Statystyczna Analiza Danych - PROJEKT

Piotr Greń [169783]

# Analiza zbioru danych nt. średniego wzrostu mężczyzn oraz kobiet w państwach świata

## Rozdział 1. Opis użytych danych.

**pogrubienie** Dane wykorzystane w przeprowadzonej analizie to dane posiadające różne informacje na temat państw świata. Informacje jakie zawiera zbiór danych to m.in. Nazwy państw, ich powierzchnie, liczebność populacji, informacje na temat wskaźnika rozwoju państwa GDP oraz średni wzrost wśród kobiet i mężczyzn w danym kraju.

Dane pochodzą z portalu Kaggle oferujący duże ilości repozytoriów danych Open Data, z zakładki

https://www.kaggle.com/datasets.

Cechy, które zostały poddane analizie z tego zbioru danych to **Średni wzrost mężczyzn[cm]** oraz **Średni wzrost kobiet[cm]** w danych krajach. Cechy te mówią o średnim wzroście mężczyzn bądź kobiet odnotowanym w każdym kraju z osobna.

Dane zawarte są w pliku **countries_data.csv**

Przykładowa próbka danych:

```
pander(head(df))
```

| Country | AVG M Height[cm] | AVG F Height[cm] |
|---|---|---|
| Afghanistan | 168.5 | 156.11 |
| Albania | 174.07 | 162.23 |
| Algeria | 175.04 | 162.35 |
| American Samoa | 177.09 | 167.55 |
| Andorra | 178.84 | 165.53 |
| Angola | 168.46 | 158.1 |

Aby przeprowadzić analizę średniego wzrostu mężczyzn oraz kobiet w państwach świata, przenosimy dane z odpowiednich kolumn, do list/wektorów, aby móc na nich pracować.

```
male_height <- Data$Average.Height.Male..cm.
female_height <- Data$Average.Height.Female..cm.
```

## Rozdział 2. Wyznaczenie podstawowych parametrów opisowych.

Poniżej znajdują się wyliczone, niektóre podstawowe parametry opisowe.

### Średnia

Funkcja **mean()** z pakietu **stats**. Funkcja oblicza i zwraca średnią arytmetyczną podanego do niej zbioru danych.

```
m_mean <- mean(male_height)
```

```
[1] 173.2012
```

```
f_mean <- mean(female_height)
```

```
[1] 161.0747
```

### Odchylenie Standardowe

Funkcja **sd()** z pakietu **stats**. Funkcja oblicza i zwraca odchylenie standardowe czyli jak szeroko wartości rozrzucone są wokół średniej arytmetycznej w danym zbiorze danych.

```
m_sd <- sd(male_height)
```

```
[1] 5.01936
```

```
f_sd <- sd(female_height)
```

```
[1] 4.153971
```

Według reguły empirycznej z otrzymanych wyników możemy wywnioskować, że:

- ok. **68%** danych znajduje się od średniej arytmetycznej w odległości:
  a) **5.019** - dla danych o średnim wzroście mężczyzn
  b) **4.154** - dla danych o średnim wzroście kobiet
- ok. **95%**:
  a) **10.038** - dla danych o średnim wzroście mężczyzn
  b) **8.308** - dla danych o średnim wzroście kobiet
- ok. **99.7%**:
  a) **15.057** - dla danych o średnim wzroście mężczyzn
  b) **12.462** - dla danych o średnim wzroście kobiet

**Współczynnik zmienności**

Współczynnik został obliczony ze wzoru $V = \frac{s}{\bar{x}}$. Informuje nas on o rozproszeniu/zróżnicowaniu danych względem ich średniej.

```
m_cv <- m_sd/m_mean
```

```
   [1] 0.02897994
```

```
f_cv <- f_sd/f_mean
```

```
   [1] 0.0257891
```

Z otrzymanych wyników można wywnioskować, że dane na temat średniego wzrostu mężczyzn mają niskie rozproszenie ponieważ współczynnik zmienności jest mniejszy niż **0.1**. Podobnie jest w przypadku średniego wzrostu wśród kobiet, gdzie współczynnik zmienności również jest mniejszy niż **0.1**.

**Mediana**

Funkcja **median()** z pakietu **stats**. Funkcja zwraca mediane danego zbioru danych czyli kwartyl rzędu 0.5.

```
m_median <- median(male_height)
```

```
   [1] 173.615
```

```
f_median <- median(female_height)
```

```
   [1] 160.8
```

Z otrzymanych wyników wiemy, że **50%** wartości w populacji średniego wzrostu mężczyzn znajduje się poniżej oraz powyżej wartości **173.615**. W populacji średniego wzrostu kobiet **50%** wartości znajduje się poniżej wartości **160.8**, a **50%** powyżej niej.

**Kwartyle**

Funkcja **quantile()** z pakietu **stats**. Funkcja zwraca kwantyle rzędów podanych w argumencie **probs**.

Aby wyliczyć kwartyle (oprócz mediany ponieważ jest już policzona), do argumentu **probs** należy podać wartości **0.25** oraz **0.75**.

```
m_quant <- quantile(male_height, probs = c(0.25, 0.75))
m_Q1 <- m_quant[1]
m_Q3 <- m_quant[2]
```

```
      25%     75%
   169.620 176.985
```

```
f_quant <- quantile(female_height, probs = c(0.25, 0.75))
f_Q1 <- f_quant[1]
f_Q3 <- f_quant[2]
```

```
      25%      75%
   158.1775 164.3225
```

Otrzymane wyniki mówią nam, że:

- **25%** wartości dla danych mówiących o średnim wzroście mężczyzn znajduje się poniżej wartości **169.62**, a **75%** powyżej niej
- **25%** wartości dla danych mówiących o średnim wzroście kobiet znajduje się poniżej wartości **158.1775**, a **75%** powyżej niej
- **75%** wartości dla danych mówiących o średnim wzroście mężczyzn znajduje się poniżej wartości **176.985**, a **25%** powyżej niej
- **75%** wartości dla danych mówiących o średnim wzroście kobiet znajduje się poniżej wartości **164.3225**, a **25%** powyżej niej

**Minimalna wartość**

Funkcja **min()** z pakietu **stats**. Funkcja zwraca minimalną wartość w próbce danych.

```
m_min <- min(male_height)
```

```
[1] 160.13
```

```
f_min <- min(female_height)
```

```
[1] 150.91
```

Najmniejsza wartości znajdująca się w zbiorze danych, mówiących o średnim wzroście mężczyzn to **160.13**.

W zbiorze danych o średnim wzrości kobiet najmniejsza wartość to **150.91**.

**Maksymalna wartość**

Funkcja **max()** z pakietu **stats**. Funkcja zwraca maksymalną wartość w próbce danych.

```
m_max <- max(male_height)
```

```
[1] 183.78
```

```
f_max <- max(female_height)
```

```
[1] 170.36
```

Największą wartością znajdującą się w zbiorze danych mówiących o średnim wzroście mężczyzn jest **182.78**. Natomiast największą wartością znajdującą się w zbiorze danych o średnim wzroście kobiet na świecie jest **170.36**.

### Dominanta

Funkcja **table()** oraz **names()** z pakietu **base**. Funkcja **table()** wykorzystuje czynnik przeklasyfikowania do stworzenia tabeli ze zliczonymi występieniami każdej wartości występującej w zbiorze danych. Funkcja **names()** służy do określenia nazwy obiektu.

```
m_mcv <- names(which.max(table(male_height)))
m_mcv <- as.numeric(m_mcv)
```

```
[1] 170.67
```

```
f_mcv <- names(which.max(table(female_height)))
f_mcv <- as.numeric(f_mcv)
```

```
[1] 154.76
```

### Wariancja

Funkcja **var()** z pakietu **cmvnorm**. Funkcja zwraca wariancję próbki danych.

```
m_var <- var(male_height)
```

```
[1] 25.19397
```

```
f_var <- var(female_height)
```

```
[1] 17.25548
```

### Rozstęp danych

Rozstęp danych został policzony zgodnie z definicją bez używania funkcji środowiska R. Wzór: $R = x_{max} - x_{min}$

```
m_range <- m_max - m_min
```

```
[1] 23.65
```

```
f_range <- f_max - f_min
```

```
[1] 19.45
```

Rozstęp danych to inaczej odległość pomiędzy największą wartością występującą w zbiorze danych, a namniejszą.

Mówi nam z jak dużego przedziału są wartości występujące w zbiorze danych.

Rozstęp danych zbioru danych na temat średniego wzrostu mężczyzn wynosi **23.65**.

Rozstęp danych zbioru danych na temat średniego wzrostu kobiet wynosi **19.45**.

**Skośność danych**

Funkcja **skew()** z pakietu **moments**. Funkcja oblicza skośność danych czyli miarę asymetrii rozkładu danych. Mówi o tym jak bardzo rozkład danych różni się od symetrycznego rozkładu normalnego.

```
m_skew <- skewness(male_height)
```

```
[1] -0.05344941
```

```
f_skew <- skewness(female_height)
```

```
[1] 0.004328793
```

Na podstawie otrzymanych wyników możemy wywnioskować, iż

- rozkład populacji średniego wzrostu mężczyzn jest nieznacznie, **lewostronnie skośny** (ogon rozkładu jest dłuższy po lewej stronie)
- rozkład popupacji średniego wzrostu kobiet jest **prawostronnie skośny** (ogon rozkładu jest dłuższy po prawej stronie)

## Podsumowanie parametrów opisowych

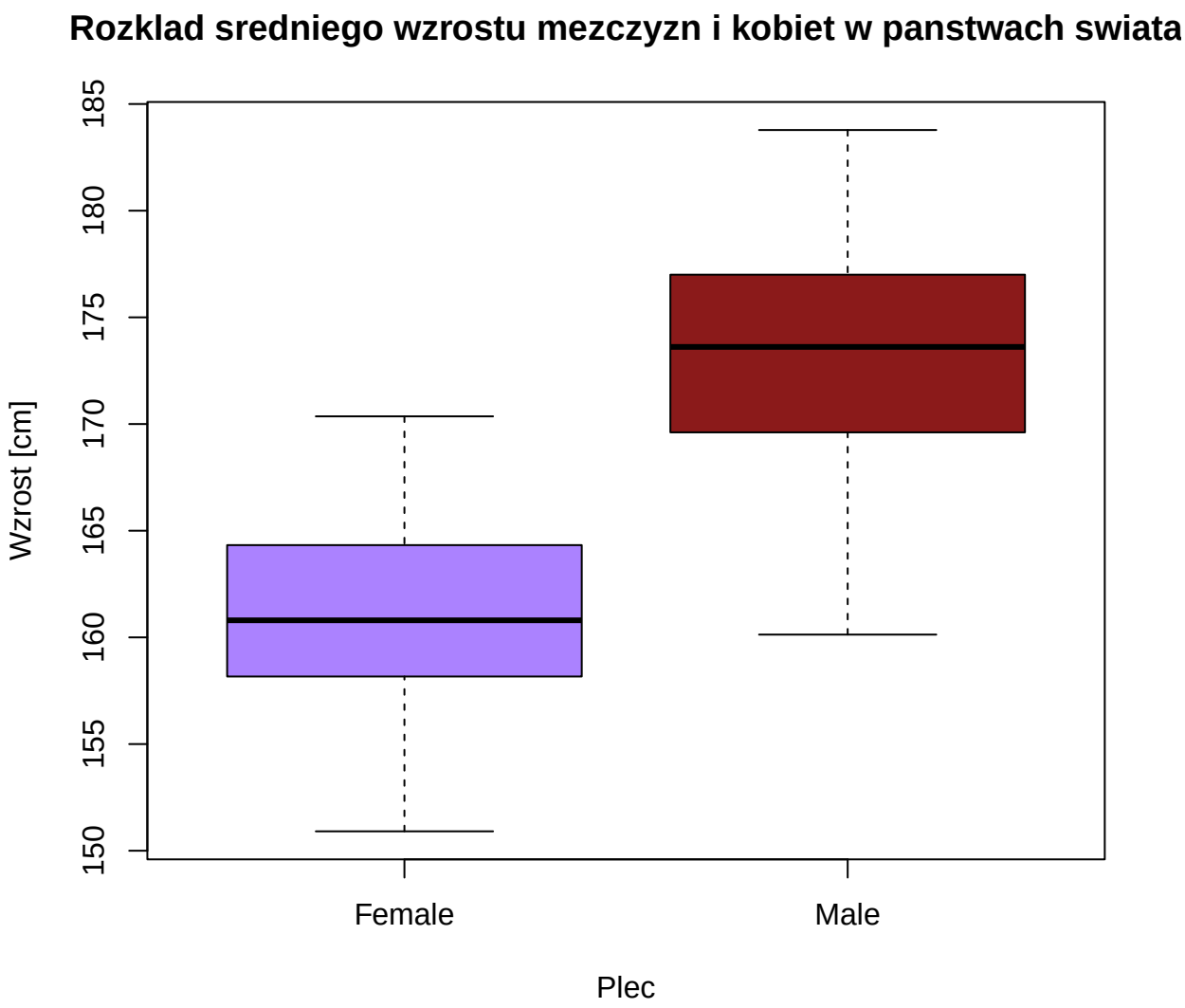
| Parametr | AVG_Male_Height | AVG_Female_Height |
|---|---|---|
| Średnia | 173.2 | 161.1 |
| Odchylenie standardowe | 5.019 | 4.154 |
| Współczynnik zmienności | 0.029 | 0.026 |
| Mediana | 173.6 | 160.8 |
| Kwartyl 25% | 169.6 | 158.2 |
| Kwartyl 75% | 177 | 164.3 |
| Min | 160.1 | 150.9 |
| Max | 183.8 | 170.4 |
| Dominanta | 170.7 | 154.8 |
| Wariancja | 25.19 | 17.25 |
| Rozstęp Danych | 23.65 | 19.45 |
| Skośność Danych | -0.053 | 0.004 |

## Rozdział 3. Graficzne przedstawienie danych.

### Wykres pudłekowy

Wykres reprezentuje rozkład analizowanych cech przedstawiając informacje dotyczące położenia i rozproszenia danych.

```
combined <- c(male_height, female_height)
gender <- c(rep("Male", length(male_height)), rep("Female", length(female_height)))
boxplot(combined ~ gender, col = c("mediumpurple1", "firebrick4"),
        main = "Rozkład średniego wzrostu mężczyzn i kobiet w państwach świata",
        xlab = "Płeć", ylab = "Wzrost [cm]")
```



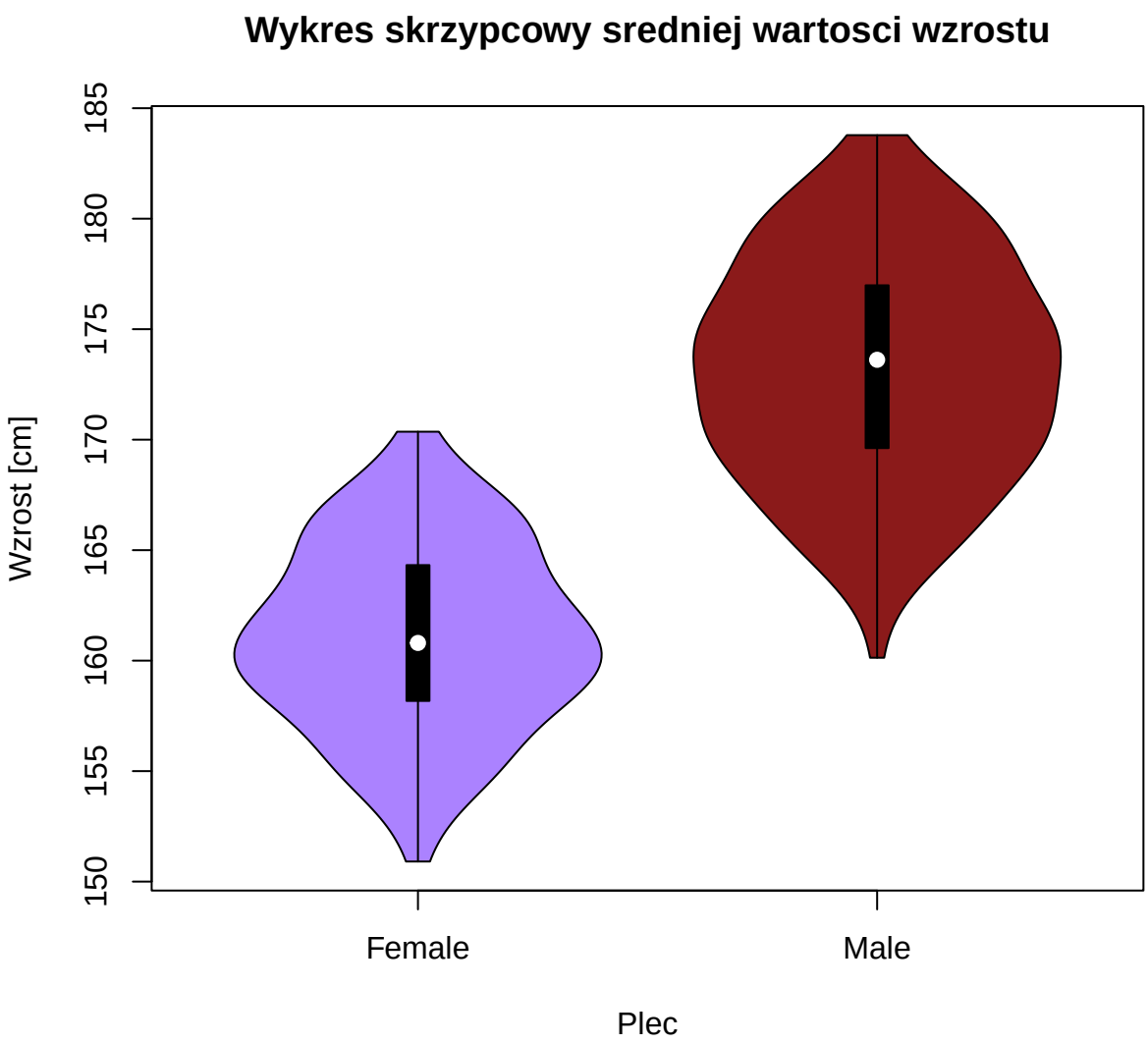

Na wykresie powyżej możemy mniej więcej zauważyć jak wygląda położenie i rozproszenie danych oraz kwartyli, dla obu badanych cech.

**Wykres skrzypcowy**

Wykres skrzypcowy jest rozszerzeniem wykresu pudełkowego. Podczas gdy wykres pudełkowy pokazuje nam tylko rozmieszczenie danych poprzez pudełko wyznaczone przez kwartyle oraz "knoty" świec/pudełek wyznaczone poprzez minimalną i maksymalną wartość w zbiorze danych, to wykres skrzypcowy przedstawia nam również gęstość rozkładu danych. Wykres skrzypcowy to nic innego jak wykres pudełkowy, tylko zamiast pudełka są skrzypce, wyznaczone przez linie KDE.

Funkcja **vioplot()** z pakietu **vioplot**. Funkcja rysuje wykres skrzypcowy dla podanego zbioru danych.

```
vioplot(combined ~ gender, col = c("mediumpurple1", "firebrick4"),
        main = "Wykres skrzypcowy średniej wartości wzrostu",
        xlab = "Płeć", ylab = "Wzrost [cm]")
```



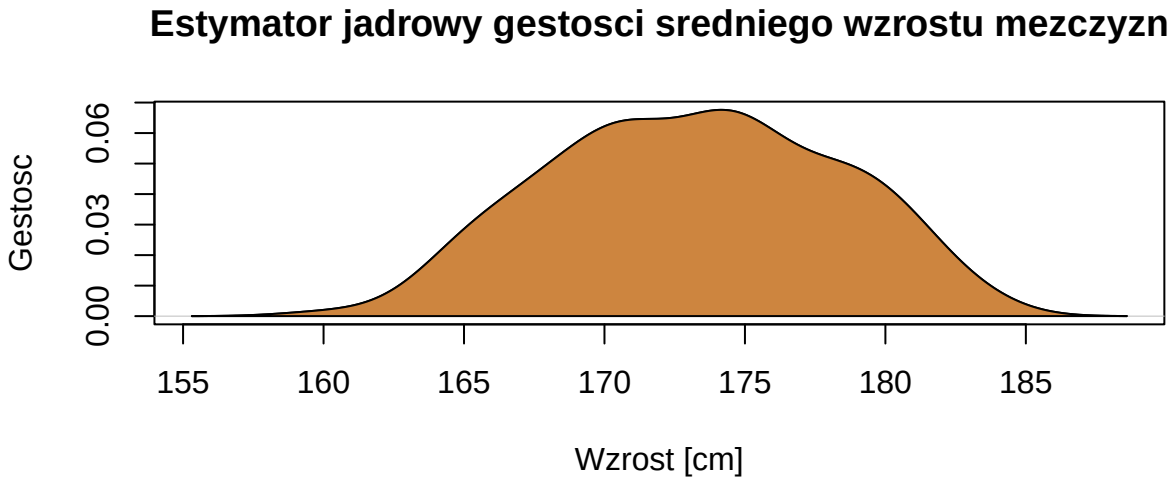

Na powyższym wykresie, biała kropka oznacza położenie mediany dla obu analizowanych cech. Oprócz białej kropki wydzimy też grubszy prostoką czyli nic innego jak pudełko, wyznaczone przez kwartyle rzędu pierwszego oraz trzeciego. Cienka pionowa kreska wyznacza rozmieszczenie/rozstęp danych. Ostatecznie skrzypce wyznaczone są przez linie KDE czyli estymatora gęstości jądrowej, który jest estymatorem funkcji gęstości prawdopodobieństwa opisującej rozkład danych.

**Wykres KDE**

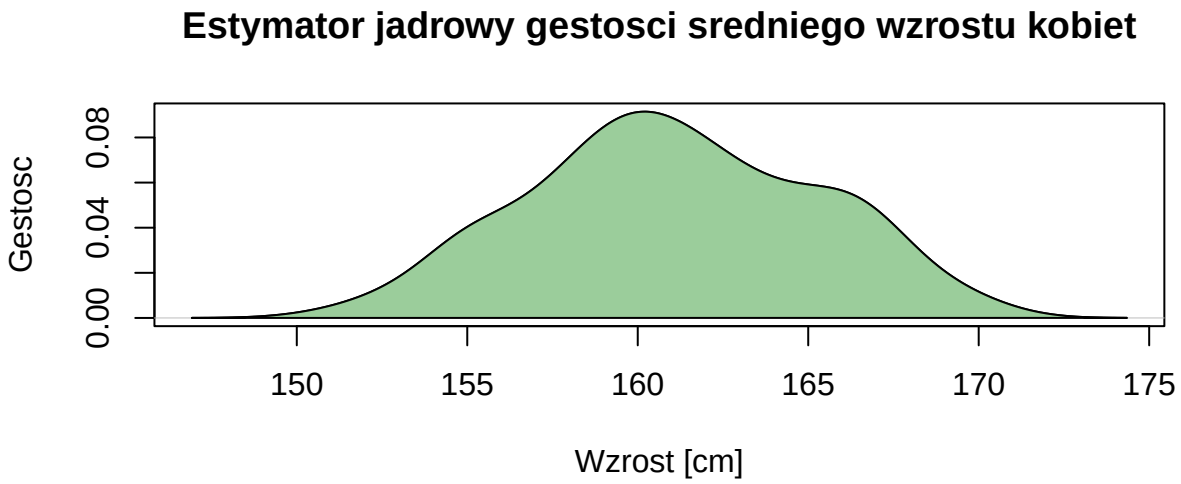
KDE jest to estymator jądrowy gęstości przeznaczony do wyznaczania gęstości rozkładu zmiennej losowej, na podstawie uzyskanej próby lub populacji. Podczas wyznaczania gęstości przy pomocy KDE nie jest wymagana informacja o typie występującego rozkładu.

Funkcja **density()** z pakietu **stats**. Funkcja oblicza/wyznacza funkcję gęstości prawdopodobieństwa.

```
d1 <- density(male_height)
plot(d1, main = "Estymator jądrowy gęstości średniego wzrostu mężczyzn",
     xlab = "Wzrost [cm]", ylab = "Gęstość")
polygon(d1, col = "tan3", border = "black")
```




```
d2 <- density(female_height)
plot(d2, main = "Estymator jądrowy gęstości średniego wzrostu kobiet",
     xlab = "Wzrost [cm]", ylab = "Gęstość")
polygon(d2, col = "darkseagreen3", border = "black")
```



**Histogram**

Histogram pozwala na analizę rozkładu danych na podstawie częstości ich występowania w określonych przedziałach. Do każdego histogramu dodany został również przeskalowany wykres gęstości prawdopodobieństwa, obliczony zakładając, że dane należą do rozkładu normalnego. Histogram przedstawia rozkład danych w sposób dyskretny, podzielony na przedziały. Dodanie wykresu gęstości pozwala na porównanie go z ciągłym rozkładem teoretycznym, takim jak właśnie rozkład normalny. Dzięki temu można zobaczyć, jak dobrze dane pasują do teoretycznego rozkładu i czy występują odstępstwa.

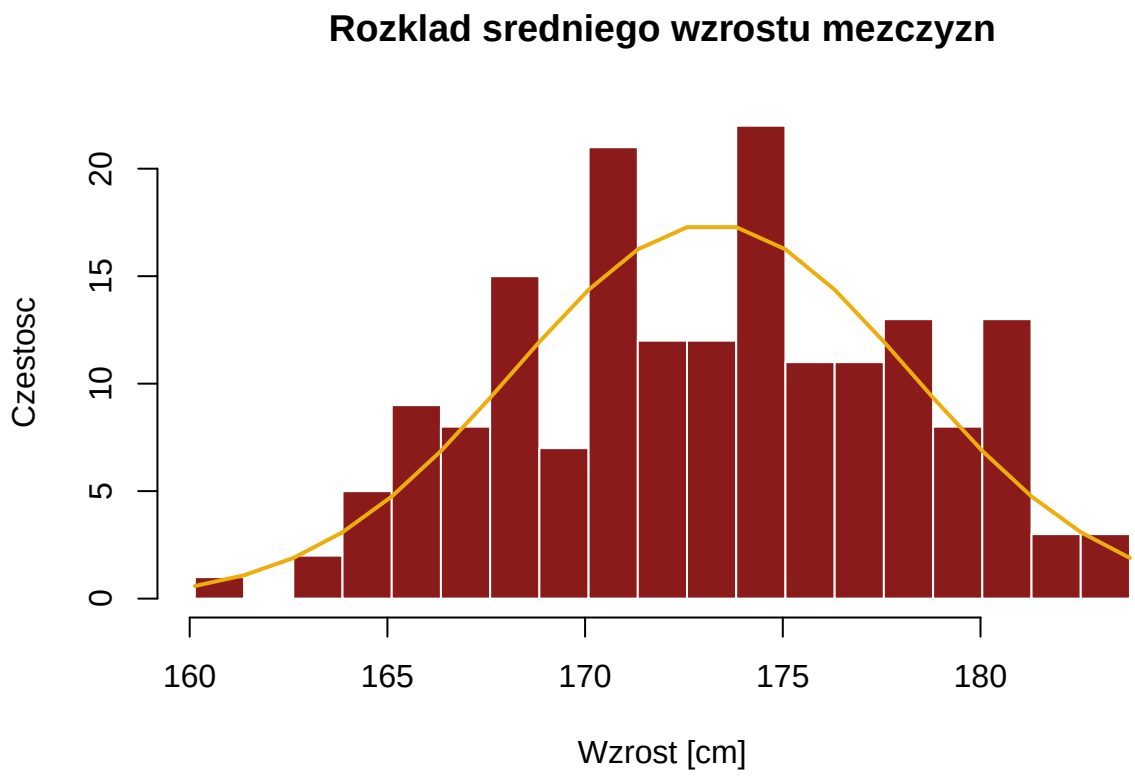
Funkcja **hist()** z pakietu **graphics**. Funkcja służy do generowania histogramu na podstawie dostarczonych danych oraz wyznaczonych przedziałów w argumencie breaks.

*Wykres dla populacji średniego wzrostu mężczyzn*

Dobrana liczba przedziałów w tym wypadku to 20. Program sam oblicza przedziały i dopasowuje do nich dane.

Obliczona szerokość pojedyńczego przedziału to ok. **1.24**.

```
h <- hist(male_height, breaks = seq(min(male_height), max(male_height), length.out = 20),
          main = "Rozkład średniego wzrostu mężczyzn", xlab = "Wzrost [cm]",
          ylab = "Częstość", col = "firebrick4", border = "white")
xfit <- seq(min(male_height), max(male_height), length.out = 20)
yfit <- dnorm(xfit,mean = m_mean, sd = m_sd)
yfit <- yfit*diff(h$mids[1:2])*length(male_height)
lines(xfit, yfit, col = "darkgoldenrod2", lwd = 2)
```



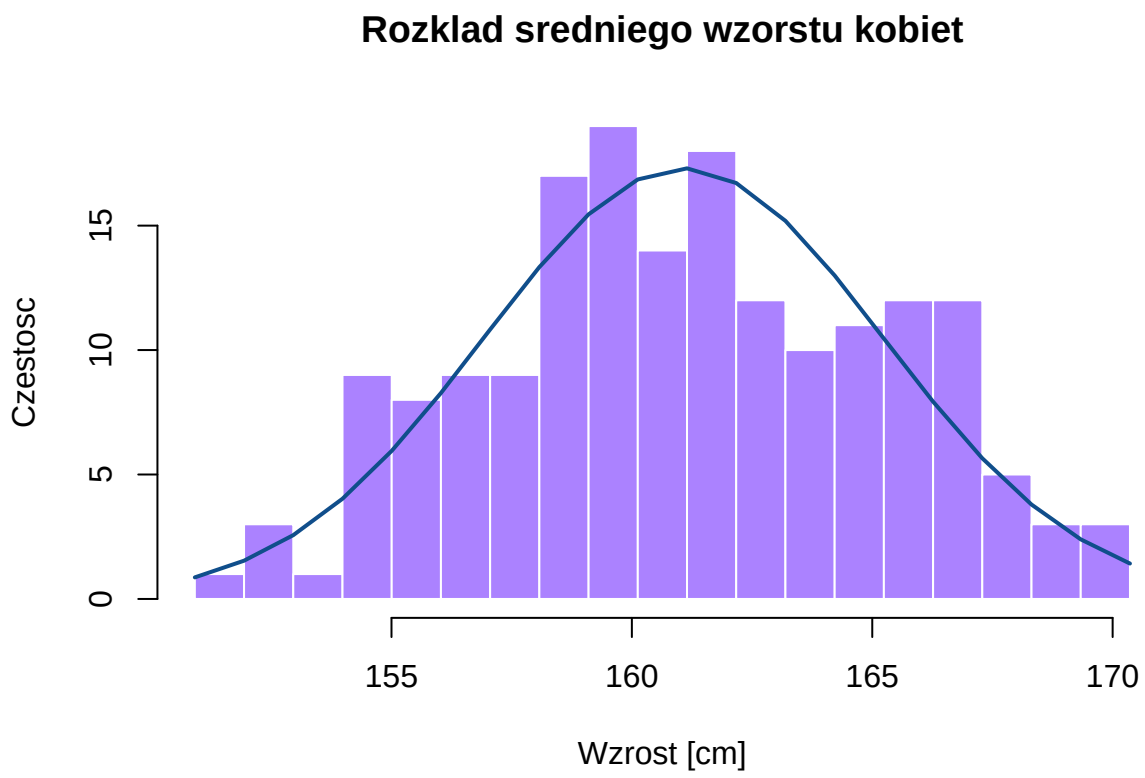

Na powyższym wykresie możemy zobaczyć rozkład danych i ich częstość występowania w danych przedziałach. Jak widzimy funkcja gęstości wyznaczona na podstawie rozkładu normalnego posiada dosyć duże odstępstwa.

*Wykres dla populacji średniego wzrostu kobiet*

Dobrana liczba przedziałów w tym wypadku to 20. Program sam oblicza przedziały i dopasowuje do nich dane.

Obliczona szerokość pojedyńczego przedziału to ok. **1.02**.

```
h1 <- hist(female_height, breaks = seq(f_min, f_max, length.out = 20),
           main = "Rozkład średniego wzorstu kobiet", xlab = "Wzrost [cm]",
           ylab = "Częstość", col = "mediumpurple1", border = "white")
xfitf <- seq(f_min, f_max, length.out = 20)
yfitf <- dnorm(xfitf, mean = f_mean, sd = f_sd)
yfitf <- yfitf*diff(h1$mids[1:2])*length(female_height)
lines(xfitf, yfitf, col = "dodgerblue4", lwd = 2)
```




Na powyższym wykresie możemy zobaczyć rozkład danych i ich częstość występowania w danych przedziałach. Jak widzimy funkcja gęstości wyznaczona na podstawie rozkładu normalnego, nawet dobrze opisuje rozkład danych średniej wartości wzrostu kobiet w krajach świata, jednak tutaj również podobnie jak w przypadku średniego wzrostu mężczyzn, występują odstępstwa.

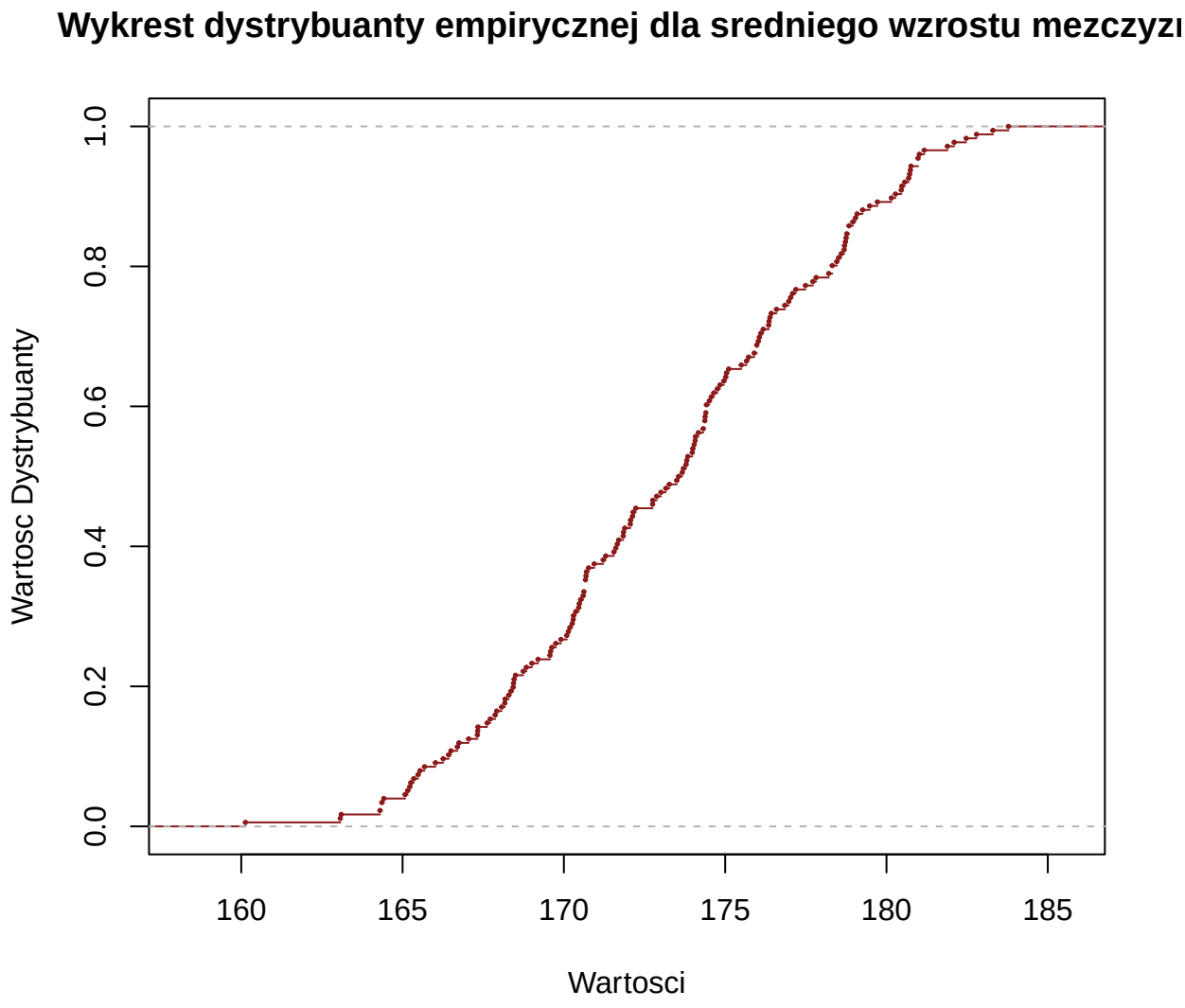
### Wykres dystrybuanty empirycznej

Dystrybuanta empiryczna to nic innego jak estymacja dystrybuanty prawdziwej dla danego zbioru danych. Do uzyskania dystrybuanty empirycznej, wykorzystuje się ECDF czyli funkcję dostarczaną przez paiet stats w środowisku R. Wykres dystrybuanty empirycznej jest graficznym narzędziem, wykorzystywanym do wizualizacji empirycznego rozkładu danych.

Funkcja **ecdf()** z pakietu **stats**. Funkcja służy do wyznaczenia empirycznej dystrybuanty rozkładu skumulowanego.

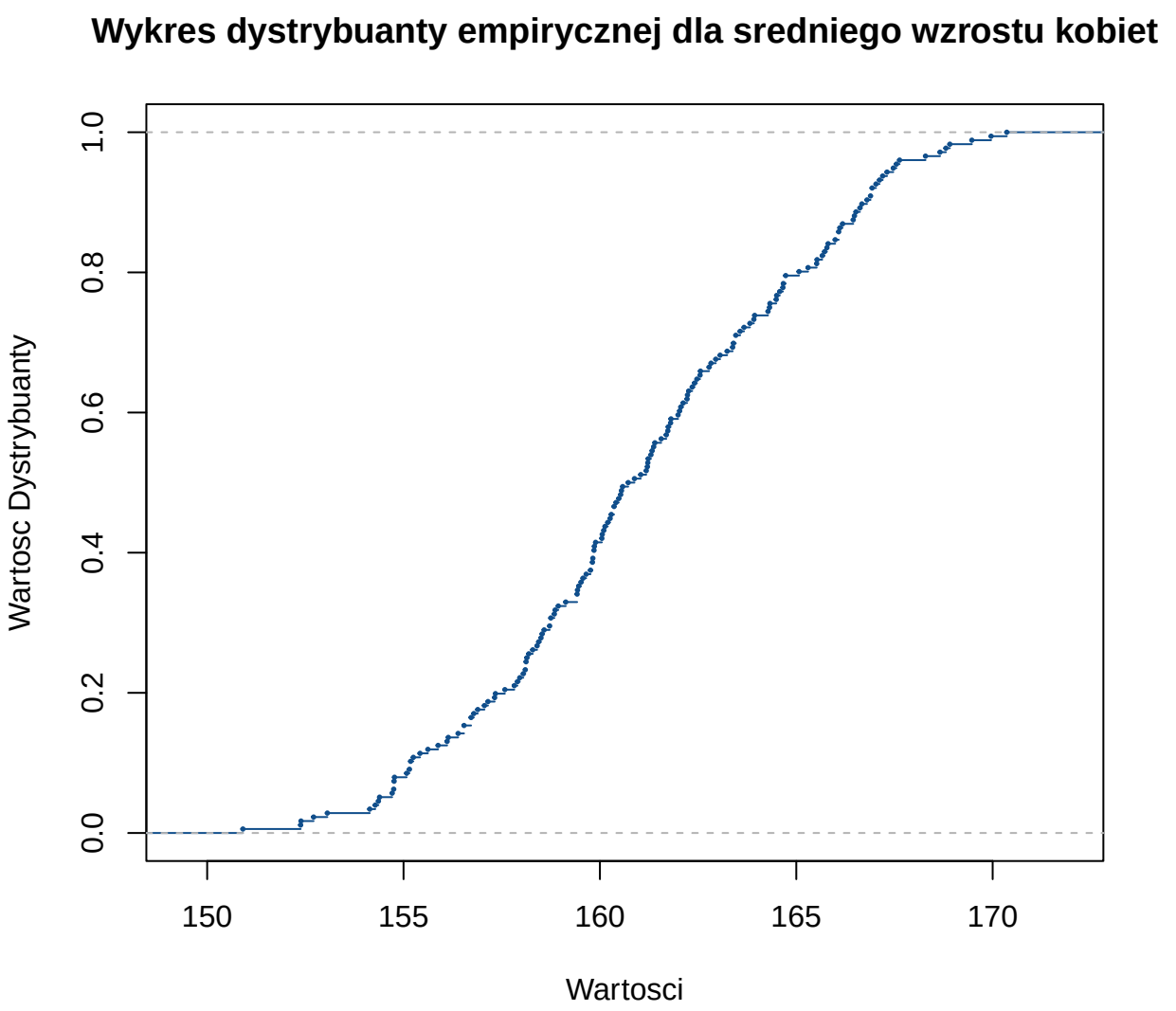
*Wykres Dystrybuanty Empirycznej dla średniego wzrostu mężczyzn na świecie*

```
emp_distm <- ecdf(male_height)
plot(emp_distm, main = "Wykrest dystrybuanty empirycznej dla średniego wzrostu mężczyzn",
     xlab = 'Wartości', ylab = "Wartośc Dystrybuanty",
     pch = 19, cex = 0.25, col = "firebrick4")
```



*Wykres Dystrybuanty Empirycznej dla średniego wzrostu kobiet na świecie*

```
emp_distf <- ecdf(female_height)
plot(emp_distf, main = "Wykres dystrybuanty empirycznej dla średniego wzrostu kobiet",
     xlab = "Wartości", ylab = "Wartość Dystrybuanty",
     pch = 19, cex = 0.25, col = "dodgerblue4")
```




Na obu powyższych wykresach widzimy dystrybuanty rozkłady danych zarówno dla zbioru nt. średniego wzrostu mężczyzn na świecie jak i zbioru nt. średniego wzrostu kobiet na świecie.

Obie dystrybuanty jednoznacznie wyznaczają rozkład prawdopodobieństwa (miarę probabilistyczną).

**Wykres korelacji pomiędzy wzrostem, a wskaźnikiem rozwoju kraju**

Na poniższym wykresie przetestowana jest korelacja pomiędzy wskaźnikiem rozwoju, kraju, a średnim wzrostem mężczyzn w danym kraju. Celem tego wykresu jest sprawdzenie czy faktycznie można mówić o jakiejś korelacji pomiędzy tymi dwiema cechami. Informacje na temat wskaźnika rozwoju danych ogólnie przyjmując, mówią nam o stopniu rozwoju państwa.

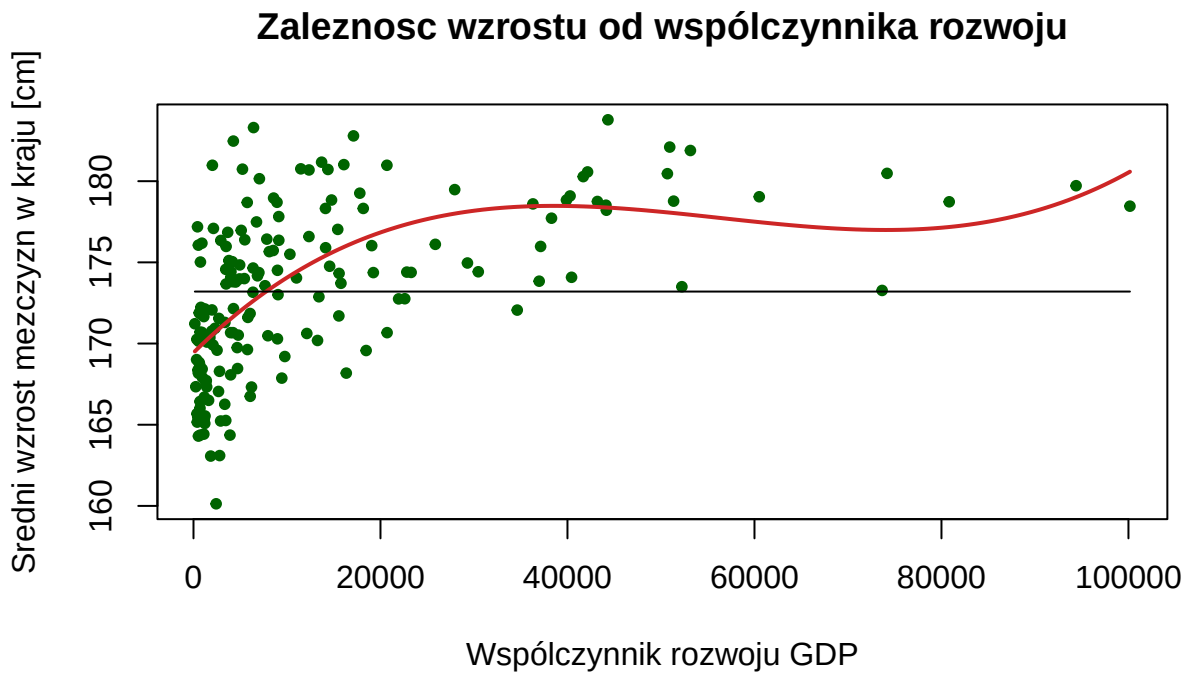
Funkcja **plot()** z pakietu **graphics**. Funkcja służy do rysowania prostych wykresów czy to liniowych, czy punktowych.

Do wykresu dorysowana została linia regresji wielomianowej (ponieważ jak widać na wykresie prosta regresji nie byłaby wystarczająco dopasowana).

Funkcja **lm()** z pakietu **stats**. Funkcja służąca do dopasowania modeli liniowych, lub wielomianowych stosowana też do wyznaczania regresji.

Funkcja **coef()** z pakietu **stats**. Funkcja służąca do wyciągania współczynników w wyznaczonym modelu.

```
wskaznik <- Data$GDP.per.capita..current.US..
model <- lm(male_height ~ poly(wskaznik, 3, raw = TRUE))
cf <- coef(model)
poly_func <- function(x) {
  y <- cf[1] + cf[2] * x + cf[3] * x^2 + cf[4] * x^3
  return(y)
}
x_pred <- seq(min(wskaznik), max(wskaznik), length.out = 1000)
y_pred <- poly_func(x_pred)
plot(wskaznik, male_height, col = "darkgreen", xlab = "Współczynnik rozwoju GDP",
     ylab = "Średni wzrost mężczyzn w kraju [cm]", main = "Zależność wzrostu od współczynnika rozwoju",
     pch = 20)
lines(x_pred, y_pred, col = "firebrick3", lwd = 2)
lines(seq(min(wskaznik), max(wskaznik), length.out=100), rep(m_mean, 100), col = "black")
```



Na wykresie powyżej możemy zauważyć, że nie ma silnej korelacji pomiędzy stopniem rozwoju państwa, a średnim wzrostem wśród mężczyzn. Co prawda można stwierdzić, że dla państw wysoko rozwiniętych faktycznie występuje jakieś powiązanie, ponieważ jak widzimy, dla każdego państwa z wysokim współczynnikiem rozwoju, średnia wzrostu mężczyzn w kraju znajduje się powyżej średniej światowej (zaznaczona na wykresie czarną linią). Jednakże jest to za mało żeby mówić tu o silnej korelacji.

## Rozdział 4. Testowanie hipotez statystycznych.

### Hipoteza 1:

Hipotezę będziemy testować na losowo pobranej próbie z popuilacji danych nt. średniego wzrostu mężczyzn na świecie. Do losowego pobrania próby używamy funkcji **sample()** z pakietu **base**. Funkcja służy do losowania próbki danych o zadanej długości, z wektora zawierającego dane generalne. Przed wylosowaniem korzystamy również z funkcji **set.seed()** z pakiet **simEd** z wartością **42**, która zapewnia powtarzalność wylosowanej próby w przypadku ponownego uruchomienia kodu.

```
set.seed(42)
proba <- sample(male_height, 70, replace=FALSE)
```

Dla wylosowanej próby o liczności 70 testujemy hipotezę:

**H0:** Średni wzrost mężczyzn na świecie wynosi 172.5 cm | $\mu = \mu_0$

**H1:** Średni wzrost mężczyzn na świecie jest większy od 172.5 cm | $\mu > \mu_0$

Hipotezę będziemy testować na poziomie istotności $\alpha = 0.05$.

Ponieważ zakładamy, że nie znamy odchylenia standardowego populacji generalnej będziemy korzystać ze statystyki **T** określonej wzorem:

$T = \frac{\overline{x}-\mu}{S}\sqrt{n-1}$

Statystyka ta ma rozkład t-Studenta.

Na początek obliczamy **średnią** oraz **odchylenie standardowe**.

```
p_mean <- mean(proba)
```

```
[1] 173.8269
```

```
p_sd <- sd(proba)
```

```
[1] 4.734018
```

Następnie musimy zadeklarować poziom istotności, liczbę wystąpień (czyli długość zbioru danych) oraz wartość oczekiwaną.

```
alfa <- 0.05
mi <- 172.5
n <- length(proba)
```

```
[1] 70
```

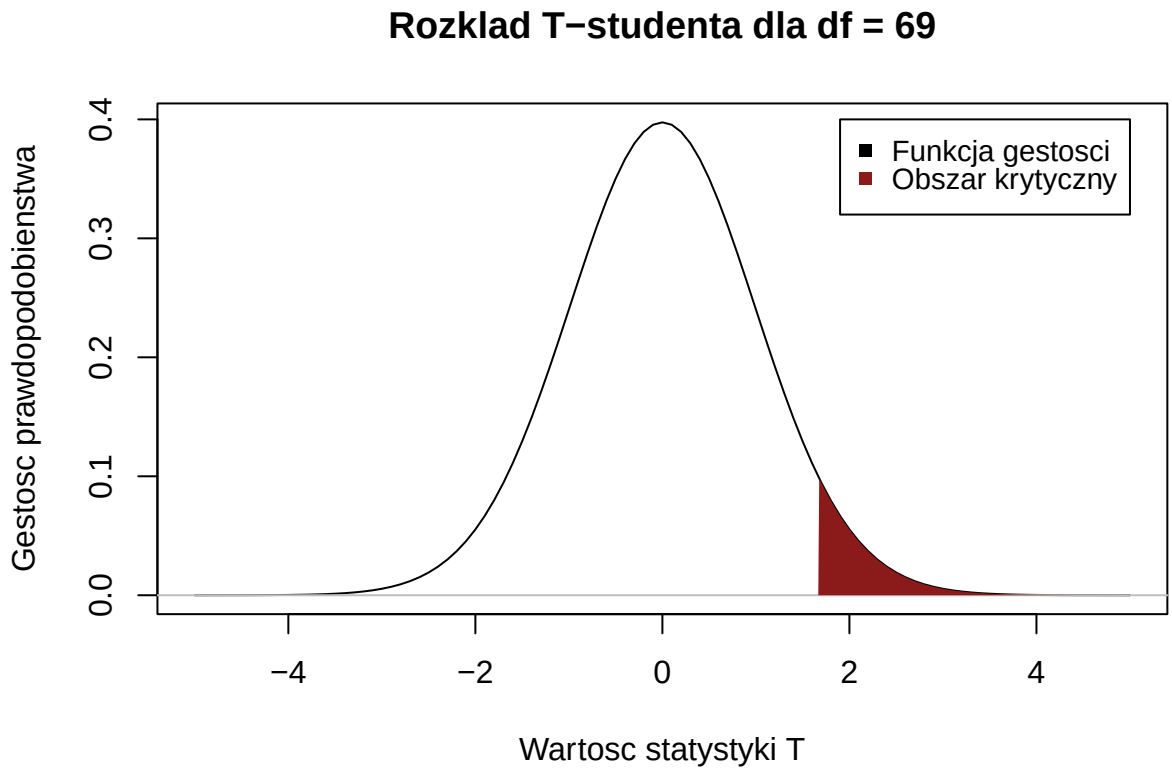
W następnym kroku musimy policzyc granicę obszaru krytycznego. Jako, że hipoteza alternatywna mówi, że średni wzrost mężczyzn jest *większy* niż zakładamy w hipotezie zerowej, musimy policzyć prawy obszar krytyczny. Granicę obszaru krytycznego wyliczamy z rozkładu t-Studenta $t_{[1-\alpha,n-1]}$. Używamy do tego funkcji **qt()** z pakietu **stats**. Funkcja **qt()** zwraca wartość odwrotnej funkcji gęstości skumulowanej (CDF) rozkładu t-Studenta dla określonej zmiennej losowej x w naszym wypadku 1 - $\alpha$ i stopni swobody df w naszym przypadku n - 1. Innymi słowy funkcja zwraca konkretny, szukany kwantyl.

```
p_kwantyl <- qt(1 - alfa, n - 1)
```

```
[1] 1.667239
```

Tak prezentuje się wykres gęstości rozkładu t-Studenta, dla stopni swobody n - 1, z zaznaczonym obszarem krytycznym w badanej przez nas hipotezie.

```
funkcja_gestosci <- function(x) dt(x, n - 1)
curve(funkcja_gestosci, from = -5, to = 5, ylab = "Gęstość prawdopodobieństwa",
      main = "Rozkład T-studenta dla df = 69", xlab = "Wartość statystyki T")
abline(h = 0, col = "gray")  # Linia bazowa
x_values <- seq(-5, 5, length.out = 1000)
y_values <- funkcja_gestosci(x_values)
obszar_krytyczny_prawy <- x_values[x_values >= p_kwantyl]
obszar_y_prawy <- y_values[x_values >= p_kwantyl]
polygon(c(obszar_krytyczny_prawy, p_kwantyl), c(obszar_y_prawy, 0), col = "firebrick4", border = NA)
legend(x = c(1.9, 5), y = c(0.4, 0.32), legend = c("Funkcja gęstości", "Obszar krytyczny"),
       col = c("black", "firebrick4"),
       pch = 15, cex = 0.9, border = NA, y.intersp = 0.8)
```



Teraz musimy obliczyć wartość statystyki **T** dla testowanej przez nas hipotezy zgodnie ze wzorem.

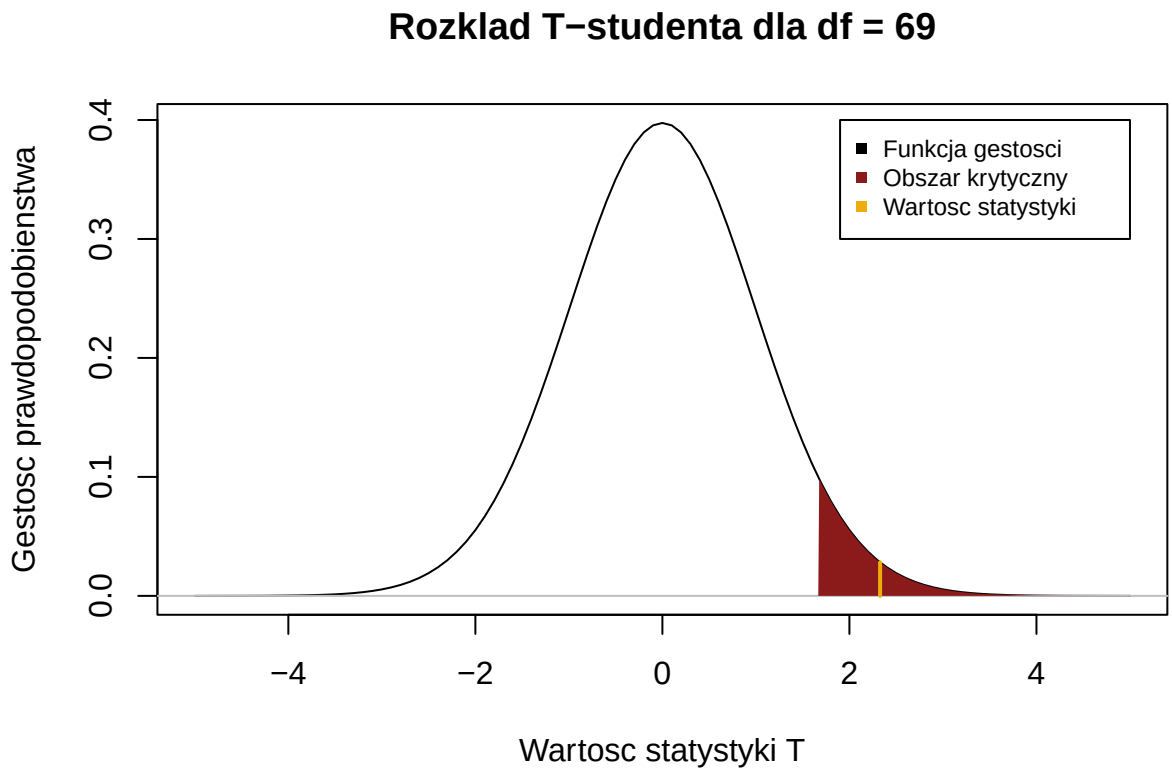
```
T1 <- ((p_mean - mi)/p_sd)*sqrt(n - 1)
```

```
   [1] 2.328192
```

Po dodaniu obliczonej statystyki do wykresu możemy zobaczyć gdzie dokładnie się znajduje na wykresie funkcji gęstości.

Statystyka została narysowana jako linia łącząca jej współrzędne x oraz y aby była dobrze widoczna.

```
x_value <- T1
y_value <- funkcja_gestosci(T1)
segments(x0 = x_value, y0 = 0, x1 = x_value, y1 = y_value, col = "darkgoldenrod2", lwd = 2)
legend(x = c(1.9, 5), y = c(0.4, 0.3), legend = c("Funkcja gęstości",
                                                  "Obszar krytyczny", "Wartość statystyki"),
       col = c("black", "firebrick4", "darkgoldenrod2"),
       pch = 15, cex = 0.75)
```




Jak widzimy wartość statystyki należy do obszaru krytycznego. **T > p_kwantyl**.

Wartość statystyki T: **2.328192**

Dolna granica obszaru krytycznego: **1.667239**

Hipotezę zerową [**H0**] odrzucamy na korzyść hipotezy alternatywnej [**H1**].

**Hipoteza 2:**

Do tej hipotezy wykorzystamy dużą próbę danych ze zbioru danych nt. średniego wzrostu kobiet na świecie. Do losowego pobrania próby używamy funkcji **sample()** z pakietu **base**. Funkcja służy do losowania próbki danych o zadanej długości, z wektora zawierającego dane populacji generalnej. Przed wylosowaniem korzystamy również z funkcji **set.seed()** z pakiet **simEd** z wartością **42**, podobnie jak w poprzedniej hipotezie.

```
set.seed(42)
proba_f <- sample(female_height, 100, replace = FALSE)
```

Testujemy hipotezę:

**H0:** Odchylenie standardowe średniego wzrostu kobiet na świecie wynosi 4 | $\sigma = \sigma_0$

**H1:** Odchylenie standardowe średniego wzrostu kobiet na świecie rózni się od 4 | $\sigma$ != $\sigma_0$

Hipotezę będziemy testować na poziomie istotności $\alpha = 0.06$.

Do obliczenia statystyki będziemy korzystać ze statystyki **V** określonej wzorem:

$V = \frac{nS^2}{\sigma^2}$

Statystyka posiada rozkład $\chi^2$.

Na początek musimy obliczyć **odchylenie standardowe**.

```
pf_sd <- sd(proba_f)
```

```
[1] 3.914264
```

Następnie deklarujemy **odchylenie standardowe**, którego hipotetyczną wartość poddajemy testowi, poziom istotności, oraz liczebność zbioru danych.

```
sigma <- 4
alfa <- 0.06
n <- length(proba_f)
```

```
[1] 100
```

W następnym kroku musimy policzyc granice obszaru krytycznego. Jako, że hipoteza alternatywna mówi, że odchylenie standardowe średniego wzrostu kobiet na świecie jest *różne* niż zakładamy w hipotezie zerowej, musimy policzyć zarówno prawy jak i lewy obszar krytyczny.

Granicę obszarów krytycznych wyliczamy z rozkładu $\chi^2$, odpowiednio:

- lewy obszar krytyczny $\chi^2_{[\frac{\alpha}{2}, n-1]}$
- prawy obszar krytyczny $\chi^2_{[1-\frac{\alpha}{2}, n-1]}$

Do obliczenia szukanych wartości używamy funkcji **qchisq()** z pakietu **stats**. Funkcja oblicza i zwraca konkretne kwantyle dla rozkładu $\chi^2$ o zadanej ilości stopni swobody, w naszym przypadku **n - 1**.

```
l_kwantyl_chi <- qchisq(alfa/2, n - 1)
```

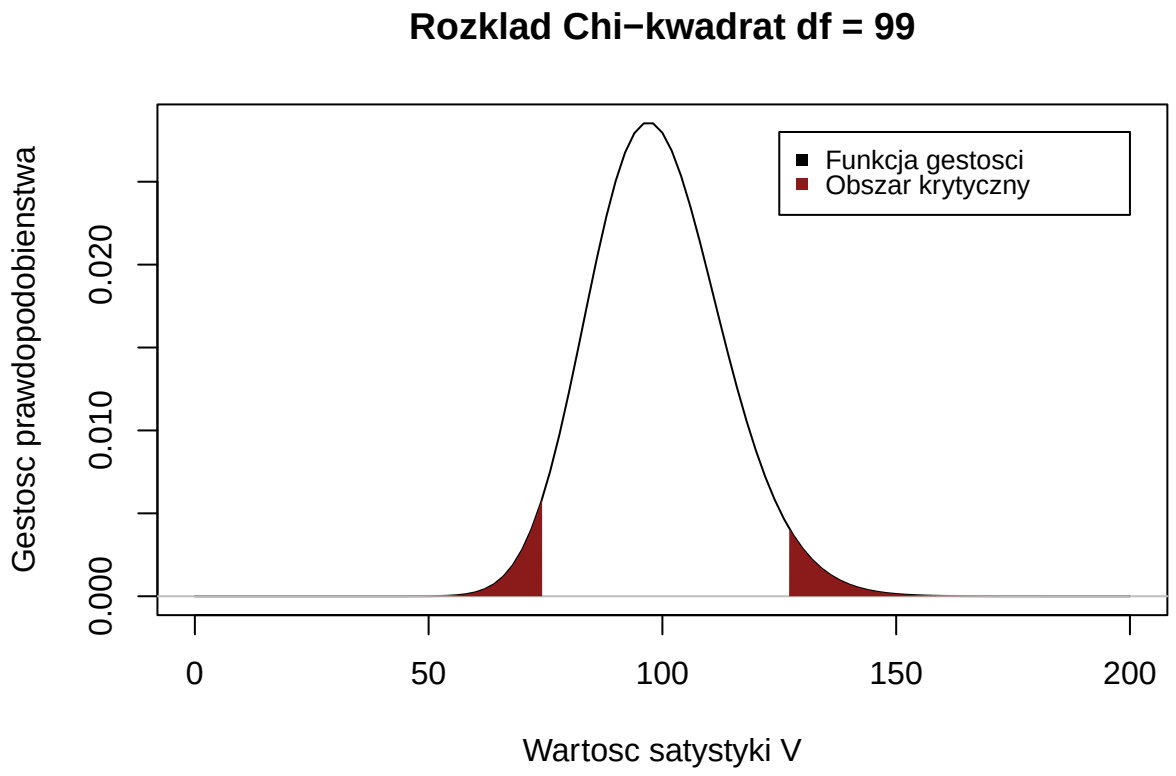
```
[1] 74.27541
```

```r
p_kwantyl_chi <- qchisq(1 - alfa/2, n - 1)
```

```
[1] 127.1031
```

Tak prezentuje się wykres gęstości rozkładu $\chi^2$, dla stopni swobody n - 1, z zaznaczonymi obszarami krytycznymi w badanej przez nas hipotezie.

```r
funkcja_chi <- function(x) dchisq(x, df = n - 1)
curve(funkcja_chi, from = 0, to = 200, xlab = "Wartość satystyki V",
      ylab = "Gęstość prawdopodobieństwa", main = "Rozkład Chi-kwadrat df = 99")
abline(h = 0, col = "grey")
x_values <- seq(0, 200, length.out = 1000)
y_values <- funkcja_chi(x_values)
obszar_krytyczny_lewy <- x_values[x_values <= l_kwantyl_chi]
obszar_krytyczny_prawy <- x_values[x_values >= p_kwantyl_chi]
obszar_y_lewy <- y_values[x_values <= l_kwantyl_chi]
obszar_y_prawy <- y_values[x_values >= p_kwantyl_chi]
polygon(c(obszar_krytyczny_prawy, p_kwantyl_chi), c(obszar_y_prawy, 0),
        col = "firebrick4", border = NA)
polygon(c(obszar_krytyczny_lewy, l_kwantyl_chi), c(obszar_y_lewy, 0),
        col = "firebrick4", border = NA)
legend(x = c(125, 200), y = c(0.028, 0.023), legend = c("Funkcja gęstości", "Obszar krytyczny"),
       col = c("black", "firebrick4"),
       pch = 15, cex = 0.8, y.intersp = 0.8)
```



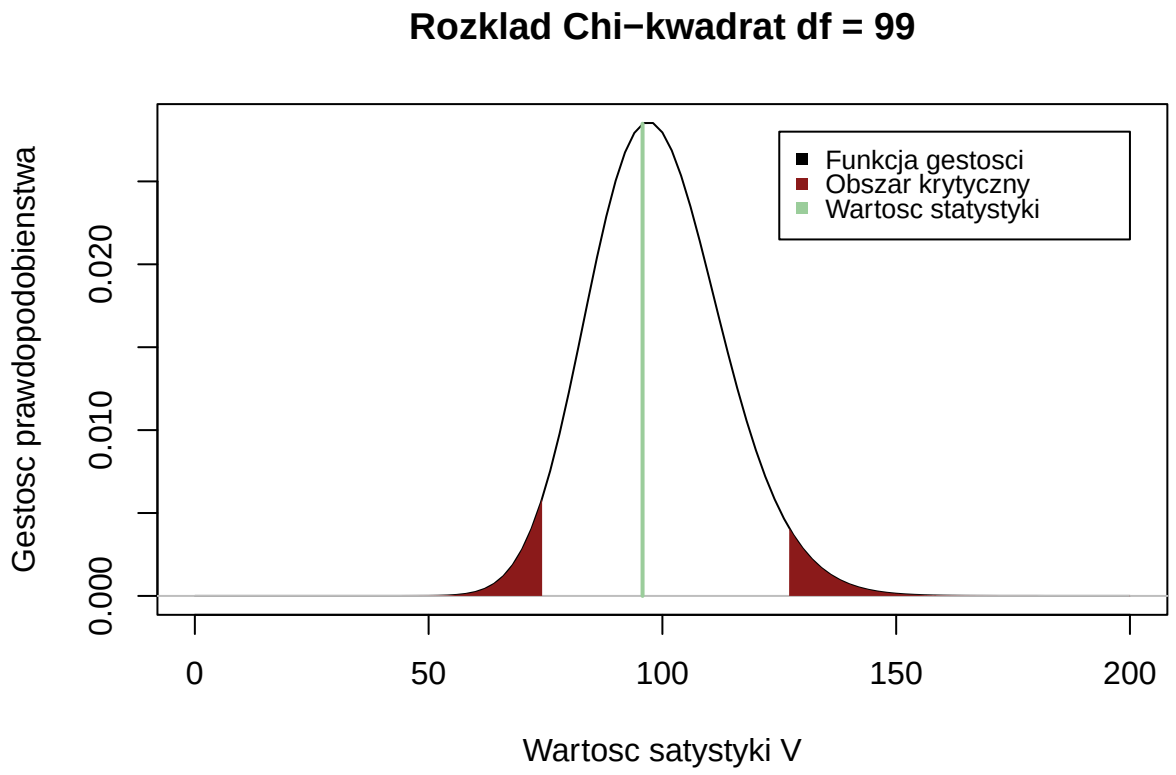
Obliczamy wartość statystyki **V** dla testowanej przez nas hipotezy zgodnie ze wzorem.

```
V1 <- (n*(pf_sd)^2)/(sigma)^2
```

```
[1] 95.75914
```

Dodajemy obliczoną statystykę do wykresu gęstości rozkładu $\chi^2$ o stopniach swobody n - 1, w analogiczny sposób jak w poprzedniej hipotezie dodawaliśmy obliczoną statystykę T do wykresu gęstości rozkładu t-Studenta.

```
x_value <- V1
y_value <- funkcja_chi(V1)
segments(x0 = x_value, y0 = 0, x1 = x_value, y1 = y_value, col = "darkseagreen3", lwd = 2)
legend(x = c(125, 200), y = c(0.028, 0.0215),
       legend = c("Funkcja gęstości", "Obszar krytyczny", "Wartość statystyki"),
       col = c("black", "firebrick4", "darkseagreen3"),
       pch = 15, cex = 0.8, y.intersp = 0.8)
```




Na wykresie jasno widzimy, że obliczona statystyka **V** nie należy do obszaru krytycznego.

Wartość statystyki V: **189.8103**

Przedział lewego obszaru krytycznego: **(0, 141.5425]**

Przedział prawego obszaru krytycznego **[211.8381,** $\infty$**)**

Nie ma podstaw, aby odrzucić hipotezę zerową [**H0**].