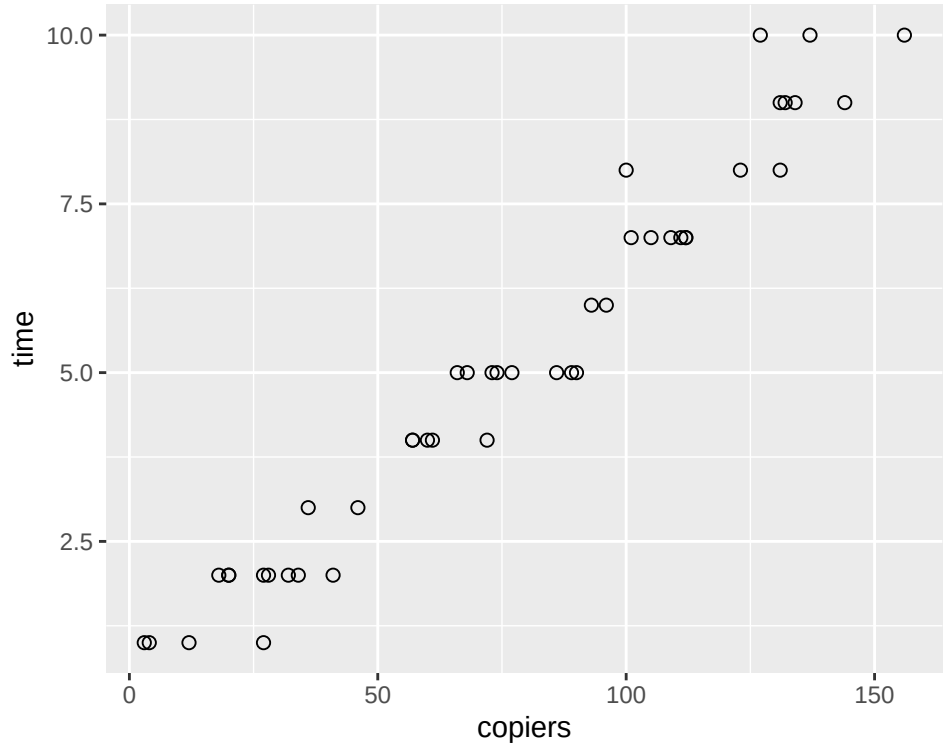
Analizujemy dane pochodzące z pliku `CH01PR20.txt` złożone z dwóch kolumn: liczby kopiarek oraz czasu ich serwisowania (w godzinach).

Zadanie 1. Poniżej przedstawimy wykres zależności czasu (w godzinach) od liczby serwisowanych kopiarek.

```
library(ggplot2)
data <- read.table('CH01PR20.txt', header = FALSE,
                   col.names = c('copiers', 'time'))
ggplot(data, aes(copiers, time)) + geom_point(shape = 1, size = 2)
```



Na podstawie powyższego wykresu możemy zaobserwować, że relacja pomiędzy czasem (w godzinach), a liczbą serwisowanych kopiarek jest w przybliżeniu liniowa.

Zadanie 2. Przejdziemy teraz do zastosowania regresji liniowej prostej dla danych z pliku `CH01PR20.txt` zakładając, że zmienną objaśnianą jest czas, a regresorem jest liczba serwisowanych kopiarek.

```
model <- lm(time ~ copiers, data)
```

Wówczas intercept oraz slope dla zbudowanego modelu wynoszą kolejno:

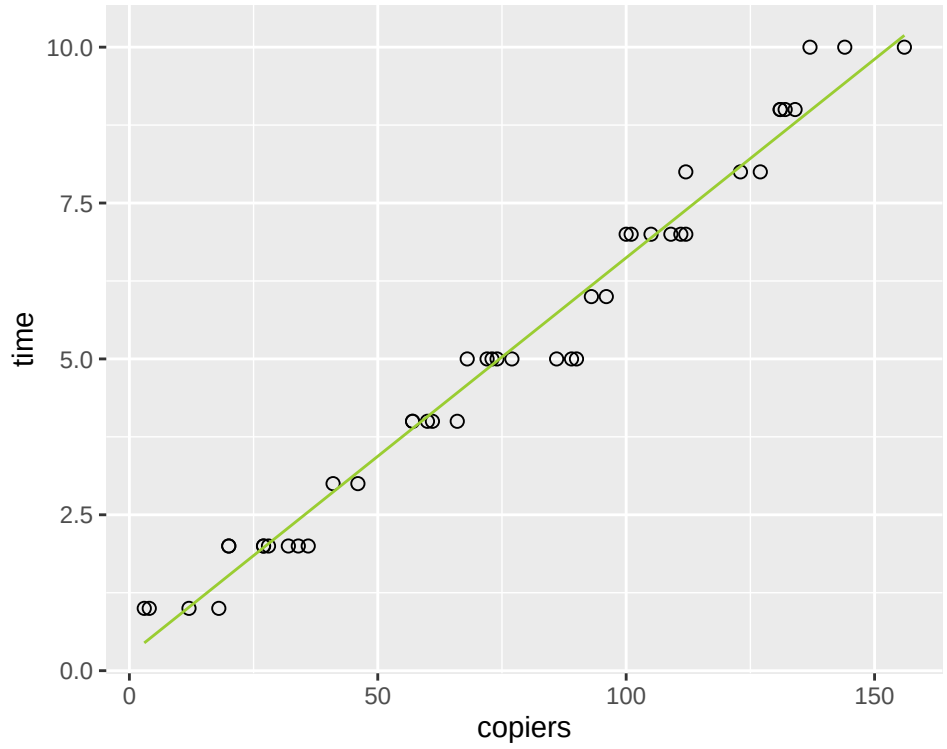
```
## (Intercept)     copiers
##  0.25419165  0.06368338
```

Możemy więc podać estymator równania regresji, tj.

$$Y_i = 0.254 + 0.064 X_i.$$

Na poniższym wykresie przedstawimy dopasowanie uzyskanej prostej do danych.

```
accuracy
```



Dodatkowo, przy użyciu komendy `confint`, poniżej podamy przedział ufności dla slope na poziomie istotności 95%.

```
##       lower      upper
##   0.0595569 0.06780987
```

Otrzymujemy więc, że z prawdopodobieństwem 0.95, estymowany parametr β_1 t.j. slope, znajduje się w przedziale [0.0596, 0.0678].

Na koniec wykonamy test istotności parametru β_1. Testujemy następujący problem

$$H_0: \ \beta_1 = 0 \text{ vs } H_1: \ \beta_1 \neq 0.$$

Do wykonania testu korzystać będziemy z funkcji `summary`.

```
summary(model)

##
## Call:
## lm(formula = time ~ copiers, data = data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.98570 -0.36780 -0.03733  0.40328  1.65802
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.254192   0.178413   1.425    0.161
## copiers     0.063683   0.002046  31.123   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5801 on 43 degrees of freedom
## Multiple R-squared:  0.9575,Adjusted R-squared:  0.9565
## F-statistic: 968.7 on 1 and 43 DF,  p-value: < 2.2e-16
```

Na podstawie uzyskanych danych możemy zaobserwować, że statystyka testowa T dla postawionego wyżej problemu wynosi 31.123, zaś liczba stopni swobody jest równa 43. Ponadto p-wartość jest mniejsza niż $2 \cdot 10^{-16}$, czyli w szczególności jest mniejsza niż 0.05, a zatem możemy odrzucić $H_0 : \ \beta_1 = 0$. Zauważmy, że odrzucenie H_0 w szczególności dostarcza nam istotnej informacji n.t. istnienia związku pomiędzy zmienną objaśnianą (czasem serwisu), a regresorem (liczbą kopiarek).

Zadanie 3. Przejdziemy teraz do estymacji wartości oczekiwanej $\mu_{11} = EY_{11}$. Przypomnijmy, że estymator μ_{11} dany jest następującą zależnością

$$\hat{\mu_{11}} = \hat{\beta}_0 + \hat{\beta}_1 X_{11},$$

gdzie $\hat{\beta}_0, \ \hat{\beta}_1$ wynoszą kolejno

```
## (Intercept)     copiers
##  0.25419165  0.06368338
```

Używając funkcji `predict` możemy otrzymać, że wartość $\hat{\mu_{11}}$ oraz odpowiadający jej przedział ufności na poziomie istotności 95% wynoszą

```
predict(model, data.frame(copiers = 11), interval = 'confidence')
```

```
##         fit       lwr      upr
## 1 0.9547089 0.6338524 1.275565
```

Podsumowując, otrzymaliśmy, że $\hat{\mu_{11}} = 0.955$ oraz rzeczywista wartość μ_{11} z prawdopodobieństwem 0.95 znajduje się w przedziale $[0.634, \ 1.276]$.

Zadanie 4. Przejdziemy teraz do estymacji wartości zmiennej zależnej Y_{11}. Podobnie jak w poprzednim zadaniu, estymator Y_{11} dany jest następującą zależnością

$$\hat{Y_{11}} = \hat{\beta}_0 + \hat{\beta}_1 X_{11},$$

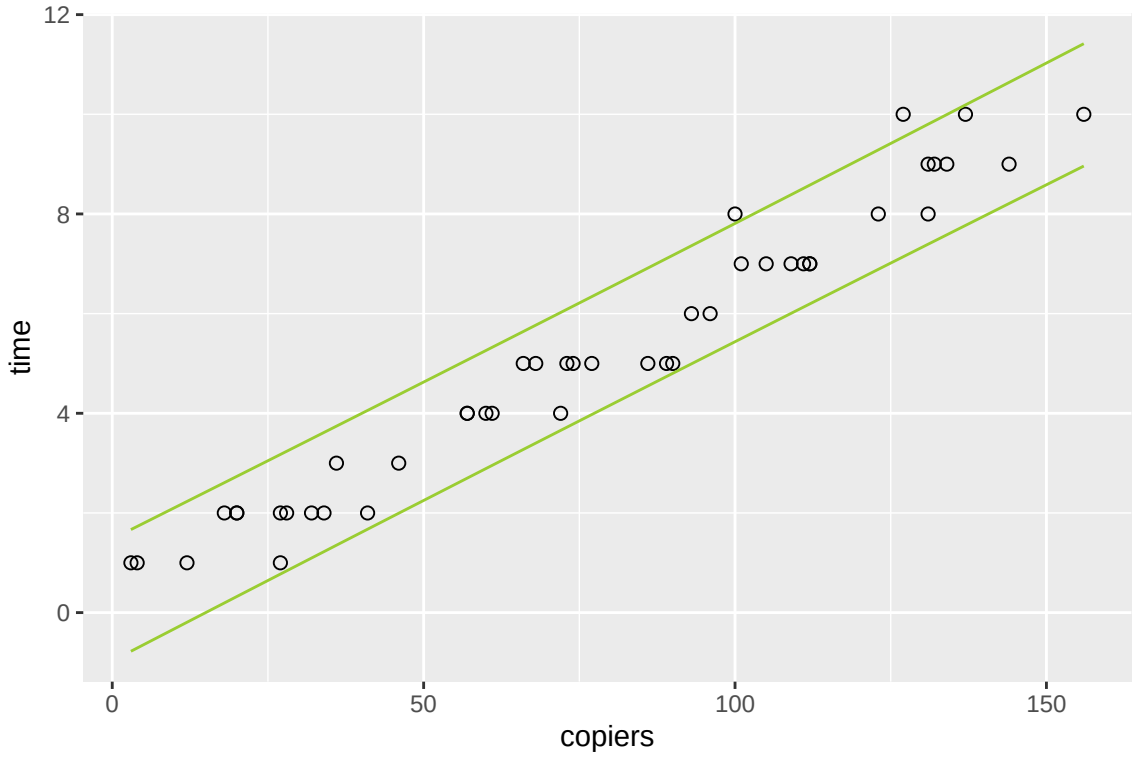
a zatem $\hat{Y_{11}} = 0.955$. W tym przypadku jednak przedział predykcyjny jest szerszy, niż uzyskany wcześniej przedział ufności dla $\hat{\mu_{11}}$, co wynika z faktu, że wariancja błędu predykcji jest większa od wariancji estymatora EY_{11}. Istotnie, używając funkcji `predict`, otrzymujemy, że

```
predict(model, data.frame(copiers = 11), interval = 'prediction')
```

```
##         fit        lwr      upr
## 1 0.9547089 -0.2583855 2.167803
```

Zadanie 5. Poniżej przedstawimy wykres danych z pliku `CH01PR20.txt` wraz z 95% przedziałami predykcyjnymi dla danych obserwacji.

```
pred
```



Możemy zauważyć, że jedynie dwie obserwacje leżą poza przedziałem predykcyjnym, co świadczy o dość dobrym dopasowaniu prostej regresji.

Zadanie 6. Załóżmy, że $n = 40,\ \sigma^2 = 120,\ SSX = \sum(X_i - \bar{X})^2 = 1000.$

```
n <- 40
sigma2 <- 120
SSX <- 1000
alpha <- 0.05
```

Naszym celem będzie wyznaczenie mocy testu $\pi(1)$ badającego następujący problem na poziomie istotności $\alpha = 0.05$

$$H_0 :\ \beta_1 = 0 \ vs\ H_1 :\ \beta_1 = 1,$$

t.j. obliczenie prawdopodobieństwa odrzucenia H_0, gdy prawdziwa jest hipoteza H_1. Przypomnijmy, że

$$\pi(1) = \mathrm{P}_{\beta_1=1}(|T| > t_c) = \mathrm{P}_{\beta_1=1}(T < -t_c) + \mathrm{P}_{\beta_1=1}(T > t_c) = \mathrm{F}_{\beta_1=1}(-t_c) + 1 - \mathrm{F}_{\beta_1=1}(t_c),$$

gdzie $t_c = t^*(1 - \frac{\alpha}{2}, n-2)$, a T ma niecentralny rozkład studenta z 38 stopniami swobody i parametrem niecentralności $\delta = 1/\sigma(\hat{\beta}_1)$. Obliczamy więc $\sigma^2(\hat{\beta}_1) = \sigma^2/SSX$

```
s2beta1 <- sigma2/SSX
```

i otrzymujemy, że $\sigma(\hat{\beta}_1)$ wynosi

```
## [1] 0.3464102
```

Stąd dostajemy, że parametr niecentralności $\delta = \beta_1/\sigma(\hat{\beta}_1)$

```
delta <- 1/sqrt(s2beta1)
```

jest równy

```
## [1] 2.886751
```

Podstawiając otrzymane wartości do podanego powyżej wzoru na $\pi(1)$

```
tc <- qt(1-alpha/2, n - 2)
power <- pt(-tc, n - 2, delta) + 1 - pt(tc, n - 2, delta)
```

otrzymujemy, że moc testu jest równa

```
## [1] 0.8032105
```

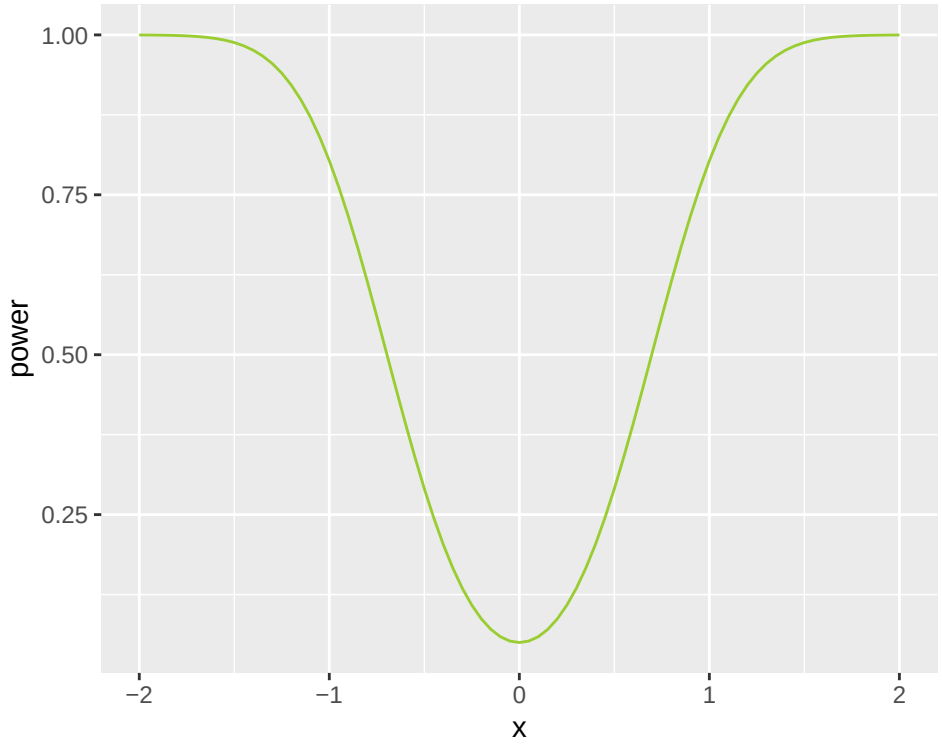
W kolejnym kroku powtarzając powyższe rozumowanie, zdefininujemy funkcję mocy π dla ustalonych wcześniej parametrów $n = 40,\ \sigma^2 = 120,\ SSX = \sum(X_i - \bar{X})^2 = 1000$.

```
n <- 40
sigma2 <- 120
SSX <- 1000
alpha <- 0.05
s2beta1 <- sigma2/SSX
tc <- qt(1-alpha/2, n - 2)

power <- function(x) {
delta <- x/sqrt(s2beta1)
return(pt(-tc, n - 2, delta) + 1 - pt(tc, n - 2, delta))
}
```

Wykres funkcji π dla wartości β_1 z przedziału od -2 do 2 wygląda wówczas następująco:

```
options(warn=-1)
x <- seq(-2.0, 2.0, by = .05)
df <- data.frame(x, power = sapply(x, power))
ggplot(df, aes(x, power)) + geom_line(col = "yellowgreen")
```




```
options(warn=0)
```

Zadanie 7. Zaczniemy od wygenerowania wektora

$$X = (X_1, \ldots, X_{200})^T \sim N(0, \frac{1}{200}\mathbb{1}).$$

```
X <- rnorm(200, 0, sqrt(1/200))
```

Następnie generujemy 1000 wektorów Y z modelu

$$Y_i = 5 + \epsilon_i,$$

gdzie $i = 1, 2, \ldots, 1000$, $\epsilon_i \sim N(0, \mathbb{1})$. Na podstawie każdego z wygenerowanych wektorów testować będziemy, na poziomie ufności 95%, hipotezę $\beta_1 = 0$. Następnie, na podstawie uzyskanych rezultatów, szacować będziemy prawdopodobieństwo odrzucenia w.w. hipotezy.

```
licznik <- 0

for (i in 1:1000) {
  Y <- sapply(X, function(x) 5 + rnorm(1, 0, 1))
  dane <- data.frame(x = X, y = Y)
  model <- lm(y ~ x,dane)
  if ((confint(model,"x")[1] > 0)||(confint(model,"x")[2] < 0)) licznik <- licznik + 1
}
```

Otrzymane prawdopodobieństwo odrzucenia hipotezy $\beta_1 = 0$ wynosi

```
## [1] 0.042
```

Widzimy więc, że jest ono bliskie teoretycznej wartości błędu I rodzaju t.j. 5%.
W kolejnej części powtórzymy powyższy eksperyment dla wektorów Y z modelu

$$Y_i = 5 + \epsilon_i,$$

gdzie $i = 1, 2, \ldots, 1000$, $\epsilon_i \sim \exp(1)$.

```
X <- rnorm(200, 0, sqrt(1/200))
licznik <- 0

for (i in 1:1000) {
  Y <- sapply(X, function(x) 5 + rexp(1))
  dane <- data.frame(x = X, y = Y)
  model <- lm(y ~ x,dane)
  if ((confint(model,"x")[1] > 0)||(confint(model,"x")[2] < 0)) licznik <- licznik + 1
}
```

Otrzymane prawdopodobieństwo odrzucenia hipotezy $\beta_1 = 0$ wynosi

```
## [1] 0.064
```

Widzimy więc, że znów jest ono bliskie teoretycznej wartości błędu I rodzaju t.j. 5%.
Przejdziemy teraz do powtórzenia eksperyment dla wektorów Y z modelu

$$Y_i = 5 + 1.5X_i + \epsilon_i,$$

gdzie $i = 1, 2, \ldots, 1000$, $\epsilon_i \sim N(0, \mathbb{1})$.

```
X <- rnorm(200, 0, sqrt(1/200))
b1 <- 1.5
licznik <- 0
power <- c()

for (i in 1:1000) {
  Y <- sapply(X, function(x) 5 + b1*x + rnorm(1, 0, 1))
  dane <- data.frame(x = X, y = Y)
  model <- lm(y ~ x,dane)
  if ((confint(model,"x")[1] > 0)||(confint(model,"x")[2] < 0)) licznik <- licznik + 1
  s <- sd(model$residuals) * sqrt(199/198)
  sigmab <- s/(var(dane$x) * 199)
  delta <- b1/sigmab
  power[i] <- 1 - pt(qt(1 - .05/2, 198), 198, delta) +
    pt(-qt(1 - .05/2, 198), 98, delta)
}

prob <- licznik/1000
```

Otrzymane prawdopodobieństwo odrzucenia hipotezy $\beta_1 = 0$ wynosi

```
## [1] 0.3
```

Widzimy więc, że jest ono bliskie teoretycznej wartości mocy testu dla $\beta_1 = 1.5$, która wynosi

```
## [1] 0.273
```

Przejdziemy teraz do powtórzenia eksperyment dla wektorów Y z modelu

$$Y_i = 5 + 1.5X_i + \epsilon_i,$$

gdzie $i = 1, 2, \ldots, 1000$, $\epsilon_i \sim \exp(1)$.

```
X <- rnorm(200, 0, sqrt(1/200))
b1 <- 1.5
licznik <- 0
power <- c()

for (i in 1:1000) {
  Y <- sapply(X, function(x) 5 + b1*x + rexp(1))
  dane <- data.frame(x = X, y = Y)
  model <- lm(y ~ x,dane)
  if ((confint(model,"x")[1] > 0)||(confint(model,"x")[2] < 0)) licznik <- licznik + 1
  s <- sd(model$residuals) * sqrt(199/198)
  sigmab <- s/(var(dane$x) * 199)
  delta <- b1/sigmab
  power[i] <- 1 - pt(qt(1 - .05/2, 198), 198, delta) +
    pt(-qt(1 - .05/2, 198), 98, delta)
}

prob <- licznik/1000
```

Otrzymane prawdopodobieństwo odrzucenia hipotezy $\beta_1 = 0$ wynosi

```
## [1] 0.309
```

Widzimy więc, że jest ono bliskie teoretycznej wartości mocy testu dla $\beta_1 = 1.5$, która wynosi

```
## [1] 0.315
```

Zadanie 8. Rozważamy próbę $n = 20$ elementową pochodzącą z modelu liniowego

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Ponadto zakładamy, że estymatorami parametrów są

$$b_0 = 1, \quad b_1 = 3, \; s(b_1) = 1 \text{ oraz } s = 4.0.$$

Zaczniemy od skonstruowania 95%-procentowego przedziału ufności dla β_1. Jego górna granica wyniesie:

$$b_1 + t_c \, s(b_1) = 3 + 2.1 = 5.1$$

zaś dolna:

$$b_1 - t_c \, s(b_1) = 3 - 2.1 = 0.9.$$

W dalszej części testować będziemy hipotezę

$$\mathrm{H}_0 : \; \beta_1 = 0$$

Statystyka testowa T wynosi $\frac{b_1}{s(b_1)} = 3$, a zatem łatwo zauważyć, że $|T| = 3 > 2.1 = t_c$. Odrzucamy więc hipotezę H_0. Zauważmy, że w ten sposób w szczególności otrzymujemy statystyczny dowód na zależność Y od X.

W dalszej części zakładać będziemy, że 95%-procentowy przedział ufności dla $E(Y)$, gdzie $X = 5$, wynosi $[13,\ 19]$. Naszym celem będzie znalezienie odpowiadającego przedziału predykcyjnego. Przypomnijmy, że wzór na przedział ufności dla $E(Y)$ jest następującej postaci

$$\hat{\mu_5} \pm t_c s(\hat{\mu_5}).$$

Zauważmy, że stąd otrzymujemy, że wartość $\hat{\mu_5}$ wynosi 16. Dodatkowo, wprost ze wzorów na $s(pred)$ i $s(\hat{\mu_5})$ dostajemy, że

$$s^2(pred) = s^2(\hat{\mu_5}) + s^2.$$

A zatem pozostaje wyznaczyć $s^2(\hat{\mu_5})$. Łatwo widzimy, że $s^2(\hat{\mu_5}) = (3/t_c)^2$ co wynosi w zaokrągleniu 2.04. Ostatecznie otrzymujemy, że szukany przedział predykcyjny jest równy w przybliżeniu

$$[7.075, 24.925].$$