Lab3 - Wczytywanie danych. Przedziały ufności

Michał Ciach

February 26, 2020

Na poprzednich zajęciach ćwiczyliśmy na danych symulowanych lub wstępnie przygotowanych do pracy w Rstudio. Na dzisiejszych zajęciach wykorzystamy poznane dotychczas techniki do analizy prawdziwych zbiorów danych.

Co to są dane?

Rachunek prawdopodobieństwa zajmuje się badaniem własności *zmiennych losowych*, czyli takich zmiennych, które przyjmują różne wartości z określonym prawdopodobieństwem. Formalnie rzecz biorąc, są to funkcje przypisujące zdarzeniom elementarnym liczby rzeczywiste. Taka definicja zmiennej X pozwala nam ściśle wyrazić co oznaczają napisy postaci $\mathbb{P}(X = 1)$.

Rachunek prawdopodobieństwa pozwala nam określić jakich wyników możemy się spodziewać, jeśli będziemy obserwować jakieś losowe zjawisko, na przykład rzut kostką. Musimy jednak w tym celu określić prawa rządzące obserwowanym zjawiskiem, czyli rozkład prawdopodobieństwa obserwowanej zmiennej losowej. To, jakie jest prawdopodobieństwo przyjęcia określonych wartości, nazywamy *rozkładem prawdopodobieństwa* zmiennej losowej.

W statystyce modelujemy zjawiska losowe również za pomocą zmiennych losowych, ale interesuje nas dokładnie odwrotne pytanie: mając zbiór obserwacji, co możemy wywnioskować na temat praw rządzących obserwowanym zjawiskiem? Formalnie rzecz biorąc, o ile w rachunku prawdopodobieństwa mamy do czynienia z ciągiem zmiennych losowych $X_1, X_2, \ldots, X_n$, to w statystyce mamy do czynienia z ciągiem *realizacji* tych zmiennych, czyli ich wartości przyjętych dla pewnego zdarzenia elementarnego ω: $X_1(\omega), X_2(\omega), \ldots, X_n(\omega)$.

Wszystkie dane, z jakimi mamy do czynienia, traktujemy jako realizacje ciągu niezależnych zmiennych losowych, a interesuje nas, co możemy powiedzieć o (nieznanym) rozkładzie prawdopodobieństwa tych zmiennych. Zdarzenie elementarne ω interpretujemy jako przeprowadzony eksperyment losowy.

Na przykład, załóżmy że obserwujemy serię rzutów kostką. Przed i-tym rzuceniem kostki interpretujemy wynik jako zmienną losowa X_i - nie wiemy, jaka liczba oczek wypadnie, możemy jedynie podać prawdopodobieństwo wyrzucenia danej liczby. Rzucenie kostką jest równoznaczne z ustaleniem zdarzenia elementarnego. Wówczas X_i zamienia się w $X_i(\omega)$ - zaobserwowaną liczbę oczek. Możemy wówczas odpowiadać na pytania dotyczące rozkładu prawdopodobieństwa zmiennej X_i, na przykład wyestymować prawdopodobieństwo wyrzucenia szóstki na tej kostce i sprawdzić, czy rzeczywiście jest równe $1/6$.

W statystyce na ogół nie rozpatrujemy wszystkich możliwych rozkładów. Zamiast tego interesują nas tzw. *rodziny parametryczne*, czyli takie zbiory rozkładów, które są opisane przez kilka ustalonych parametrów. Na przykład, rodzina jednowymiarowych rozkładów normalnych jest w pełni opisana przez dwa parametry: średnią oraz wariancję. Dzięki temu żeby odpowiedzieć na pytanie, jaki rozkład ma badana zmienna losowa, wystarczy wyestymować te parametry, tak jak to robiliśmy na poprzednich zajęciach.

Wczytywanie danych

Jako pierwsze ćwiczenie wczytamy dane z pliku `Zadluzenie gmin.csv` dostępnego na stronie przedmiotu. Plik zawiera dane dotyczące procentowego zadłużenia (w stosunku do dochodów) gmin w Polsce w roku 2015. Rstudio umożliwia dwa sposoby wczytywania danych.

Sposób 1. Do wczytania pojedynczego pliku najlepiej skorzystać z wbudowanego narzędzia, dostępnego w zakładce *Environment -> Import Dataset -> From Text (base)...* (konkretne wyrażenia mogą różnić się w zależności od wersji Rstudio): Po uruchomieniu narzędzia wyświetli się okno pozwalające wybrać plik do wczytania. Po wybraniu pliku wyświetli się okno zawierające m.in. podgląd pliku, podgląd ramki danych jaką utworzy Rstudio, oraz dodatkowe opcje takie jak wybór separatora kolumn: Kliknięcie na przycisk *Import* utworzy nową zmienną typu *data frame* o nazwie `Zadluzenie.gmin`.
Nazwę zmiennej możemy zmienić w polu *Name* w lewym górnym rogu okna importowania danych.
Podgląd tabeli wyświetli się w nowej zakładce w oknie skryptów.
Z kolei w zakładce *History* obok zakładki *Environment* pojawi się kod, którym R wczytał dane.
Możemy łatwo przekopiować go do skryptu lub notatnika markdown, zaznaczając odpowiednią linijkę w zakładce *History* i klikając przycisk *To Source*.
Sposób 2. W przypadku większej liczby zbiorów danych używanie wbudowanego narzędzia jest niewydajne. Dużo lepszym pomysłem jest wówczas skorzystanie z jednej z funkcji służących do wczytywania danych: `read.table`, `read.delim` lub `read.csv`. Pierwsza służy do wczytywania dowolnych tabel, druga - tabel, w których separatorem jest tabulator, trzecia - tabel, w których separatorem jest przecinek. Różne są też opcje domyślne: `read.csv` oraz `read.delim` domyślnie zakładają, że w danych znajduje się nagłówek zawierający nazwy kolumn, a `read.table` nie.

W dalszej części zajęć możecie sami wybrać, z którego sposobu wczytywania danych będziecie korzystać.

Wizualizacja i analiza danych

Zadanie 1. Wczytaj dane z pliku `Zadluzenie gmin.csv` i sprawdź, co zawierają poszczególne kolumny. Za pomocą funkcji `summary` sprawdź podstawowe statystyki tych danych. Jaka jest mediana zadłużenia? Jaka jest najwyższa wartość zadłużenia wśród 75% najmniej zadłużonych gmin? A jakie jest najwyższe zadłużenie?

```
Zadluzenie.gmin <- read.csv("~/Files/Math/SAD21_22/Lab3/Zadluzenie gmin.csv",sep = "\t",fileEncoding =
summary(Zadluzenie.gmin$Zadłużenie.gmin)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   15.18   26.09   27.69   37.80  405.66
```

Oblicz średnią i odchylenie standardowe zadłużenia. Na tej podstawie odpowiedz, w przybliżeniu, o ile punktów procentowych zadłużenie średnio rzecz biorąc odchyla się od średniej?

```
mean(Zadluzenie.gmin$Zadłużenie.gmin)
```

```
## [1] 27.68867
```

```
sd(Zadluzenie.gmin$Zadłużenie.gmin)
```

```
## [1] 19.1929
```

```
ggplot(Zadluzenie.gmin,aes(x=Zadłużenie.gmin))+geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

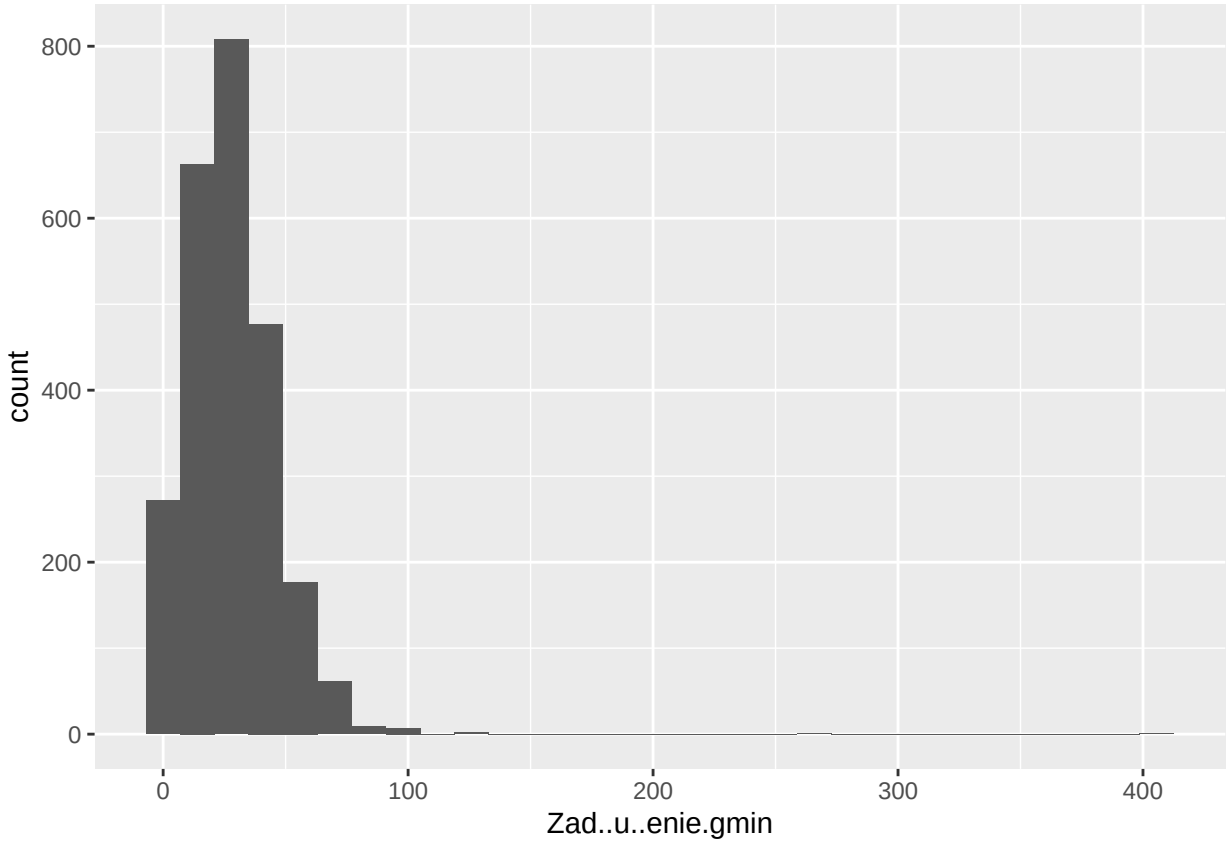
```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
```

```
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <82>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <c5>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Zadłużenie.gmin' in 'mbcsToSbcs': dot substituted for
## <bc>
```



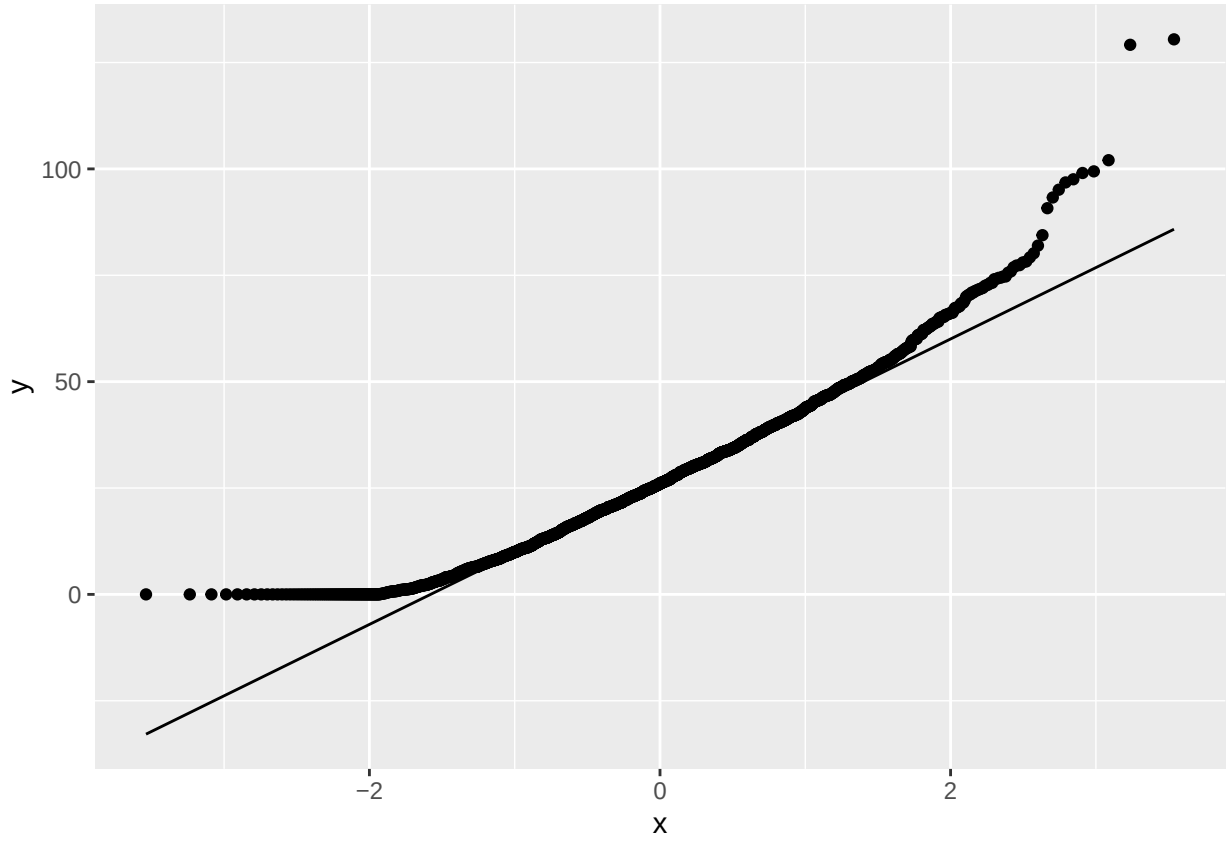
Zadanie 2. Po podjęciu decyzji o odrzuceniu (lub nie) ewentualnych obserwacji odstających, zwizualizuj zadłużenie ponownie na histogramie. Za pomocą wykresu kwantylowego oceń, czy i jak rozkład zadłużenia odbiega od rozkładu normalnego. Wykres kwantylowy porównujący dane ze standardowym rozkładem normalnym utworzysz korzystając z warstwy `stat_qq` z pakietu `ggplot2`.
Ta warstwa przyjmuje jedną estetykę o nazwie `sample`, równą nazwie kolumny z której ma utworzyć QQplot.

Do wykresu kwantylowego warto dodać prostą obrazującą ogólny trend w wykresie.
Można w tym celu wykorzystać warstwę `stat_qq_line` o tych samych estetykach.
Ta warstwa przeprowadzi prostą przez punkty odpowiadające kwantylom o wartości 0.25 i 0.75.
Więcej o wykresach kwantylowych i jak je interpretować przeczytasz tutaj.

```
Zadluzenie.gmin[order(Zadluzenie.gmin$Zadłużenie.gmin,decreasing=TRUE)[1:3],]
```

```
##         Region Kod.Teryt    Kod.BDR Zadłużenie.gmin Jednostka Okres
## 2478 Ostrowice   3203042 4326303042        405.6593         %  2015
## 2477     Rewal   3205072 4326405072        264.7442         %  2015
## 2476   Bielice   3212012 4326412012        130.4456         %  2015
```

```
filtred<-Zadluzenie.gmin[order(Zadluzenie.gmin$Zadłużenie.gmin,decreasing=TRUE)[3:nrow(Zadluzenie.gmin)]
ggplot(filtred)+stat_qq(aes(sample=Zadłużenie.gmin))+stat_qq_line(aes(sample=Zadłużenie.gmin))
```



Przedziały ufności

Zazwyczaj dane reprezentują podzbiór jakiejś populacji. Przedział ufności pozwala nam ocenić jakie wartości może przyjmować badany parametr w rzeczywistości, jeśli estymujemy go na podstawie danych.
W praktyce najczęściej oblicza się przedziały ufności dla średniej lub wariancji, przyjmując założenie, że dane pochdzą z rozkładu normalnego.

Zadanie 3. Wczytaj dane `iris` i wybierz wiersze odpowiadające gatunkowi *versicolor*. Korzystając z wykresu kwantylowego sprawdź, czy zmienna `Sepal.Width` (mierząca szerokość działki kielicha) ma rozkład normalny.

```
# Rozwiązanie
```

Przy założeniu, że szerokość działki kielicha ma rozkład normalny, oblicz przedział ufności dla średniej wartości tej zmiennej na poziomie $\alpha = 0.95$. W tym celu wykorzystaj wzór na tzw. *studentyzowany przedział ufności*:

$$\left(\bar{X} - \frac{t(1-\alpha/2, n-1)}{\sqrt{n}}\hat{S}, \bar{X} + \frac{t(1-\alpha/2, n-1)}{\sqrt{n}}\hat{S}\right),$$

gdzie $\bar{X}$ to średnia, $\hat{S}$ to pierwiastek z *nieobciążonego* estymatora wariancji (czyli wynik funkcji `sd()`), a $t(1-\alpha/2, n-1)$ to kwantyl na poziomie $1-\alpha/2$ dla rozkładu t Studenta o $n-1$ stopniach swobody (możesz go obliczyć korzystając z funkcji `qt()`).

Co możemy powiedzieć o średniej szerokości działki kielicha gatunku *Iris versicolor* na podstawie naszych danych, przy założeniu, że zmienna ta ma rozkład normalny? Co się stanie, jeśli okaże się, że to założenie nie jest spełnione?

```
#Rozwiązanie
```

Zadanie 4. Porównaj wyniki z poprzedniego zadania z tzw. *asymptotycznym przedziałem ufności*, danym wzorem

$$\left(\bar{X} - \frac{q(1-\alpha/2)}{\sqrt{n}}\hat{S}, \bar{X} + \frac{q(1-\alpha/2)}{\sqrt{n}}\hat{S}\right),$$

gdzie $q(1-\alpha/2)$ jest kwantylem na poziomie $1-\alpha/2$ ze standardowego rozkładu normalnego.
Matematycznie, asymptotyczny przedział ufności otrzymujemy, biorąc wzór na przedział dla znanego odchylenia standardowego, $(\bar{X} - \sigma q(1-\alpha/2)/\sqrt{n},\ \bar{X} + \sigma q(1-\alpha/2)/\sqrt{n})$, i podstawiając estymator $\hat{S}$ za ochylenie standardowe σ.
Daje to mniej dokładne wyniki, ponieważ nie uwzględniamy w tym wzorze zmienności estymatora $\hat{S}$. Innymi słowy, stosując ten wzór zakładamy, że estymator $\hat{S}$ dał nam odchylenie standardowe z nieskończoną dokładnością.
Co z tego wynika jeśli chodzi o szerokość przedziału ufności? Co możemy zaobserwować, jeśli wykorzystamy mniej danych, np. 10 pomiarów *Iris versicolor*? A co się stanie, jeśli weźmiemy bardzo dużo danych? Dlaczego ten przedział nazywa się *asymptotycznym* przedziałem ufności?

```
#Rozwiązanie
```

Z rozkładem t Studenta spotkamy się ponownie na następnych zajęciach, na których zajmiemy się testowaniem hipotez statystycznych.
Ten rozkład najczęściej pojawia się wtedy, gdy nie znamy odchylenia standardowego w naszej próbie i wykorzystujemy estymator średniej dzielony przez estymator odchylenia standardowego. Mowi się wówczas o statystyce t, *t-value* lub *t-score*:

$$T = \frac{\bar{X} - \mu}{\hat{S}}\sqrt{n}.$$

gdzie $\hat{S}$ to pierwiastek z *nieobciążonego* estymatora wariancji. Taka statystyka ma rozkład t Studenta z $n-1$ stopniami swobody. Stopnie swobody należy (na razie) rozumieć po prostu jako parametr tego rozkładu - im więcej stopni swobody, tym bardziej przypomina od standardowy rozkład normalny, a im mniej, tym bardziej jego wartości są odległe od zera.

W przypadku gdy wykorzystujemy średnią dzieloną przez znane odchylenie standardowe mówi się często o teście z i statystyce z, *z-value* lub *z-score*:

$$Z = \frac{\bar{X} - \mu}{\sigma}\sqrt{n}$$

Czynnik $\sqrt{n}$ pojawia się tutaj, ponieważ dla próby statystycznej z rozkładu normalnego mamy $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. Wówczas (oczywiście jeśli μ oraz σ są poprawne) mamy $Z \sim \mathcal{N}(0,1)$.

Więcej na temat rozkładu t, porównanie go z rozkładem normalnym, i kilka wskazówek kiedy używać jednego lub drugiego można usłyszeć tutaj. Więcej na temat *z-score* można przeczytać tutaj, a na temat *t-score* tutaj.

Zadania dodatkowe.

Zadanie 1. Wylosuj 1000 prób po 10 obserwacji z rozkładu jednostajnego na przedziale $[0, a]$ dla wybranej wartości parametru a. Następnie:
1. Wykorzystaj te próbki do estymacji parametru a metodą największej wiarygodności: $\hat{a} = \max_i X_i$. Przedstaw otrzymane wartości $\hat{a}$ na histogramie.
2. Oblicz przedział ufności dla próby numer 1 korzystając ze wzoru

$$(\max(X_n), \frac{\max(X_n)}{\alpha^{1/n}})$$

.
3. Oblicz przedział ufności dla każdej próby i sprawdź, dla ilu prób przedział zawiera prawdziwą wartość parametru a.

```
# Rozwiązanie
```