Laboratorium 3

Martyna Konopacka

Zadanie 1

Przedział ufności dla wartości oczekiwanej na poziomie $1 - \alpha$ jest postaci

$$(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

gdzie $z_{1-\alpha/2}$ jest kwantylem rzędu $1 - \alpha/2$ rozkładu standardowego (dla poziomu ufności 95% mamy $z_{\alpha/2} = 1.96$), $\bar{X}$ to średnia z próby. W zadaniu zakładamy, że odchylenie standardowe jest znane i wynosi $\sigma = 1$.

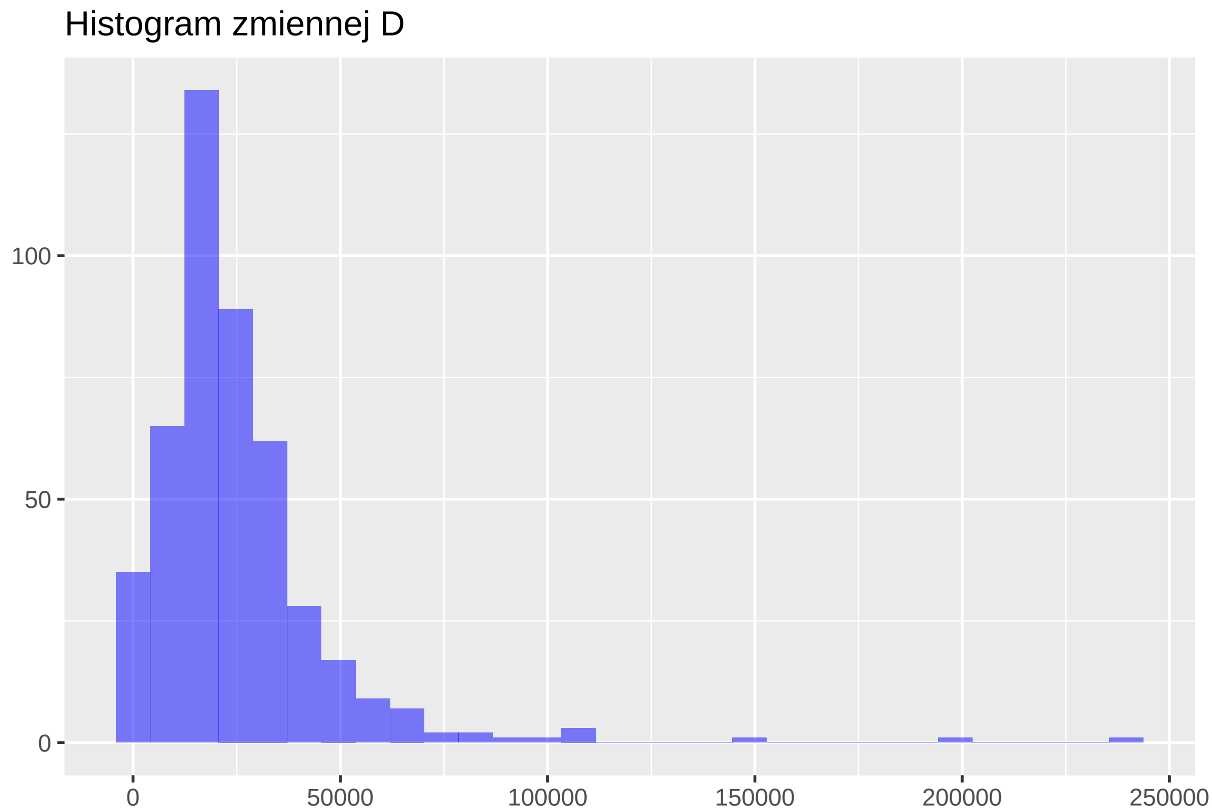
Celem eksperymentu jest sprawdzenie, w jakim odsetku powtórzeń eksperymentu średnia populacji $\mu = 0$ rzeczywiście mieści się w wyznaczonym przedziale.

Table 1: Wyniki dla podpunktow b i c

mu	sd	n	poziom	iter	trafienia	sr_szer_przedzialu
0	1	100	0.95	1000	0.956	0.3919928
0	1	200	0.95	1000	0.945	0.2771808

W obu przypadkach odsetki są bardzo zbliżone do wyznaczonych teoretycznie wartości - jest tak dlatego, że n powyżej 30 jest już wystarczająco duże. Dzięki sprawdzeniu średnich szerokości przedziałów można zauważyć, jak zwiększenie liczebności próby wpływa na zawężenie przedziału.

Zadanie 2



Patrząc na histogram można domyślać się, że rozkład nie będzie normalny - w szczególności normalność psują obserwacje odstające z prawej strony wykresu. W części pierwszej sprawdzimy, czy odsetki obserwacji zmiennej D w odległościach $\sigma, 2\sigma, 3\sigma$ od średniej są zbliżone do wartości wynikających z reguły trzech sigm. Odchylenie standardowe w populacji z zadania wynosi 22087.09, a średnia 24977.43.

```
## [1] 0.6800000 0.8689956
```

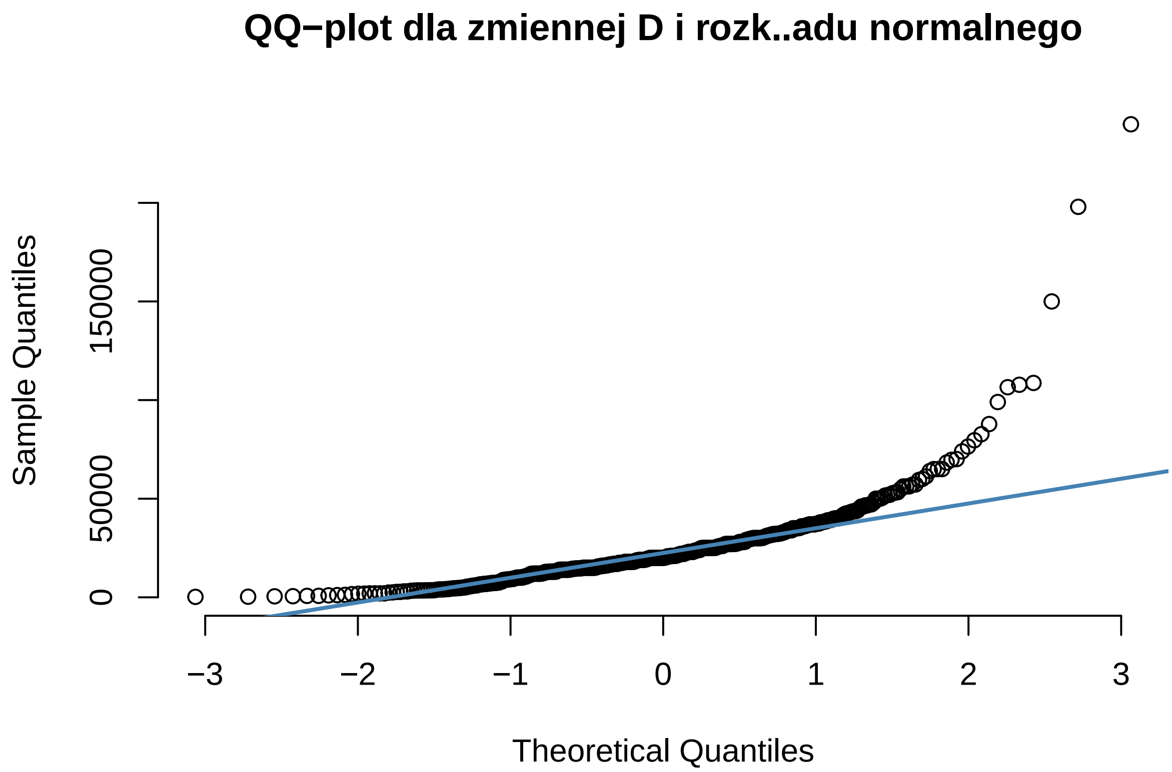
```
## [1] 0.9500000 0.9694323
```

```
## [1] 0.9970000 0.9847162
```

Największą różnicę widać dla pierwszego przedziału - leży w nim znacznie większy odsetek obserwacji, niż wynikałoby z zasady trzech sigm. Jest tak dlatego, że przez udział obserwacji odstających przedział ten jest szerszy, niż powinien.

```
## Warning in title(...): conversion failure on 'QQ-plot dla zmiennej D i rozkładu
## normalnego' in 'mbcsToSbcs': dot substituted for <c5>
```

```
## Warning in title(...): conversion failure on 'QQ-plot dla zmiennej D i rozkładu
## normalnego' in 'mbcsToSbcs': dot substituted for <82>
```



Na wykresie można zauważyć obserwacje odstające zaburzające normalność rozkładu. Bliżej środka punkty leżą blisko prostej.

Zadanie 3a

Lemat

Niech $D = U^2$, $E[U] = \mu$ oraz $Var[U] = \sigma^2$. Wtedy $E[D] = \mu^2 + \sigma^2$. Dowód:

$$E[D] = E[U^2] = E[U - \mu + \mu]^2 = E[U - \mu]^2 - 2E[(U - \mu)\mu] + E[\mu]^2 = Var[U] - 2\mu E[U - \mu] + \mu^2 = \sigma^2 + \mu^2$$

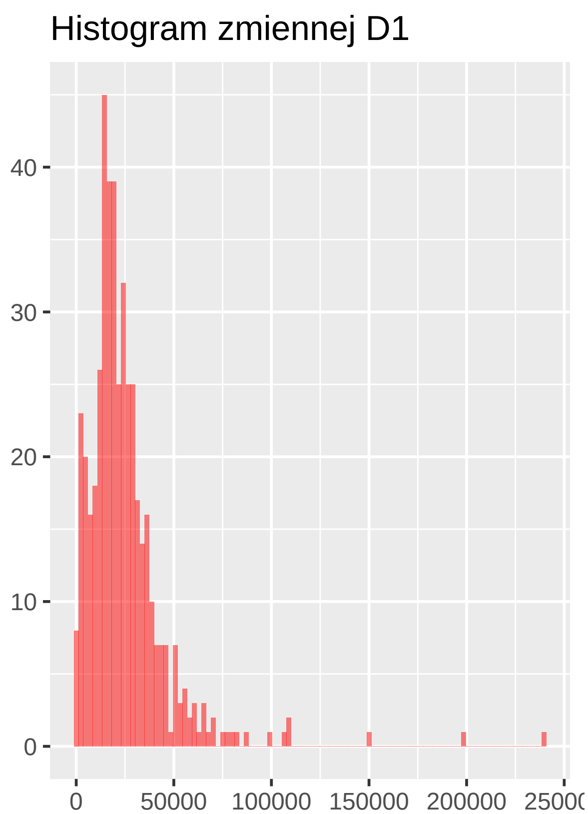
Stąd $E[D] - (E[U])^2 = \sigma^2 + \mu^2 - \mu^2 = \sigma^2 \geq 0$.

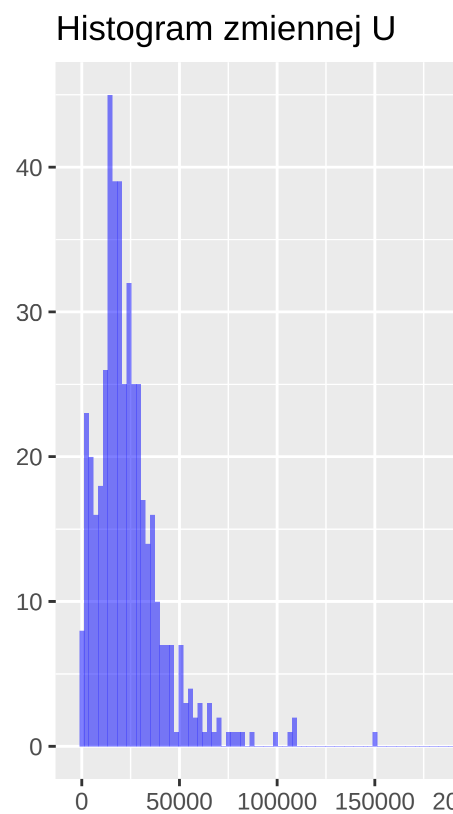
W celu obliczenia pierwiastka, najpierw przesuwamy cały wektor D o jego najniższą wartość - w ten sposób unikniemy problemu z wartościami ujemnymi. średnia D_1 powinna być większa niż kwadrat średniej U o wartość wariancji U. Uwaga: wbudowana funkcja `var` oblicza nieobciążony estymator wariancji z $n-1$ w mianowniku - w tym przypadku chodzi nam jednak o wariancję liczoną "dla populacji", czyli `mean((U-mean(U))^2` i tylko taki sposób daje dobre wyniki.

Ponieważ dane zostały przesunięte, w podpunkcie a. porównałam kwadrat średniej $U = \sqrt{D_1}$ ze średnią zmiennej D_1. Różnica pomiędzy średnimi $D1$ i D jest równa początkowemu przesunięciu.

```
##   sq_mean_U  mean_D1  sigma_U diff_mD1_mU2
## 1  21874.92 25207.43 3332.503     3332.503
```

W dalszej części polecenia rysujemy histogram zmiennej U - dla porównania obok zamieściłam histogram D_1, jednak okazało się, że różnica w skali tych danych jest zbyt mała, żeby była widoczna na takim wykresie, nawet





dla dużej liczby przedziałów.
Ostatnią częścią podpunktu jest wyznaczenie frakcji osób z wykształceniem wyższym - zadanie realizuje poniższy kod:

```
pw <- nrow(filter(income, Education >= 5))/nrow(income)
pw
```

```
## [1] 0.268559
```

Zadanie 3b

Estymatorem μ jest średnia próbkowa, a estymatorem p proporcja w próbie. Konstrukcja przedziału ufności dla μ została opisana w poprzednich zadaniach. Klasyczny przedział ufności dla p w próbie o wystarczająco dużej liczebności (pozwalającej na przybliżanie rozkładem normalnym, powiedzmy $n > 30$) jest postaci:

$$(\bar{p} - u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}}), \bar{p} + u_{1-\frac{\alpha}{2}}\sqrt{\frac{\bar{p}(1-\bar{p})}{n}})$$

gdzie $\bar{p}$ jest estymatorem p.

```
przedzial_mu <- function(V, sd_pop, poziom = 0.95){
  # Funkcja zwraca lewy i prawy brzeg przedziału ufności dla wartosci oczekiwanej; znany parametr sd_po
  s <- sd_pop
  meanX <- mean(V)
  n <- length(V)
  u <- qnorm(p = 1 - (1-poziom)/2)
  low <- meanX - u*s/sqrt(n)
  high <- meanX + u*s/sqrt(n)
  return(c(low = low, high = high))
```

```
}

przedzial_p <- function(V, poziom = 0.95){
  # Funkcja zwraca lewy i prawy brzeg przedziału ufności dla p; V powinien być wektorem TRUE / FALSE
  p <- sum(V)/length(V) # estymator p
  n <- length(V)
  u <- qnorm(p = 1 - (1-poziom)/2)
  low <- p - u*sqrt((p*(1-p)/n))
  high <- p + u*sqrt((p*(1-p)/n))
  return(c(low = low, high = high))
}
```

Sprawdzimy, czy przedziały konstruowane o próbę o liczebności 200 zawierają rzeczywiste wartości parametrów:

```
##        low real_mu_U     high inside
## 1 143.6264  147.9017 159.6275   TRUE
```

```
##        low real_mu_D     high inside
## 1 22956.83  24977.43 26998.02   TRUE
```

```
##         low   real_p      high inside
## 1 0.2224383 0.268559 0.3475617   TRUE
```

Wszystkie przedziały zawierały prawdziwe wartości parametru. Szanse na to w każdym z przypadków wynosiły 95%.

Zadanie 3c

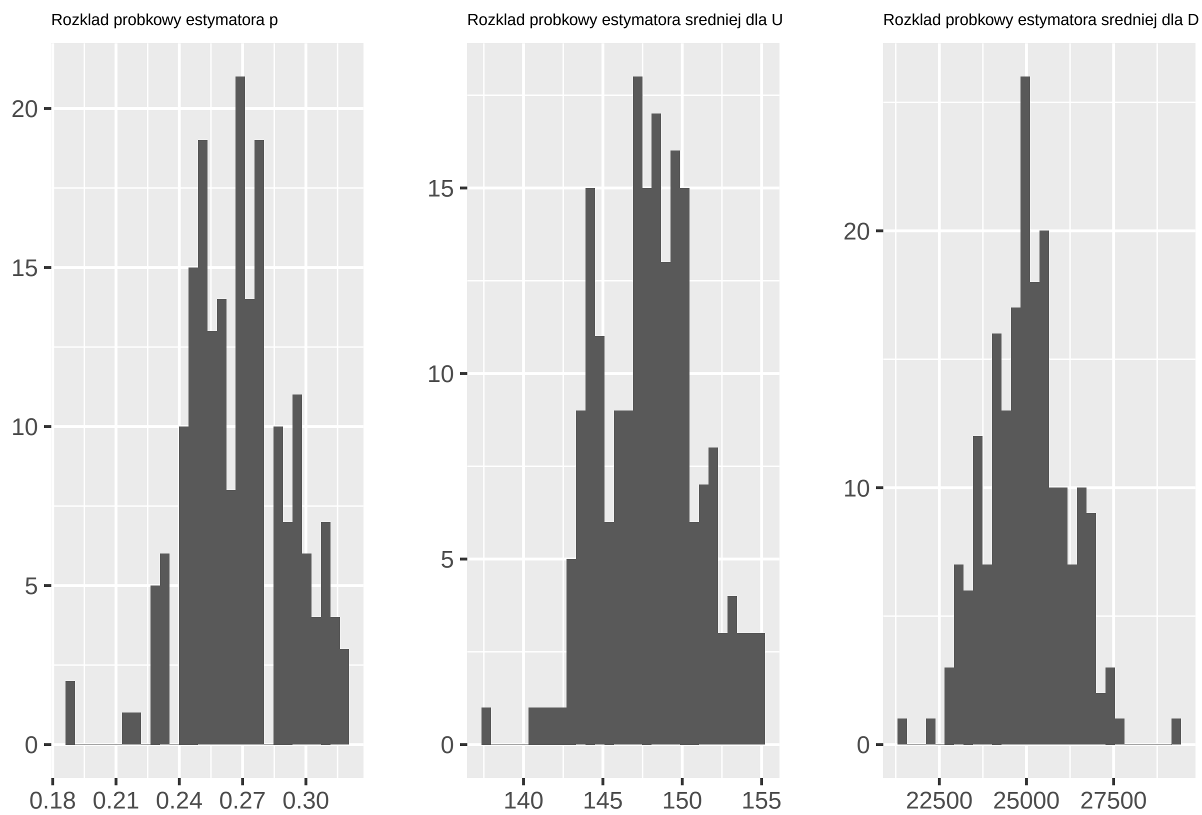
```
it = 200
est_p <- numeric(length = it)
est_mu_U <- numeric(length = it)
est_mu_D <- numeric(length = it)

# liczniki: ile razy prawdziwy parametr zawiera sie w przedziale?
zawiera_p <- 0
zawiera_U <- 0
zawiera_D <- 0

for (i in 1:it){
  # pobranie prób
  sampleV <- sample(Vp,200)
  sampleU <- sample(U,200)
  sampleD <- sample(D,200)
  # obliczenie estymatorow
  p <- sum(sampleV)/200
  mu_U <- mean(sampleU)
  mu_D <- mean(sampleD)
  # zapisanie ich do wektorow w celu pozniejszego rysowania histogramu
  est_p[i] <- p
  est_mu_U[i] <- mu_U
  est_mu_D[i] <- mu_D
  # sprawdzenie przedzialow
  pp <- przedzial_p(sampleV)
  pU <-przedzial_mu(sampleU, sd_U)
  pD <- przedzial_mu(sampleD, sd_D)
```

```
  if (between(pw, pp["low"], pp["high"])){zawiera_p <- zawiera_p + 1}
  if (between(mean(U), pU["low"], pU["high"])){zawiera_U <- zawiera_U + 1}
  if (between(mean(D), pD["low"], pD["high"])){zawiera_D <- zawiera_D + 1}
}

p1 <- qplot(est_p, geom = 'histogram', main = 'Rozklad probkowy estymatora p', xlab = '') + theme(plot.t
p2 <- qplot(est_mu_U, geom = 'histogram', main = 'Rozklad probkowy estymatora sredniej dla U', xlab = '
p3 <- qplot(est_mu_D, geom = 'histogram', main = 'Rozklad probkowy estymatora sredniej dla D', xlab = '
grid.arrange(p1,p2,p3, ncol = 3)
```



Rozkłady przypominają rozkład normalny. W dalszej części zadania sprawdzamy jeszcze, jak często przedziały ufności zawierały prawdziwy parametr:

```
data.frame(frakcja = zawiera_p / 200, srednia_U = zawiera_U / 200, srednia_D = zawiera_D / 200)
```

```
##   frakcja srednia_U srednia_D
## 1    0.99     0.995      0.99
```

Odsetki są znacząco wyższe niż teoretyczne - myślę, że przedział jest błędnie skonstruowany i zbyt szeroki.

Zadanie 4

W tym zadaniu, ponieważ nieznane są parametry populacji, korzystamy z kwantyli rozkładu t-Studenta.

```
library(stats)
load('grades.RData')
IQ <- grades$IQ
```

```r
P <- grades$TestPsych
przedzial_mu_t <- function(V, poziom = 0.95){
  # Funkcja zwraca lewy i prawy brzeg przedziału ufności dla wartosci oczekiwanej z nieznanym parametrem
  s <- sd(V) ## z dokumentacji wynika, że to liczy estymator z mianownikiem n-1
  meanX <- mean(V)
  n <- length(V)
  t <- qt(p = 1 - (1-poziom)/2, df = n-1) # kwantyl rozkładu studenta z n-1 df
  low <- meanX - t*s/sqrt(n)
  high <- meanX + t*s/sqrt(n)
  return(c(low = low, high = high))
}
przedzial_mu_t(IQ)
```

```
##      low     high
## 105.9535 111.8927
```

```r
przedzial_mu_t(P)
```

```
##      low     high
## 54.16301 59.76006
```