c(29,25,16,19,11)
> wilcox.test(x,y,paired=TRUE)

        Wilcoxon signed rank exact test

data:  x and y
V = 8, p-value = 1
alternative hypothesis: true location shift is not equal to 0
```

13. Wygenerować próby o liczebności 100 obserwacji według rozkładów:

i) $\mathcal{N}(900; 50)$,

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.

Dla rozkładu normalnego:

```
> sort(date)
  [1] 762.8224 795.2190 812.0676 812.5280 816.6873 817.9988
  [7] 818.7031 821.7786 826.5124 827.5425 832.5884 836.7686
 [13] 840.9068 847.9874 853.6303 853.9072 854.2297 855.2192
 [19] 856.5705 860.5095 864.1862 865.6014 867.3600 871.2605
 [25] 871.7699 871.9944 872.7814 873.0952 873.2818 874.3267
 [31] 877.5630 878.4061 878.9451 879.0116 881.8725 883.1790
 [37] 883.5506 885.7507 886.5955 887.7584 889.1169 892.5621
 [43] 894.3013 896.7187 898.7820 898.9566 900.6439 901.6194
 [49] 903.7041 905.6801 908.1926 908.7540 908.8792 909.2324
 [55] 909.5669 909.9604 912.0305 913.0430 914.3613 914.7565
 [61] 914.9882 916.5597 916.6616 919.1968 920.3461 921.0485
 [67] 922.2701 923.1088 923.2785 923.6226 923.8394 924.3442
 [73] 927.3990 931.0796 932.9407 935.7692 937.4641 938.0855
 [79] 938.7435 939.2103 939.3551 939.8678 940.4628 940.7337
 [85] 941.1653 943.4416 947.0704 947.1939 948.4291 949.2159
 [91] 950.2742 951.9223 955.0377 958.4702 960.4376 960.9944
 [97] 963.8929 983.3848 987.0670 998.5160
```

Dla rozkładu trójkątnego:

```
> date_t<-rtri(100,725,1075,(1075-725)/2+725)
> sort(date_t)
  [1]  739.6934  763.0699  766.8297  770.6123  772.4035  776.9512
  [7]  777.0574  779.0585  780.2686  788.6298  794.6963  795.1362
 [13]  796.6256  797.4395  798.2365  799.6118  804.5042  807.5155
 [19]  808.3255  808.3343  811.0745  821.0158  823.0435  825.7199
 [25]  826.3051  829.3342  830.6366  833.6097  835.5920  840.9048
 [31]  841.3055  848.9529  850.5779  852.8613  854.1918  855.6240
 [37]  855.9167  858.2456  860.2985  860.7450  861.9434  863.3612
 [43]  865.4240  867.0157  867.0387  868.0315  868.1484  870.5039
 [49]  871.9375  874.8073  875.2354  875.6995  878.2417  878.7602
 [55]  880.3122  883.2170  883.3158  884.0816  890.5218  891.3411
 [61]  891.3832  892.8608  898.6297  901.4584  901.4690  902.0057
 [67]  903.5756  905.0158  908.6303  909.9223  915.8190  916.9651
 [73]  917.1613  918.8422  928.3995  932.3522  952.0288  960.4119
 [79]  960.7768  961.5987  962.3628  963.2061  968.2601  980.1751
 [85]  980.2772  983.0821  983.4616  984.3403  984.3847  984.4152
 [91] 1000.6156 1011.4155 1013.3454 1014.7753 1016.8237 1025.8685
 [97] 1029.4746 1032.2460 1034.8860 1044.5740
```

a) Korzystam z wspomagania komputerowego – język R:

i)

```
> date<-rnorm(100,900,50)
> mean(date)
[1] 898.6022
> var(date)
[1] 2060.591
> sqrt(var(date))
[1] 45.39373
```

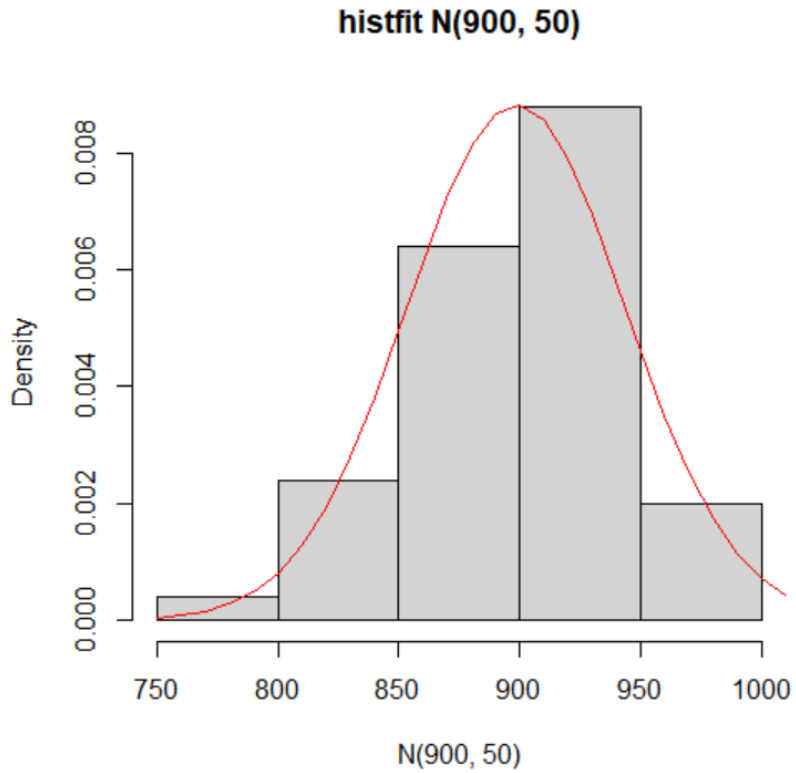
$$\bar{X} = 898{,}6022\,, \sigma^2 = 2060{,}591;\ \ \sigma = 45{,}39373$$

ii)

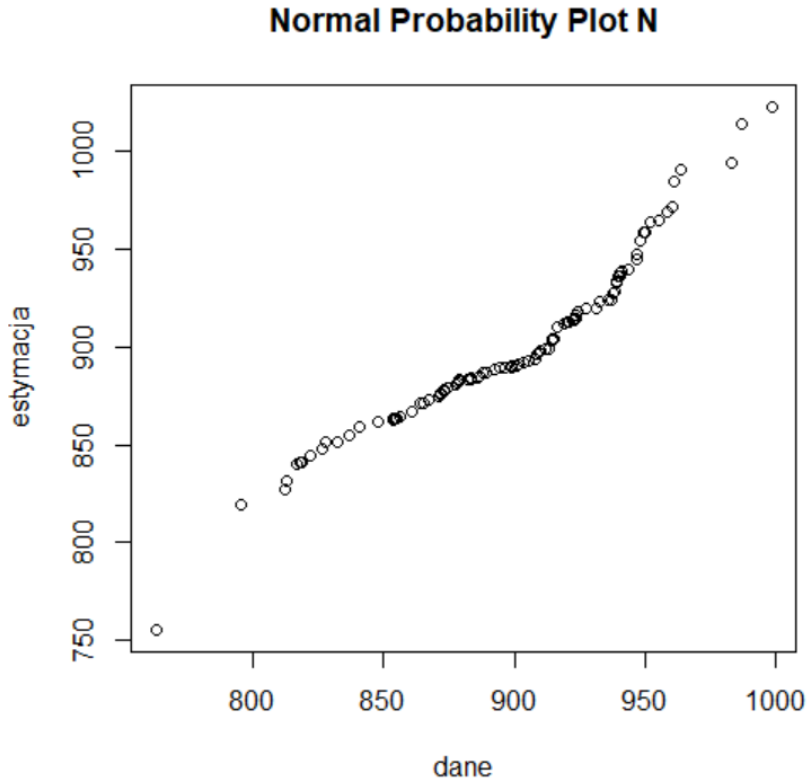
```
> mean(date_t)
[1] 884.1719
> var(date_t)
[1] 5633.238
> sqrt(var(date_t))
[1] 75.0549
```

$$\bar{X} = 884{,}1719\,, \sigma^2 = 5633.238\,;\ \ \sigma = 75.0549$$

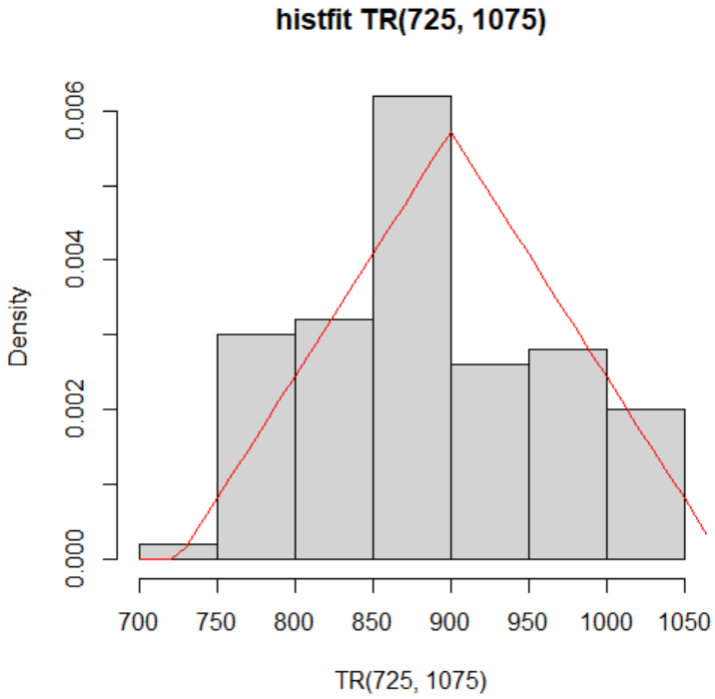
b) Następnie generuje wykresy:



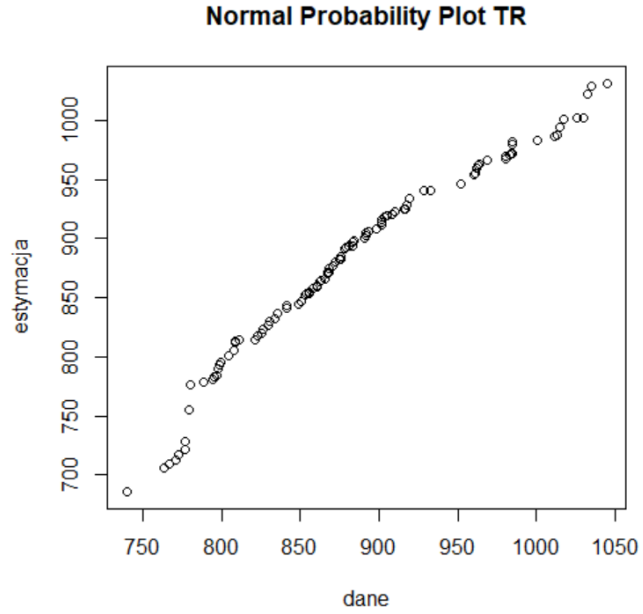

```
hist(date, prob = TRUE, main = "histfit N(900, 50)", xlab = "N(900, 50)")
lines(seq(750, 1050, by = 10), dnorm(seq(750, 1050, by = 10), n$estimate[[1]], n$estimate[[2]]), col = "red")
```



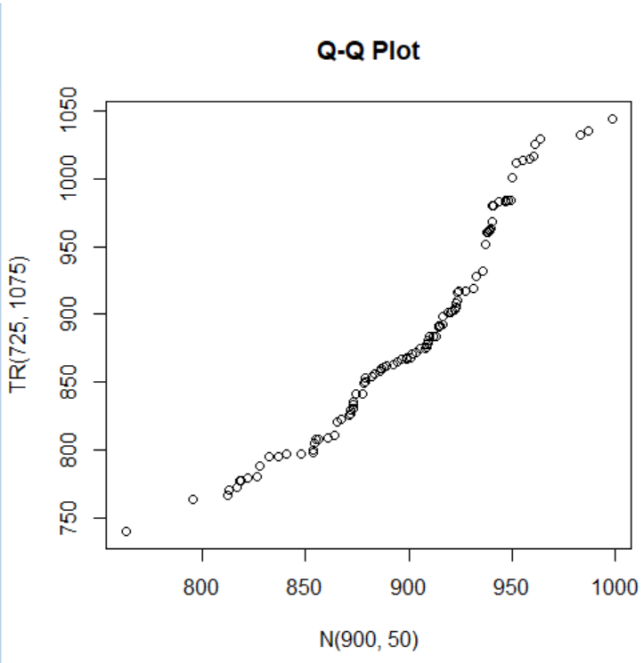

```
qqplot(date, rnorm(100, n$estimate[[1]], n$estimate[[2]]), xlab = "dane", ylab = "estymacja", main = "Normal Probability Plot N")
```



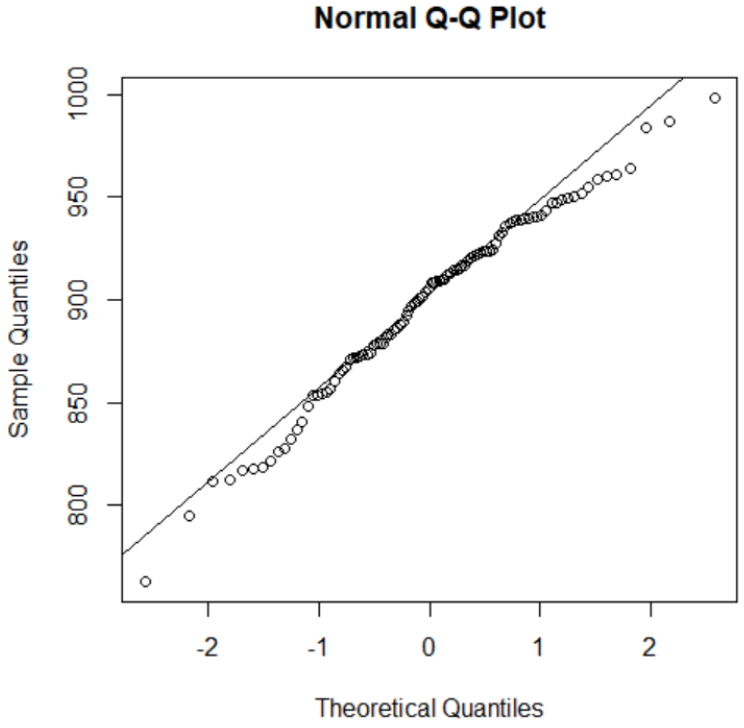

```
hist(date_t, prob = TRUE, main = "histfit TR(725, 1075)", xlab = "TR(725, 1075)")
lines(seq(700, 1100, by = 10), dtri(seq(700, 1100, by = 10), 725,1075,(1075-725)/2 + 725), col = "red")
```




```
qqplot(date_t, rnorm(100, n$estimate[[1]], n$estimate[[2]]), xlab = "dane", ylab = "estymacja", main = "Normal Probability Plot TR")
```




```
qqplot(date, date_t, xlab = "N(900, 50)", ylab = "TR(725, 1075)", main = "Q-Q Plot")
```




```
qqnorm(date)
qqline(date)
```

f)

```
> ks.test(date,date_t, alternative='two.sided')

        Two-sample Kolmogorov-Smirnov test

data:  date and date_t
D = 0.25, p-value = 0.003861
alternative hypothesis: two-sided
```