Estymacja parametryczna

1. Średnie wynagrodzenie 50 losowo wybranych programistów wyniosło 6000 zł. Wiadomo, że odchylenie standardowe wynagrodzenia programistów wynosi 2100 zł. Wyznacz 95% przedział ufności dla średniego wynagrodzenia programistów, zakładając, że rozkład ich wynagrodzeń jest rozkładem normalnym.

X - wynagrodzenie programistów w zł

Zakładamy, że X ~N(mu, sigma)

sigma=2100
średnia w próbie = 6000
n = 50
1-alpha =0.95, alpha=0.05

Wyznaczyć: 95%CI dla średniego wynagrodzenia wszystkich programistów

ESTYMACJA PARAMETRYCZNA

Przedziały ufności dla średniej μ na poziomie ufności $1-\alpha$

Model 1 $X \sim N(\mu, \sigma)$, σ znane - funkcja `z.test(x)` w R lub ze wzoru $\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Model 2 $X \sim N(\mu, \sigma)$, σ nieznane - funkcja `t.test(x)` w R lub ze wzoru $\left[\bar{x} - t^{n-1}_{1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t^{n-1}_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

Model 3 $X \sim$ rozkład dowolny $(n > 25)$ - funkcja `t.test(x)` w R lub ze wzoru $\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

Model 1 (rozkład normalny, sigma jest znane, sigma = 2100)

```
# 95%CI
z <- qnorm(1-0.05/2)
6000 - z*2100/sqrt(50); 6000 + z*2100/sqrt(50)
```

Odp. **[5417.92, 6582.08]** zł

Interpretacja: Mamy 95% pewność, że średnia wynagrodzenie wszystkich programistów znajdzie się w przedziale od 5417.9 do 6582.1zł.

Czy 99%CI będzie szerszy czy węższy?

```
z <- qnorm(1-0.01/2)
6000 - z*2100/sqrt(50); 6000 + z*2100/sqrt(50)
```

Odp. szerszy **[5235.02, 6764.98]** zł

Z 99% prawdopodobieństwem średnia wynagrodzenia znajdzie się w przedziale od 5235 do 6765 zł.

2. Dla wybranego użytkownika zarejestrowano czasy między naciśnięciami klawiszy, gdy wpisywał login i hasło. Pobrano z nich losową próbę 18 pomiarów (w sekundach):

$$0.24, 0.22, 0.26, 0.34, 0.35, 0.32, 0.33, 0.29, 0.19, 0.36, 0.30, 0.15, 0.17, 0.28, 0.38, 0.40, 0.37, 0.27.$$

Zakładając, że czasy pochodzą z rozkładu normalnego, wyznacz

a) 99% przedział ufności dla średniego czasu między naciśnięciami klawiszy tego użytkownika,

b) 95% przedział ufności dla odchylenia standardowego czasu między naciśnięciami klawiszy tego użytkownika.

Przedziały ufności dla średniej μ na poziomie ufności $1-\alpha$

Model 1 $X \sim N(\mu, \sigma)$, σ znane - funkcja `z.test(x)` w R lub ze wzoru $\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Model 2 $X \sim N(\mu, \sigma)$, σ nieznane - funkcja `t.test(x)` w R lub ze wzoru $\left[\bar{x} - t^{n-1}_{1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t^{n-1}_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

Model 3 $X \sim$ rozkład dowolny $(n > 25)$ - funkcja `t.test(x)` w R lub ze wzoru $\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

Przedziały ufności dla wariancji σ^2 na poziomie ufności $1-\alpha$

Model 1 $X \sim N(\mu, \sigma)$, σ nieznane - funkcja `sigma.test(x)` w R lub ze wzoru $\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}},\ \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$

Model 2 $X \sim$ rozkład dowolny $(n > 25)$ - własna funkcja w R lub ze wzoru $\left[\frac{s^2(2n-2)}{(\sqrt{2n-3}+z_{1-\alpha/2})^2},\ \frac{s^2(2n-2)}{(\sqrt{2n-3}-z_{1-\alpha/2})^2}\right]$

a) 99%CI dla średniej

X - czas między naciśnięciami klawiszy [s]
Zakładamy, że: X~N(mi, sigma),
sigma - nieznane
n=18

```
czas <- c(.24,.22,.26,.34,.35,.32,.33,.29,.19,.36,
          .30,.15,.17,.28,.38,.40,.37,.27)
```

Model 2 (X~N(mi, sigma), sigma nieznane)

```
t.test(czas, conf.level = 0.99)$conf.int
```

odp. **[0.2394741, 0.3405259]**

b) 95%CI dla odchylenia standardowego

Model 1 dla odchylenia standardowego (X~N(mi, sigma), sigma nieznane)

```
library(TeachingDemos)
sigma.test(czas, conf.level = 0.95)$conf.int          # zwraca przedział dla wariancji!
sqrt(sigma.test(czas, conf.level = 0.95)$conf.int)
```

odp. **[0.05550124, 0.11088182]**

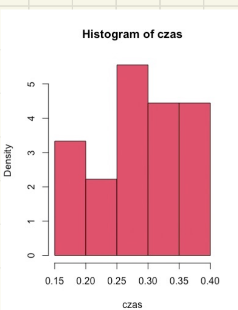
Uwaga 1 Możemy też wyestymować rozrzut wokół średniej (**reguła 3-sigma**):

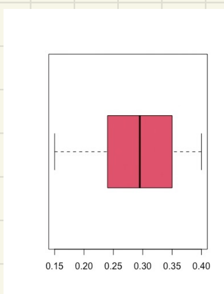
```
# rozrzut wokół średniej większości czasów
# [srednia +/- odch.std.] - typowy obszar zmienności (pokrywający 66,7% obs. w przypadku normalności rozkładu)
 mean(czas)-sd(czas); mean (czas)+sd(czas)
[0.2160366, 0.3639634]

# [srednia +/- 2*odch.std.]  - (pokrywa 95%, w przypadku r. norm)
 mean(czas)-2*sd(czas); mean(czas)+2*sd(czas)
[0.1420732, 0.4379268]

 # [srednia +/- 3*odch.std.]  - (pokrywa 99,7%, w przypadku r. norm)
 mean(czas)-3*sd(czas); mean(czas)+3*sd(czas)
[0.06810973, 0.5118903]
```

Uwaga 2. Czy założenie o normalności rozkładu było zasadne? Możemy to sprawdzić na 3 sposoby:

1) Sprawdzamy kształt rozkładu i relację średniej i mediany:

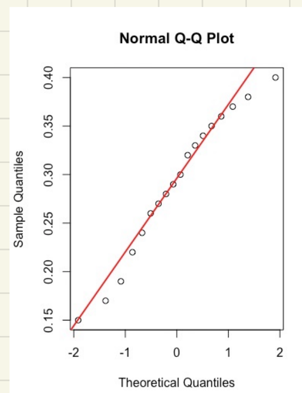
```
# Czy rozkład jest symetryczny? Czy me blisko m?
Czy me wpada do przedziału dla średniej?
hist(czas, prob=T, col=2)
boxplot(czas, horizontal=TRUE,col=2)
summary(czas)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1500  0.2450  0.2950  0.2900  0.3475  0.4000
```





2) Porównujemy kwantyle z próbki z "modelowymi" dla r. normalnego

```
# Q-Q plot
qqnorm(czas)
qqline(czas, col="red",lwd=2)
```



3) Test normalności

H0: X ma rozkład normalny
H1: X nie ma r. normalnego

```
shapiro.test(czas)

	Shapiro-Wilk normality test

data:  czas
W = 0.96112, p-value = 0.6233
```

# # **p-value = 0.6233 > alpha**
>>> Można przyjąć H0, że rozklad jest normalny

3. Zmierzono czas świecenia 69 świetlówek i stwierdzono, że dla 14 z nich był on krótszy niż 1000 godzin, dla 15 był w przedziale [1000, 2000), dla 29 świetlówek był dłuższy niż 2000, ale krótszy niż 3000 godzin, zaś dla pozostałych 11 - czas świecenia był dłuższy niż 3000, ale nie dłuższy niż 4000 godzin. Oszacuj przedziałowo średnią i odchylenie standardowe czasu świecenia świetlówek. Przyjmij poziom ufności 0.95.

dane zagregowane (**szereg rozdzielczy**):

liczność	czas_świecenia
14	0 - 1000
15	1000 - 2000
29	2000 - 3000
11	3000 - 4000

X - czas świecenia świetlówek, X~nieznany rozkład, n=69

```
srodki <- c(500, 1500, 2500, 3500)
licznosci <- c(14, 15, 29, 11)
n <- sum(licznosci)
```

Przedziały ufności dla średniej μ na poziomie ufności $1-\alpha$

Model 1 $X \sim N(\mu, \sigma)$, σ znane - funkcja `z.test(x)` w R lub ze wzoru $\left[\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

Model 2 $X \sim N(\mu, \sigma)$, σ nieznane - funkcja `t.test(x)` w R lub ze wzoru $\left[\bar{x} - t^{n-1}_{1-\alpha/2}\frac{s}{\sqrt{n}},\ \bar{x} + t^{n-1}_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

Model 3 $X \sim$ rozkład dowolny $(n > 25)$ - funkcja `t.test(x)` w R lub ze wzoru $\left[\bar{x} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{s}{\sqrt{n}}\right]$

Przedziały ufności dla wariancji σ^2 na poziomie ufności $1-\alpha$

Model 1 $X \sim N(\mu, \sigma)$, σ nieznane - funkcja `sigma.test(x)` w R lub ze wzoru $\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}},\ \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}\right]$

Model 2 $X \sim$ rozkład dowolny $(n > 25)$ - własna funkcja w R lub ze wzoru $\left[\frac{s^2(2n-2)}{(\sqrt{2n-3}+z_{1-\alpha/2})^2}, \frac{s^2(2n-2)}{(\sqrt{2n-3}-z_{1-\alpha/2})^2}\right]$

```
# średnia w szeregu rozdzielczym
srednia <- sum(srodki*licznosci)/n
2036.232
```

95% CI dla średniej: **Model 3** (rozkład nieznany, n=69>25)

```
alpha<-0.05
srednia - qnorm(1-alpha/2)*S/sqrt(n); srednia + qnorm(1-alpha/2)*S/sqrt(n)
```

Odp. **[1801.743, 2270.721]**

```
# odch.std. w szeregu rozdzielczym
S <- sqrt(sum(licznosci*(srodki-srednia)^2)/(n-1))
 993.8
```

95% CI dla odch. std. : **Model 2** (rozkład nieznany, n=69>25)

```
sqrt(2*n-2)*S/(sqrt(2*n-3)+qnorm(1-alpha/2))
sqrt(2*n-2)*S/(sqrt(2*n-3)-qnorm(1-alpha/2))
```

Odp. **[853.4998, 1199.878]**

```
# można też stworzyć funkcje
CI_S_M2 <- function(n, S, alpha=0.05){
  L <- sqrt(2*n-2)*S/(sqrt(2*n-3)+qnorm(1-alpha/2))
  P <- sqrt(2*n-2)*S/(sqrt(2*n-3)-qnorm(1-alpha/2))
  c(L, P)
}

CI_S_M2(69, S, 0.05)
```

4. Ramka danych *faithful* zawiera dane dotyczące czasu trwania erupcji gejzera Old Faithful (zmienna *eruptions*) oraz czasu oczekiwania na kolejną erupcję (zmienna *waiting*). Utwórz 99% przedział ufności dla średniego czasu oczekiwania na kolejną erupcję.

```
data(faithful)
dim(faithful)
[1] 272   2

summary(faithful)
  eruptions       waiting
Min.   :1.600   Min.   :43.0
1st Qu.:2.163   1st Qu.:58.0
Median :4.000   Median :76.0
Mean   :3.488   Mean   :70.9
3rd Qu.:4.454   3rd Qu.:82.0
Max.   :5.100   Max.   :96.0
```

X - czas oczekiwania na kolejną erupcję
X~rozkład nieznany, n=272 >> **Model 3**

t.test(faithful$waiting, conf.level = 0.99)$conf.int
Odp. 99% CI dla średniej **[68.75871, 73.03541]**

Uwaga 1: Funkcja t. test() jest wystarczająco dobrym przybliżeniem tylko, gdy n jest duże (n>100)

Formalnie należy użyć wzoru:

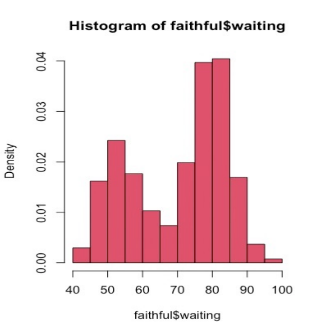
```
mean(faithful$waiting) - qnorm(1-0.01/2)*sd(faithful$waiting)/sqrt(272);
mean(faithful$waiting) + qnorm(1-0.01/2)*sd(faithful$waiting)/sqrt(272)
```

Odp. 99% CI dla średniej **[68.77376, 73.02036]**

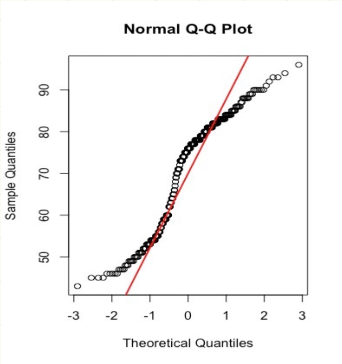
Duża próbka n=272, więc Model 3 działa.

Ale gdybyśmy sprawdzili normalność rozkładu (bo me=76, m=70.9, różnica jest spora, oraz me nie należy do [68.77376, 73.02036]).

1) hist(faithful$waiting,prob=T,col=2)



2) qqnorm(faithful$waiting)
qqline(faithful$waiting, col=2,lwd=2)



3) shapiro.test(faithful$waiting)

\# H0: X ma rozkład normalny
\# H1: X nie ma rozkładu N
\# p-value = 1.015e-10 < alpha >>> H1 (**r. nie jest norm.**), wiec cały czas **Model 3**

5. Ramka danych *Pima.te* z pakietu *MASS* zawiera dane dotyczące zdrowia kilkuset Indianek z plemienia Pima mających co najmniej 21 lat. Zmienna *type* zawiera informację, czy kobieta jest chora na cukrzycę, czy nie.

a) Utwórz 95% przedział ufności dla odsetka Indianek dotkniętych cukrzycą.

b) Utwórz 95% przedział ufności dla odsetka Indianek dotkniętych cukrzycą mających co najmniej 35 lat.

```
library(MASS)
data(Pima.te)
View(Pima.te)
```

Przedział ufności dla odsetka (procentu) p na poziomie ufności $1-\alpha$

Model $X \sim Bern(p)$, p - nieznane - funkcje `binom.test(k,n)`, `prop.test(k,n)` lub (gdy $np > 5, n(p-1) > 5$) ze wzoru $\left[\hat{p} - z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]$, gdzie $\hat{p} = \frac{k}{n} = \frac{\text{liczba sukcesów}}{\text{liczność próby}}$

(a)

X -cecha (0/1), X=1 dotyczy Indianki z cukrzycą

X~Bern(p), p-nieznane,

p oznacza prawdopodobieństwo sukcesu p=P(X=1)

```
k <- sum(Pima.te$type == 'Yes')
109
n <- length(Pima.te$type)
 332
k/n
```

estymator p = 0.3283133

95% CI dla odsetka:

```
binom.test(k, n, conf.level = 0.95)$conf.int
```

odp. **[0.2780256, 0.3816971]**

```
prop.test(k, n, conf.level = 0.95)$conf.int
```

odp. **[0.2785847, 0.3820858]**

czyli mamy 95% pewności, że odsetek kobiet chorych na cukrzycę w tej populacji wynosi od 27.9% do 38.2%.

(b)

Y -cecha (0/1), Y=1 dotyczy Indianki 35+ z cukrzycą

Y~Bern(p), p-nieznane

p oznacza prawdopodobieństwo sukcesu p=P(X=1)

```
k <- sum(Pima.te$type == 'Yes' & Pima.te$age >= 35)
53
n <- sum(Pima.te$age >= 35)
102
k/n
```

estymator p = 0.5196078

95% CI dla odsetka:

```
binom.test(k, n, conf.level = 0.95)$conf.int
```

odp. **[0.4184415, 0.6196060]**

```
prop.test(k, n, conf.level = 0.95)$conf.int
```

odp. **[0.4189563, 0.6187654]**

czyli u starszych kobiet ryzyko zachorowania jest większe niż dla ogółu.

7. Jak dużą próbę należy pobrać, aby z maksymalnym błędem ±2% oszacować na poziomie ufności 0.99 procent kierowców, którzy nie zapinają pasów bezpieczeństwa? Uwzględnij rezultaty wstępnych badań, z których wynika, że interesująca nas wielkość jest rzędu 16%. Porównaj otrzymaną liczność próby z licznością, jaka byłaby wymagana, gdyby pominąć rezultaty wstępnych badań.

Minimalna liczność próby do oszacowania odsetka p na poziomie ufności $(1-\alpha)$ z max. błędem d

Model 1 Jeśli znany jest szacunkowy procent p_0 - ze wzoru $n \geq \frac{p_0(1-p_0)}{d^2} z^2_{1-\alpha/2}$

Model 2 Jeśli nie jest znany szacunkowy procent p_0 - ze wzoru $n \geq \frac{1}{4d^2} z^2_{1-\alpha/2}$

Pierwszy przypadek:

X - cecha (0/1), gdzie X=1 dotyczy kierowcy nie zapinającego pasów
X~Bern(p), p0-znane, p0=0.16

```
d <- 0.02
alpha <- 0.01
p0 <- 0.16
minN <- p0*(1-p0)*(qnorm(1-alpha/2)/d)^2          # znane p0 (Model 1)
minN
2229.325
ceiling(minN)
odp. 2230
```

Minimalna liczność próby, potrzebna do oszacowania na poziomie ufności 99%, z błędem d=2% odsetka kierowców, którzy nie zapinają pasów bezpieczeństwa, wynosi 2230.

Drugi przypadek:

X~Bern(p), p0-nieznane
Czy przy p0 nieznanym potrzebna będzie liczność większa czy mniejsza?

```
minN <- (qnorm(1-alpha/2)/(2*d))^2          # nieznane p0 (Model 2)
minN
 4146.81
ceiling(minN)
odp. znacznie większa  4147
```

Po pominięciu rezultatów wstępnych badań, minimalna wielkość próby wzrasta niemal dwukrotnie, do 4147.

8. Poniższe dane przedstawiają zarejestrowaną przez radar drogowy prędkość 10 losowo wybranych pojazdów, jadących pewną autostradą (km/h):

$$106, 115, 99, 109, 122, 119, 104, 125, 107, 111.$$

Zakładając normalność rozkładu prędkości, wyznacz liczność próby potrzebną do wyestymowania średniej prędkości z dokładnością ± 2 km/h na poziomie ufności 0.95.

Minimalna liczność próby do oszacowania średniej μ na poziomie ufności $(1-\alpha)$ z max. błędem d

Model 1 $X \sim N(\mu, \sigma)$, σ znane - ze wzoru $n \geq \left(\frac{\sigma}{d} z_{1-\alpha/2}\right)^2$

Model 2 $X \sim N(\mu, \sigma)$, σ nieznane - ze wzoru $n \geq \left(\frac{s}{d} t_{1-\alpha/2}^{n_0-1}\right)^2$, gdzie n_0 - liczność pobranej próby wstępnej

X - prędkość aut na autostradzie przy radarze [km/h]
Zakładamy, że X~N(mu, sigma),
sigma nieznane

Model 2 (rozkład normalny, sigma nieznane, n0=10)

```
x <- c(106, 115, 99, 109, 122, 119, 104, 125, 107, 111)

S <- sd(x)
 8.367264

d <- 2
n0 <- length(x)          # liczność pobranej próby wstępnej n0=10
alpha <- 0.05

minN <- (S*qt(1-alpha/2, n0-1)/d)^2
minN
89.56793

ceiling(minN)
```

odp. **90**

Minimalna liczność próby potrzebna do wyestymowania średniej prędkości z dokładnością d=2 km/h, na poziomie ufności 95% wynosi 90.