Statystyka - raport 4

Łukasz Rębisz

22.12.2022

Zadania 1, 3, 5

Wyznaczmy przedziały ufności dla **różnicy dwóch** średnich w modelu normalnym. Załóżmy, że **wariancje są znane**. Niech $X_1, ..., X_n \;\; i.i.d. \sim N(\mu_X, \sigma_X^2)$; $\;\; Y_1, ..., Y_m \;\; i.i.d. \sim N(\mu_Y, \sigma_Y^2)$ to niezależne próby losowe.

Wówczas zmienna $Z = \frac{(\bar{X}-\mu_X)-(\bar{Y}-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \sim N(0,1)$.

Czyli $1-\alpha = P(Z_{\alpha/2}^{-1} \leq Z \leq Z_{1-\alpha/2}^{-1})$, stąd przedział ufności dla $(\mu_x - \mu_y)$ na poziomie ufności $1-\alpha$ to $(\bar{X}-\bar{Y}) \pm Z_{1-\alpha/2}^{-1}\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}$.

Załóżmy następnie, że **wariancje nie są znane**, ale mimo tego zakładamy, że są **równe**.

Wówczas nieobciążony estymator wariancji $s^2 = \frac{(n-1)s_X^2+(m-1)s_Y^2}{n+m-2} \sim \sigma^2 \cdot \chi^2(n+m-2)$. Uśredniony błąd standardowy $(U)SE = s\sqrt{\frac{1}{n}+\frac{1}{m}}$.

Zmienna $T = \frac{Z}{s/\sigma} \sim t(n+m-2)$.

Zatem przedział ufności dla różnicy średnich to $(\bar{X}-\bar{Y}) \pm t_{1-\alpha/2}^{-1}(n+m-2) \cdot (U)SE$.

Jeśli natomiast **nie zakładamy równości wariancji** (i nie znamy wariancji), to chcąc przybliżyć zmienną T rozkładem Studenta, należy wyznaczyć liczbę stopni swobody dla tego rozkładu. Mamy:

$$T = \frac{(\bar{X}-\mu_X)-(\bar{Y}-\mu_Y)}{\sqrt{\frac{s_X^2}{n}+\frac{s_Y^2}{m}}} = \frac{Z \cdot \sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}{\sqrt{\frac{1}{n}\frac{\sigma_X^2 \cdot \chi^2(n-1)}{n-1}+\frac{1}{m}\frac{\sigma_Y^2 \cdot \chi^2(m-1)}{m-2}}} \stackrel{?}{=} \sqrt{\frac{k}{\chi^2(k)}} \cdot Z \sim t(k)$$

Wyznaczmy k, dla jakiego zachodzi powyższa równość.

Oznaczmy $L := \frac{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}{\sqrt{\frac{1}{n}\frac{\sigma_X^2 \cdot \chi^2(n-1)}{n-1}+\frac{1}{m}\frac{\sigma_Y^2 \cdot \chi^2(m-1)}{m-2}}}, \quad P := \sqrt{\frac{k}{\chi^2(k)}}$.

Obliczając wartości oczekiwane mamy: $\mathbb{E}[L] = 1 = \mathbb{E}[P]$.

$$Var[P] = \frac{2}{k}$$

$$Var[L] = \frac{\frac{\sigma_X^4}{n^2}\frac{2}{n-1}+\frac{\sigma_Y^4}{m^2}\frac{2}{m-1}}{\left(\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}\right)^2}$$

$$\text{Stąd } k \approx \frac{\left(\frac{s_X^2}{n}+\frac{s_Y^2}{m}\right)^2}{\frac{s_X^4}{n^2}\frac{1}{n-1}+\frac{s_Y^4}{m^2}\frac{1}{m-1}}.$$

Zatem przedział ufności w tym przypadku to:

$$(\bar{X} - \bar{Y}) \pm t^{-1}_{1-\alpha/2}(k) \cdot (N)SE,$$

gdzie $(N)SE = \sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}$.

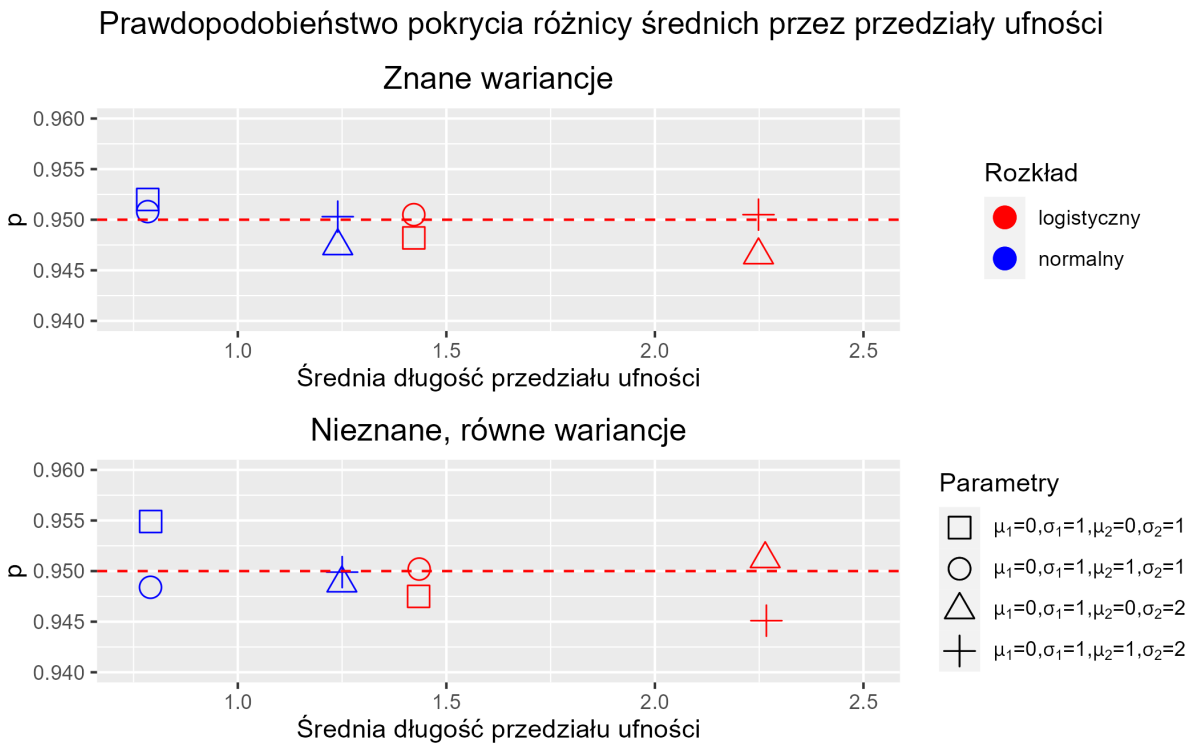
Zadania 2, 4, 6

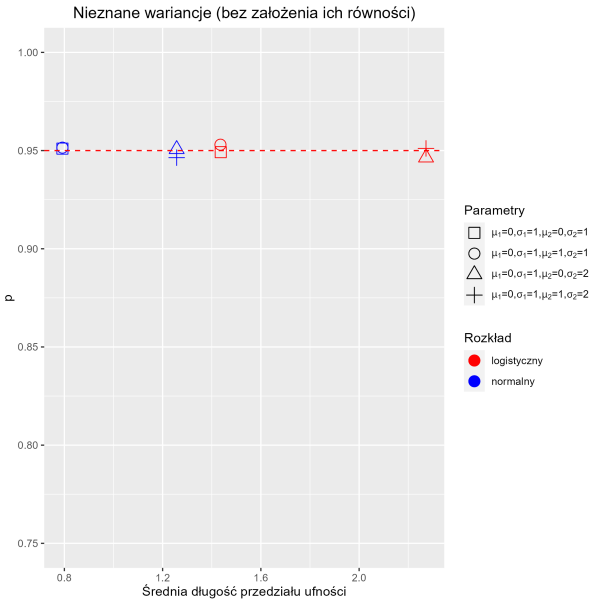
Implementacja zgodnie z powyższymi wzorami:

```
conf_interval <- function(n1, n2, distribution1, parameters1, distribution2, parameters2,
                          alpha, var_known, var_equal){

  if(distribution1=="norm"){
    X1 <- rnorm(n1, mean = parameters1[1], sd = parameters1[2])
    mu1 <- parameters1[1]
    sigma1 <-  (parameters1[2])^2
  }

  if(distribution2=="norm"){
    X2 <- rnorm(n2, mean = parameters2[1], sd = parameters2[2])
    mu2 <- parameters2[1]
    sigma2 <-  (parameters2[2])^2
  }

  if(distribution1=="logis"){
    X1 <- rlogis(n1, location = parameters1[1], scale = parameters1[2])
    mu1 <- parameters1[1]
    sigma1 <- ((parameters1[2]*pi)^2)/3
  }

  if(distribution2=="logis"){
    X2 <- rlogis(n2, location = parameters2[1], scale = parameters2[2])
    mu2 <- parameters2[1]
    sigma2 <- ((parameters2[2]*pi)^2)/3
  }

  if(var_known){
    u <- mean(X1) - mean(X2)
    Z_c <- qnorm(1-alpha/2)
    SE <- sqrt(sigma1/n1 + sigma2/n2)
    l <- u - Z_c*SE
    r <- u + Z_c*SE
    len <- 2*Z_c*SE
    diff <- mu1 - mu2
    diff_in_interval <- ((diff>=l) && (diff<=r))
    return(c(l, r, len, diff_in_interval))
  } else{
    if(var_equal){
      u <- mean(X1) - mean(X2)
      t_c <- qt(1-alpha/2, n1 + n2 - 2)
      s_c <- sqrt(((n1-1)*var(X1)+(n2-1)*var(X2))/(n1 + n2 - 2))
      USE <- s_c*sqrt(1/n1 + 1/n2)
      l <- u - t_c*USE
```

```
      r <- u + t_c*USE
      len <- 2*t_c*USE
      diff <- mu1 - mu2
      diff_in_interval <- ((diff>=l) && (diff<=r))
      return(c(l, r, len, diff_in_interval))
    }
    if(! var_equal){
      u <- mean(X1) - mean(X2)
      s1_2 <- var(X1)
      s2_2 <- var(X2)
      k <- ((s1_2/n1 + s2_2/n2)^2)/((s1_2^2/n1^2)/(n1-1) + (s2_2^2/n2^2)/(n2-1))
      t_c <- qt(1-alpha/2, k)
      NSE <- sqrt(s1_2/n1 + s2_2/n2)
      l <- u - t_c*NSE
      r <- u + t_c*NSE
      len <- 2*t_c*NSE
      diff <- mu1 - mu2
      diff_in_interval <- ((diff>=l) && (diff<=r))
      return(c(l, r, len, diff_in_interval))
    }
  }
}
```

Dla danych rozkładów (normalnego, logistycznego) o zadanych parametrach wyznaczmy przedziały ufności dla różnicy średnich. Powtórzmy doświadczenie 10 000 razy i na tej podstawie oszacujmy prawdopodobieństwo pokrycia nieznanej różnicy średnich przez zadane powyższymi wzorami przedziały ufności.



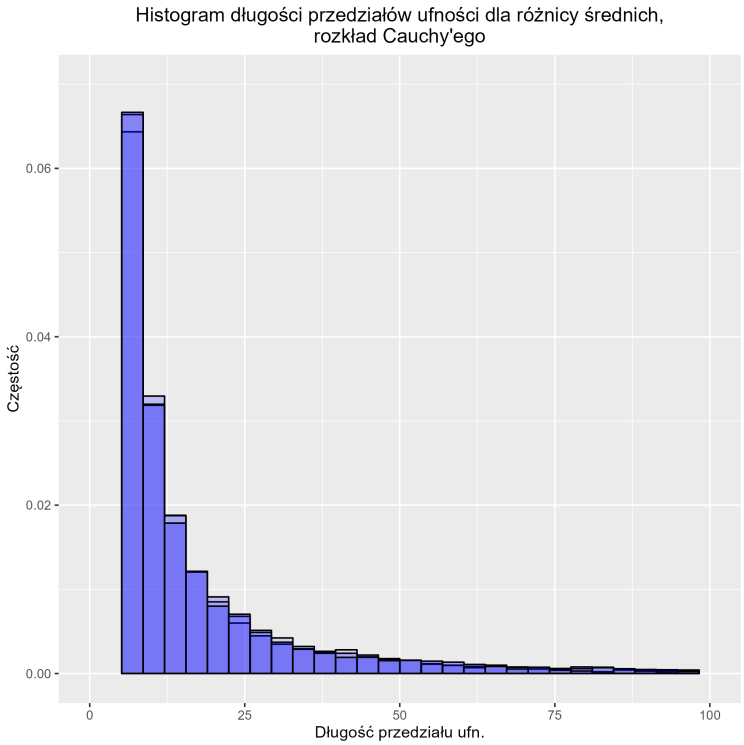


Analiza powyższych wykresów pozwala wysunąć następujące wnioski:

- Konstrukcja przedziału ufności dla różnicy średnich przy założeniu, że wariancje są znane jest dokładna zarówno dla rozkładu normalnego jak i logistycznego. Dla obu rozkładów prawdopodobieństwo pokrycia różnicy przez przedział ufność jest w przybliżeniu równe założonemu poziomowi ufności 0,95. Dla rozkładu logistycznego otrzymaliśmy jednak dłuższe przedziały ufności.
- Założenie, że wariancje nie są znane, ale są równe prowadzi do analogicznych wyników i wniosków jak powyżej. Jest to jednak niepoprawne podejście dla ostatnich dwóch podpunktów. W przyjętym modelu zakładmy wówczas równość wariancji przy założonym parametrze skali $\sigma_1 = 1 \neq 2 = \sigma_2$.
- Metoda niezakładająca równości wariancji dobrze sprawdza się dla obu rozkładów w pierwszych dwóch podpunktach tzn. dla równych parametrów skali $\sigma = 1$. W kolejnych podpunktach $\sigma_1 = 1, \sigma_2 = 2$, co znaczącą wpływa na rozrzut wyników. W tych przypadkach otrzymujemy przedziały ufności na poziomie istotności ok. 0,8.

Rozkład Cauchy'ego

W przypadku **rozkładu Cauchy'ego** brak wartości oczekiwanej i wariancji. Nie możemy sprawdzić, czy teoretyczna wartość różnicy średnich znajduje się w przedziale ufności. Możemy natomiast zbadać **długości przedziałów** otrzymywane dla tego rozkładu w modelu normalnym o nieznanej wariancji (załóżmy ponadto, że wariancje nie muszą być równe).



Analiza powyższych histogramów dla rozkładów Cauchy'ego o parametrach:

i) $\mu_1 = 0, \quad \sigma_1 = 1, \quad \mu_2 = 0, \quad \sigma_2 = 1;$

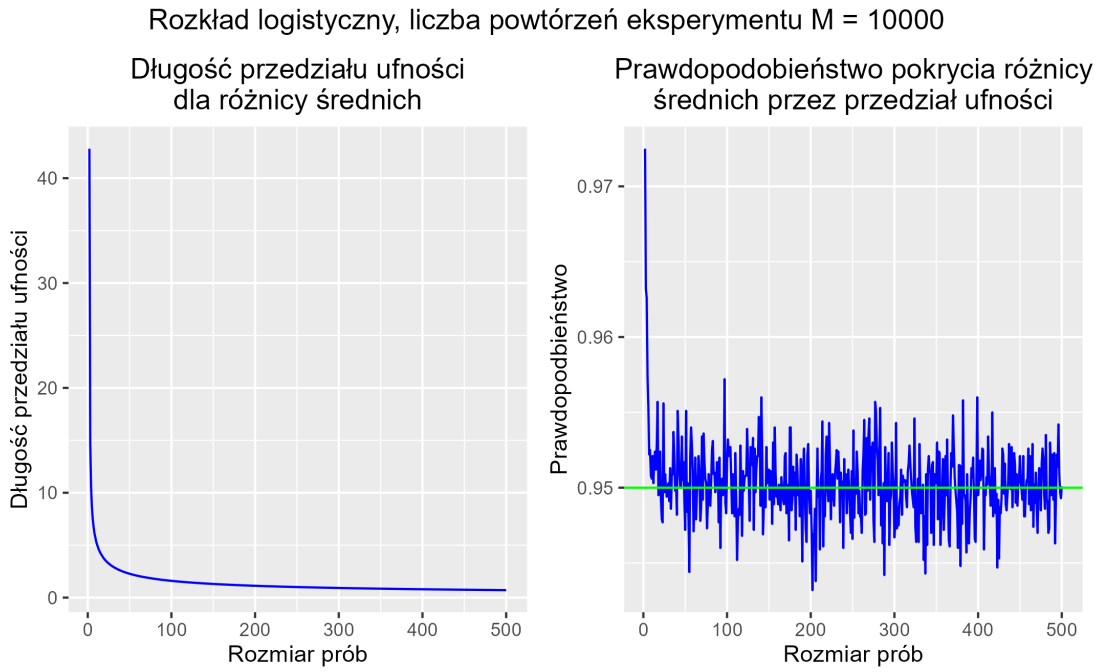
ii) $\mu_1 = 0, \quad \sigma_1 = 1, \quad \mu_2 = 0, \quad \sigma_2 = 2;$

iii) $\mu_1 = 0, \quad \sigma_1 = 1, \quad \mu_2 = 1, \quad \sigma_2 = 2;$

wskazuje na ciężkie ogony dla rozkładu długości przedziałów ufności dla różnicy średnich. Uzyskiwane długości przedziałów są często bardzo duże. Skala na osi została ograniczona do długości przedziału równej 100. Nie jest to maksimum uzyskiwanych wyników. Uzyskane wyniki świadczą o tym, że konstruowanie przedziałów ufności dla różnicy średnich w modelu normalnym **NIE jest wiarygodne dla rozkładu Cauchy'ego**.

Dokładność wyników w zależności od rozmiaru próby

Ustalmy, że rozkład, który chcemy przybliżyć rozkładem normalnym to **rozkład logistyczny**. Sprawdźmy, jak dokładne przedziały ufności dla różnicy średnich uzyskamy w modelu normalnym (o nieznanych wariancjach, bez założenia ich równości) dla zmiennych o parametrach $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 1, \sigma_2 = 2$



Analiza powyższych wykresów pokazuje, że dla **rozkładu logistycznego** pomimo spadku długości przedziałów ufności wraz ze wzrostem rozmiaru prób (w tempie $\frac{1}{\sqrt{n}}$), prawdopodobieństwo pokrycia różnicy średnich przez przedział ufności spada i stabilizuje się na poziomie ok. 0,7. Świadczy to o niedokładym przybliżeniu rozkładu logistycznego modelem normalnym.

Zadania 7, 9

Naszym zadaniem jest wyznaczenie przedziału ufności dla **ilorazu dwóch wariancji** w modelu normalnym. Załóżmy, że **nie są znane średnie** badanych rozkładów.

Niech $X_1, ..., X_n \;\; i.i.d. \sim N(\mu_X, \sigma_X^2)$; $\;\; Y_1, ..., Y_m \;\; i.i.d. \sim N(\mu_Y, \sigma_Y^2)$ to niezależne próby losowe. Wówczas $\frac{(n-1)s_X^2}{\sigma_X^2} \sim \chi^2_{n-1}$, $\;\; \frac{(m-1)s_Y^2}{\sigma_Y^2} \sim \chi^2_{m-1}$. Wówczas (przy założeniu niezależności zmiennych) statystyka

$$F = \frac{\frac{(m-1)s_Y^2}{\sigma_Y^2}/(m-1)}{\frac{(n-1)s_X^2}{\sigma_X^2}/(n-1)} = \frac{\sigma_X^2 \cdot s_Y^2}{\sigma_Y^2 \cdot s_X^2} \sim F(m-1, n-1).$$

Zatem $1-\alpha = P\left[F^{-1}_{\alpha/2}(m-1,n-1) \leq \frac{\sigma_X^2 \cdot s_Y^2}{\sigma_Y^2 \cdot s_X^2} \leq F^{-1}_{1-\alpha/2}(m-1,n-1)\right]$,

stąd $1-\alpha = P\left[\frac{s_X^2}{s_Y^2} \cdot F^{-1}_{\alpha/2}(m-1,n-1) \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{s_X^2}{s_Y^2} \cdot F^{-1}_{1-\alpha/2}(m-1,n-1)\right]$.

W przypadku **znanych średnich** estymatory wariancji są nieobciążone, tzn.
$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2$, $s_Y^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \mu_Y)^2$. Wówczas przedział ufności dla ilorazu wariancji to:

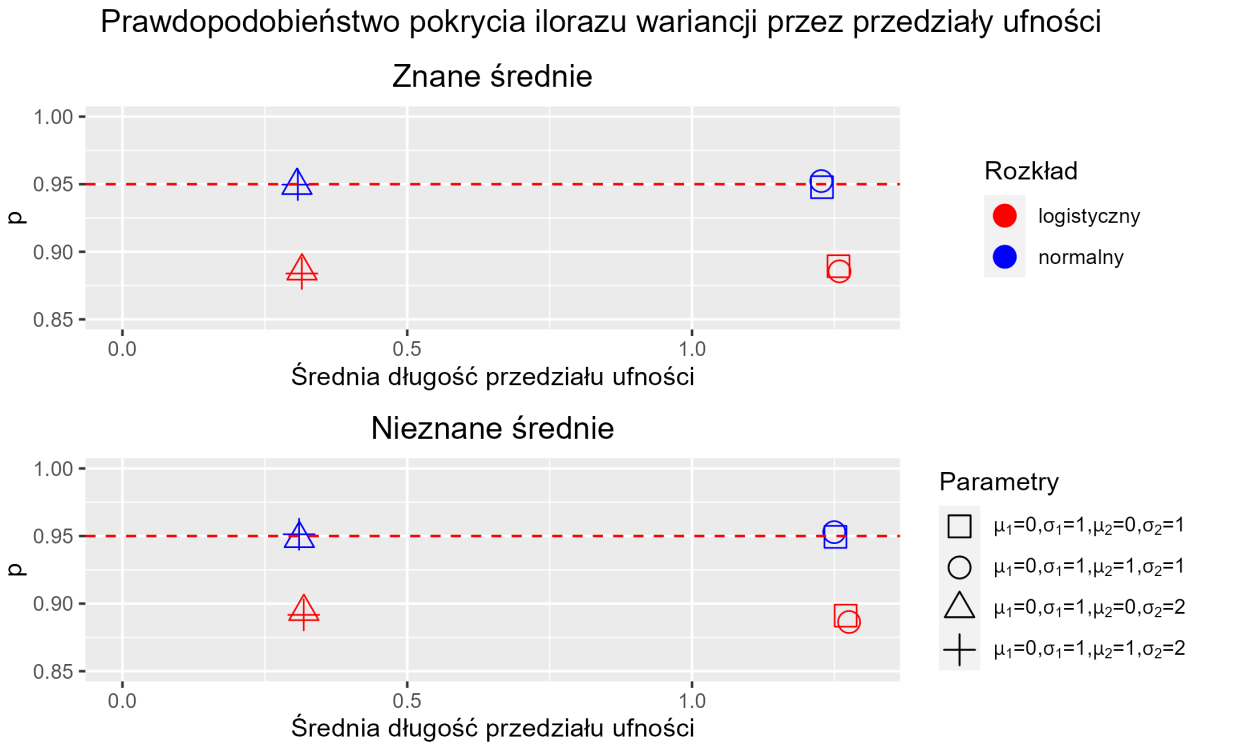
$$1 - \alpha = P\left[\frac{s_X^2}{s_Y^2} \cdot F^{-1}_{\alpha/2}(m,n) \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{s_X^2}{s_Y^2} \cdot F^{-1}_{1-\alpha/2}(m,n)\right].$$

Zadania 8, 10

Implementacja powyższych wzorów:

```
conf_interval_s <- function(n1, n2, distribution1, parameters1, distribution2, parameters2,
                            alpha, mean_known){

  if(distribution1=="norm"){
    X1 <- rnorm(n1, mean = parameters1[1], sd = parameters1[2])
    mu1 <- parameters1[1]
    sigma1 <-  (parameters1[2])^2
  }

  if(distribution2=="norm"){
    X2 <- rnorm(n2, mean = parameters2[1], sd = parameters2[2])
    mu2 <- parameters2[1]
    sigma2 <-  (parameters2[2])^2
  }

  if(distribution1=="logis"){
    X1 <- rlogis(n1, location = parameters1[1], scale = parameters1[2])
    mu1 <- parameters1[1]
    sigma1 <- ((parameters1[2]*pi)^2)/3
  }

  if(distribution2=="logis"){
    X2 <- rlogis(n2, location = parameters2[1], scale = parameters2[2])
    mu2 <- parameters2[1]
    sigma2 <- ((parameters2[2]*pi)^2)/3
  }

  if(mean_known){
    s_X1_2 <- (1/n1)*sum((X1-mu1)^2)
    s_X2_2 <- (1/n2)*sum((X2-mu2)^2)
    F_1 <- qf(alpha/2, n2, n1)
    F_2 <- qf(1 - alpha/2, n2, n1)
    l <- (s_X1_2/s_X2_2)*F_1
    r <- (s_X1_2/s_X2_2)*F_2
    len <- r - l
    quot <- sigma1/sigma2
    quot_in_interval <- ((quot >= l) && (quot <= r))
    return(c(l, r, len, quot_in_interval))
  } else{
    s_X1_2 <- (1/n1-1)*sum((X1-mean(X1))^2)
    s_X2_2 <- (1/n2-1)*sum((X2-mean(X2))^2)
    F_1 <- qf(alpha/2, n2-1, n1-1)
    F_2 <- qf(1 - alpha/2, n2-1, n1-1)
    l <- (s_X1_2/s_X2_2)*F_1
    r <- (s_X1_2/s_X2_2)*F_2
```

```
    len <- r - l
    quot <- sigma1/sigma2
    quot_in_interval <- ((quot >= l) && (quot <= r))
    return(c(l, r, len, quot_in_interval))
  }
}
```

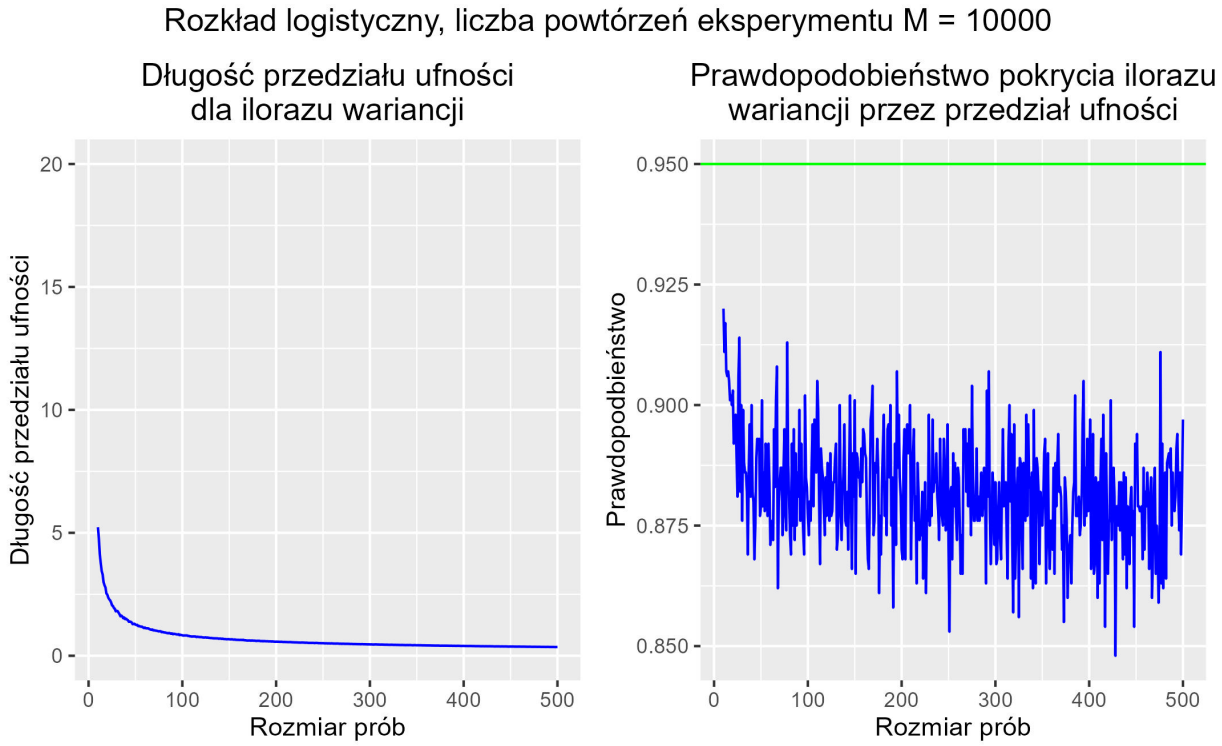


Analiza powyższych wykresów pokazuje, że przedziały ufności dla ilorazu wariancji dla **rozkładów normalnych** są wiarygodne - prawdopodbieństwo pokrycia ilorazu przez przedział ufności jest równe w przybliżeniu założonemu poziomowi istotnosci 0,95. W przypadku różniących się wariancji przedziały ufności dla ilorazu wariancji mają mniejsze długości. To samo zjawisko obserwujemy w **rozkładzie logistycznym**. Rozkład ten ma jednak mniejsze prawdopodobieństwo pokrycia ilorazu przez przedział ufności. W każdym podpunkcie wynosi niespełna 0,9.

Dla **rozkładu Cauchy'ego** wariancja jest nieokreślona. Nie ma więc sensu konstruować przedziałów ufności dla ilorazu wariancji dla prób z tego rozkładu.

Dokładność wyników w zależności od rozmiaru próby

Ustalmy, że rozkład, który chcemy przybliżyć rozkładem normalnym to **rozkład logistyczny**. Sprawdźmy, jak dokładne przedziały ufności dla ilorazu wariancji uzyskamy w modelu normalnym (o nieznanych średnich) dla zmiennych o parametrach $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 1, \sigma_2 = 1$



Analiza powyższych wykresów pokazuje, że długości przedziałów ufności dla ilorazu wariancji rosną wraz ze wzrostem rozmiaru prób. Prawdopodobieństwo pokrycia ilorazu wariancji przez przedział ufności rośnie, uzyskując maksimum bliskie założonemu poziomowi ufności dla rozmiaru prób równego około 50. Następnie prawdopodobieństwo spada do wartości równej 0.

Zadanie 12

W sytuacji, gdy znamy rozkład pewnej zmiennej losowej i chcemy wyznaczyć rozkład funkcji tej zmiennej losowej, przydatna jest następująca metoda (zwana **Metodą Delta**):

Twierdzenie:

Niech $\{X_n\}$ będzie ciągiem zmiennych losowych, takich że:

$$\sqrt{n}(X_n - \theta) \xrightarrow{D} N(0, \sigma^2).$$

Załóżmy, że funkcja $g(x)$ jest różniczkowalna w punkcie θ oraz $g'(\theta) \neq 0$. Wówczas:

$$\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2(g'(\theta))^2).$$

Przykład wykorzystania

Niech $Z_n \sim \chi^2(n)$. Wówczas, wiedząc że zmienne Z_n spełniają warunek

$$\sqrt{n}\left(\frac{1}{\sqrt{2n}}Z_n - \frac{1}{\sqrt{2}}\right) \xrightarrow{D} N(0, \sigma^2),$$

mamy:

Niech $g(t) = \sqrt{t}$ oraz $W_n = g(Z_n/(\sqrt{2}n)) = (Z_n/(\sqrt{2}n))^{1/2}$.

Zauważmy, że $g(1/\sqrt{2}) = (\frac{1}{2})^{1/4}, \quad g'(1/\sqrt{2}) = 2^{-3/4}$. Wówczas spełnione są założenia Metody Delta, więc zgodnie z powyższym twierdzeniem:

$$\sqrt{n}\left[W_n - \left(\frac{1}{2}\right)^{1/4}\right] \xrightarrow{D} N(0, 2^{-3/2}).$$