Lab4: Testowanie hipotez

Michał Ciach, Anna Macioszek

Nadobnice alpejskie w Bieszczadach

Zadanie przykładowe.. Za pomocą testu dwumianowego zweryfikuj, czy na podstawie tych danych można stwierdzić, że *R. longicorn* preferencyjnie wybiera składy drewna.

Rozwiązanie. Test dwumianowy służy do zweryfikowania hipotezy, że prawdopodobieństwo wystąpienia jakiegoś zdarzenia jest równe p_0. Formalnie, rozpatrujemy zmienną losową Y równą 1 gdy dane zdarzenie wystąpiło i 0 w przeciwnym przypadku. Nasza hipoteza to

$$H_0 : \mathbb{P}(Y = 1) = p_0,$$

hipoteza alternatywna natomiast zależy od danego pytania badawczego. Zajmiemy się nią za chwilę.

Tak jak zawsze przy testowaniu hipotez, zakładamy, że hipoteza H_0 jest prawdziwa i przy tym założeniu sprawdzamy, czy możemy w danych zaobserwować coś o podejrzanie małym prawdopodobieństwie. Jeśli tak, to odrzucamy H_0 na korzyść H_1. W tym przypadku badamy liczbę wystąpień danego zdarzenia, którą oznaczymy jako X. Zmienna Y mówi nam, czy zdarzenie wystąpiło w jednym eksperymencie, zmienna X natomiast oznacza liczbę wystąpień zdarzenia w serii eksperymentów statystycznych (na przykład pomiarów). Jeśli przeprowadziliśmy N pomiarów, to, przy założeniu H_0, zmienna X ma rozkład dwumianowy $Bin(N, p_0)$, skąd pochodzi nazwa testu.

W naszym przypadku będziemy badać, czy na podstawie zebranych danych możemy stwierdzić, że *R. longicorn* wybiera składy drewna z prawdopodobieństwem większym niż $p_0 = 1/2$. Stąd, nasza hipoteza alternatywna to

$$H_1 : \mathbb{P}(Y = 1) > 1/2.$$

Jakie zdarzenia mogą świadczyć o tym, że prawdziwa jest raczej H_1 niż H_0?
Skoro H_1 mówi, że *R. longicorn* wybiera składy drewna *częściej* niż wybrane p_0, to sprawdzimy, czy zaobserwowana liczba wystąpień w składach drewna, czyli X, jest nieprawdopodobnie wysoka.

Nasza liczba przeprowadzonych eksperymentów (pomiarów) to $N = 110$, ponieważ tyle historycznych wystąpień zbadał zespół ekologów. Stąd, przy założeniu H_0, liczba wystąpień w składach drewna powinna mieć rozkład $Bin(110, 0.5)$.
Żeby przeprowadzić test dwumianowy musimy jeszcze otrzymać zaobserwowaną wartość zmiennej losowej X. Skoro *R. longicorn* wystąpiła w składach drewna w 66,7% przypadków, to obliczamy $X = 0.667 * 110 = 73.37$, co zaokrąglamy do 73. Tworzymy zmienne w R które przechowają nam te wartości:

```
N = 110
X = as.integer(110*0.667)
```

Dygresja dotycząca R. W powyższym kawałku kodu wykorzystałem znak równości `=` zamiast strzałki w lewo `<-`.
Obie metody są poprawne i w tym przypadku nie ma między nimi żadnej różnicy.
Różnica występuje natomiast przy wywoływaniu funkcji. Komenda `ggplot(data=X)` może zadziałać inaczej, niż `ggplot(data <- X)`, i w 99% przypadków należy stosować tę pierwszą.
Są to szczegóły techniczne i na tym kursie nie musicie się nimi przejmować. Zainteresowani mogą dowiedzieć się więcej tutaj.

Sformułowaliśmy już nasze hipotezy badawcze i otrzymaliśmy wartość naszej *statystyki testowej* X.
Teraz należy sprawdzić, czy X przyjęła nietypowo wysoką wartość.
W tym celu obliczymy, jakie jest prawdopodobieństwo, że X wyniosłaby *co najmniej* 73 gdyby prawdziwa była H_0.
Takie prawdopodobieństwo nazywamy ***p-wartością***. Jeśli to prawdopodobieństwo będzie małe (na ogół przyjmuje się próg 0.05), to odrzucimy hipotezę zerową i uznamy, że składy drewna bukowego rzeczywiście stanowią pułapkę ekologiczną dla *R. longicorn*.

Bardzo ważne jest, aby rozpatrzeć *co najmniej* tak nietypowe wartości, jak te zaobserwowane. To dlatego, że zaobserwowanie *dokładnie* danej, konkretnej liczby na ogół samo w sobie jest bardzo małe.
Z tego powodu obliczenie $\mathbb{P}(X = 73)$ niewiele by nam powiedziało o tym, czy X jest nietypowo wysoka.

Do obliczenia p-wartości $\mathbb{P}(X \geq 73)$ wykorzystamy funkcję `pbinom`.
Dokumentację, jak zawsze, możecie przeczytać wpisując w konsolę komendę `?pbinom`. Tutaj jednak czai się kolejna pułapka: funkcja `pbinom(73)` zwróci nam wartość $\mathbb{P}(X \leq 73)$. Nas natomiast interesuje $\mathbb{P}(X \geq 73) = 1 - \mathbb{P}(X < 73) = 1 - \mathbb{P}(X \leq 72)$.
Z tego powodu w poniższym kawałku kodu musimy odjąć 1 od wartości zmiennej X.

```
p.value = 1 - pbinom(X - 1, N, 0.5) # = P(X >= successes)
p.value
```

```
## [1] 0.0003844925
```

Widzimy, że prawdopodobieństwo zaobserwowania co najmniej 73 wystąpień w składach drewna jest szalenie małe.
Gdyby H_0 była prawdziwa, to takie zjawisko wystąpiłoby raz na $1/0.00038 \approx 2600$ powtórzeń takiego eksperymentu.
Na tej podstawie stwierdzamy, że *R. longicorn* preferencyjnie osiedla się w składach drewna.

Całą procedurę możemy dodatkowo zilustrować na wykresie, na którym zaznaczymy typowe i nietypowe wartości zmiennej X.
W pierwszym kroku należy utworzyć ramkę danych z której stworzymy wykres.
Po pierwsze, potrzebujemy wektora wartości jakie może przyjąć zmiena X, czyli wektora `1:N`. Następnie, obliczymy prawdopodobieństwo przyjęcia każdej z tych wartości komendą `dbinom(1:N, N, 0.5)`.
Na końcu policzymy wektor p-wartości odpowiadających każdej z potencjalnych wartości zmiennej X, `1 - pbinom(1:N - 1, N, 0.5)`, i stworzymy wektor logiczny mówiący nam, czy p-wartość jest większa niż 0.05, komendą `1 - pbinom(1:N - 1, N, 0.5) >= 0.05`.
Otrzymane wektory ustawimy w ramkę danych:

```
data.to.plot <- data.frame('X' = 1:N, 'dbinom' = dbinom(1:N, N, 0.5), 'Typical' = 1 - pbinom(1:N - 1, N
```

Otrzymaną ramkę danych możemy wykorzystać do stworzenia następującego wykresu (po załadowaniu biblioteki `ggplot2`):

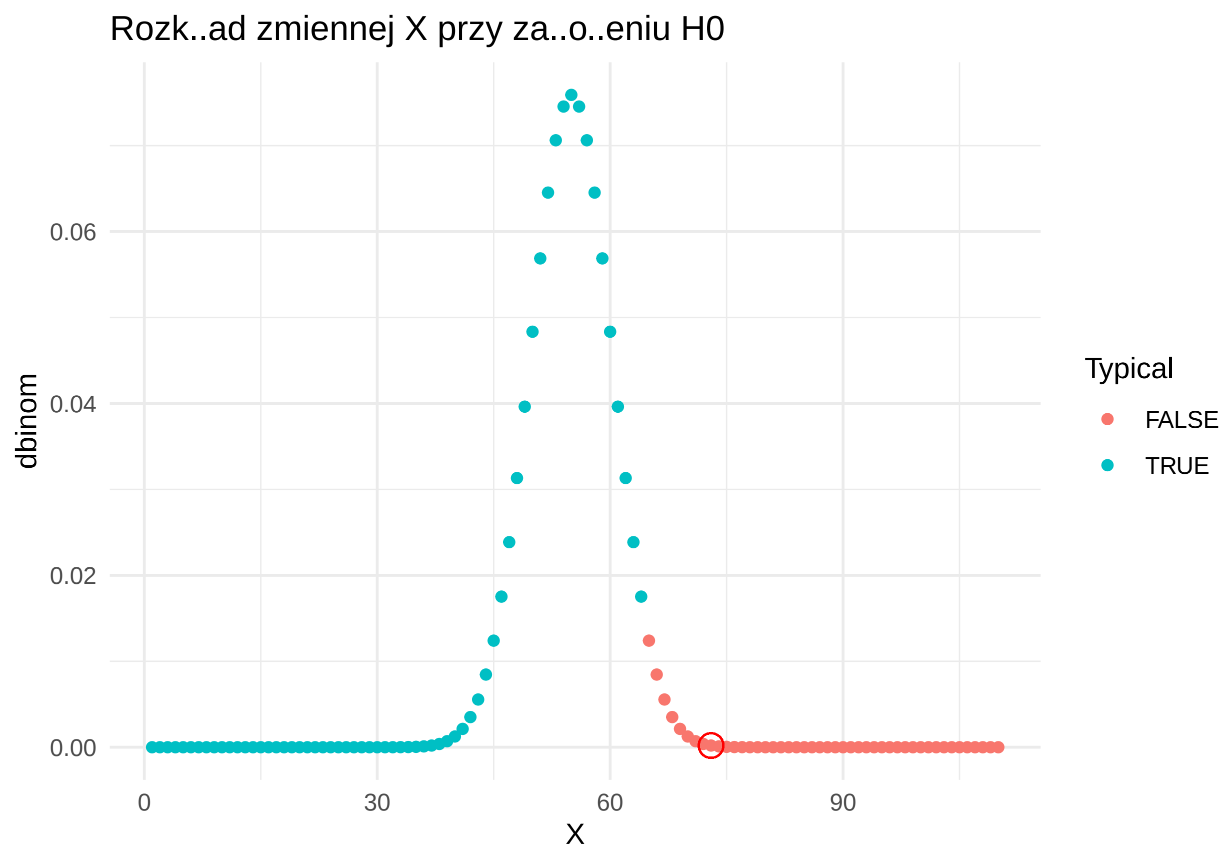
```
ggplot(data.to.plot) + geom_point(aes(x=X, y=dbinom, col=Typical)) + ggtitle('Rozkład zmiennej X przy za
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
```

```
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

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

```
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <82>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <c5>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'Rozkład zmiennej X przy założeniu H0' in 'mbcsToSbcs':
## dot substituted for <bc>
```



Powyższy wykres przedstawia prawdopodobieństwo zaobserwowania poszczególnych wartości X przy założeniu H_0.
Na przykład, widzimy, że najbardziej prawdopodobne są wartości bliskie 50.
Na czerwono oznaczyłem tzw. *obszar krytyczny*. Są to te wartości zmiennej X, które odpowiadają p-wartości mniejszej niż 0.05.
Na niebiesko zaznaczone są wartości 'typowe'. Zwróć uwagę, że to, jaka wartość jest typowa a jaka nie, zależy m.in. od wybory hipotezy alternatywnej H_1. Dodatkowo za pomocą czerwonego okręgu zaznaczona jest zaobserwowana wartość zmiennej X.

Zadanie 1. Dane zebrane przez zespół wskazują, że przed rokiem 2000 *R. longicorn* zamieszkujące składy drewna stanowiły 0.40 wszystkich wystąpień tego owada. Po 2000 liczba ta wzrosła do 0.76.
Za pomocą testu niezależności chi-kwadrat zweryfikuj, czy wybór siedliska zależy od okresu.
Oblicz p-wartość korzystając z funkcji `pchisq`, a następnie zweryfikuj swoje wyniki za pomocą `chisq.test`.

Test niezależności chi-kwadrat był omówiony na Wykładzie 3. Więcej informacji na jego temat możesz znaleźć również tutaj.

Wskazówka. Aby przeprowadzić test niezależności chi-kwadrat, musimy najpierw obliczyć *tabelę kontyngencji*, podsumowującą liczby wystąpień *R. longicorn* w różnych siedliskach w zależności od okresu:

Liczba wystąpień	Przed 2000 r.	Po 2000 r.
Skład drewna	O_{11}	O_{12}
Siedlisko naturalne	O_{21}	O_{22}

Z poprzedniego zadania wiemy, że zbadano $N = 110$ wystąpień owada, a 66.7% spośród nich było w składzie drewna. Żeby uzupełnić tabelę kontyngencji musimy jeszcze obliczyć procent owadów znalezionych przed rokiem 2000.

Oznaczmy przez A zdarzenie polegające na tym, że losowo wybrany owad zamieszkuje skład drewna, a przez B zdarzenie, że losowo wybrana obserwacja jest sprzed roku 2000. Zdarzenie przeciwne oznaczmy przez B'. Korzystając ze *wzoru na prawdopodobieństwo całkowite* możemy teraz napisać

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B')\mathbb{P}(B') = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B')(1 - \mathbb{P}(B)).$$

Zwróćmy uwagę, że z treści zadania znamy zarówno $\mathbb{P}(A|B)$, jak i $\mathbb{P}(A|B')$ – jest to odpowiednio 0.40 i 0.76. Mamy zatem

$$0.667 = 0.40\mathbb{P}(B) + 0.76(1 - \mathbb{P}(B)),$$

a po przekształceniu tej równości otrzymujemy $\mathbb{P}(B) = (0.667 - 0.76)/(0.4 - 0.76) \approx 0.26$.

W tabeli kontyngencji oznaczyliśmy O_{11} jako liczbę owadów znalezionych w składzie drewna przed rokiem 2000.
Obliczymy tę wartość jako $N\mathbb{P}(A \wedge B)$. Prawdopodobieństwo zajścia obu zdarzeń naraz, $\mathbb{P}(A \wedge B)$, obliczymy korzystając z *reguły łańcuchowej* jako

$$\mathbb{P}(A \wedge B) = \mathbb{P}(A|B)\mathbb{P}(B) = 0.40 \cdot 0.26 = 0.1033.$$

Wobec tego na podstawie danych z zadania mamy $O_{11} = 110 * 0.1033 = 11.36$, co zaokrąglamy do 11.

Pozostałe elementy macierzy kontyngencji należy obliczyć analogicznie. Statystyka testowa i jej rozkład pod warunkiem H_0 podane są na slajdach z wykładu oraz na podlinkowanej wcześniej stronie. Aby wykorzystać funkcję `chisq.test`, należy utworzyć tabelę kontyngencji korzystając z funkcji `matrix` omówionej na poprzednich zajęciach. **Koniec wskazówki.**

```
N<-110
pt<-0.667
pakb<-0.4
pakbp<-0.76
pb<-(pt-pakbp)/(pakb-pakbp)
mat<-sapply(c(N*pakb*pb,N*pakbp*(1-pb),N*pb*(1-pakb),N*(1-pb)*(1-pakbp)),as.integer)
mat<-mat+c(0,0,0,1)
mat<-matrix(mat,nrow = 2,ncol=2,byrow=TRUE)
print(mat)
```

```
##      [,1] [,2]
## [1,]   11   62
## [2,]   17   20
```

```
chisq.test(mat,correct=F)
```

```
##
##  Pearson's Chi-squared test
##
## data:  mat
## X-squared = 12.338, df = 1, p-value = 0.0004439
```

Uwaga, opcja `correct=T` zwróci inny wynik niż obliczony ręcznie.

Zadłużenie gmin c.d.

W następnym zadaniu wykorzystamy test t Studenta aby porównać średnie zadłużenia gmin z wybranych województw.

Zadanie 2. *Przygotowanie danych.*
Zaczniemy od wczytania danych znajdujące się w pliku `Zadluzenie gmin.csv`, dostępnym na stronie przedmiotu.

Ogólny sposób wczytywania danych został opisany na poprzednich zajęciach.
W tym zadaniu jednak czai się na nas kolejna pułapka.
Druga kolumna danych zawiera *kod terytorialny* gminy, którego pierwsze dwie cyfty to identyfikator województwa. Część kodów terytorialnych zaczyna się od zera. R domyślnie uzna, że ta kolumna zawiera liczby, więc te zera usunie.
Żeby je zachować, musimy ręcznie podać typy danych w kolejnych kolumnach za pomocą argumentu `colClasses`:

```
Zadluzenie.gmin <- read.delim("Zadluzenie gmin.csv", colClasses = c('factor', 'factor', 'factor', 'nume
```

Ponieważ kazaliśmy R-owi interpretować drugą kolumnę jako factor, to zachowa on początkowe zera z kodów terytorialnych. W następnym kroku dodamy do naszych danych kolumnę zawierającą nazwy województw w których znajdują się gminy. Na początku musimy wybrać dwa pierwsze znaki z każdego identyfikatora.
Do obsługi napisów w R najlepiej nadaje się biblioteka `stringr`. Zainstaluj ją komendą `install.packages("stringr")`, a następnie załaduj i wykorzystaj funkcję `str_sub()`, aby otrzymać wektor zawierający po dwa pierwsze znaki z każdego identyfikatora terytorialnego. Dołącz ten wektor do danych `Zadluzenie.gmin` jako nową kolumnę. Sposób dołączania kolumn do ramki danych był omówiony na pierwszych zajęciach.

Następnym krokiem będzie przetłumaczenie otrzymanego wektora na nazwy województw.
Poniższa komórka z kodem utworzy wektor, którego pola nazwane są identyfikatorem województw, a zawierają ich nazwy.

```
slownik <- c('02' = 'Dolnośląskie', '04' = 'Kujawsko-pomorskie',
             '06' = 'Lubelskie', '08' = 'Lubuskie',
             '10' = 'Łódzkie', '12' = 'Małopolskie',
             '14' = 'Mazowieckie', '16' = 'Opolskie',
             '18' = 'Podkarpackie', '20' = 'Podlaskie',
             '22' = 'Pomorskie', '24' = 'Śląskie',
             '26' = 'Świętokrzyskie', '28' = 'Warmińsko-mazurskie',
             '30' = 'Wielkopolskie',  '32' = 'Zachodniopomorskie')
```

Jedną z opcji indeksowania wektorów w R jest indeksowanie po nazwach pól.
Dzięki temu stworzony powyżej wektor umożliwia nam bardzo proste przetłumaczenie pierwszych dwóch znaków kodu terytorialnego na nazwę województwa.
Wystarczy napisać `slownik[c('02', '02', '04')]`, aby automatycznie przetłumaczyć wektor `c('02', '02', '04')` na wektor `c("Dolnośląskie", "Dolnośląskie", "Kujawsko-pomorskie")`.

Wykorzystaj zmienną `slownik`, aby utworzyć w danych `Zadluzenie.gmin` nową kolumnę zawierającą nazwy województw.

```
#library(stringr)
Zadluzenie.gmin$Wojewodztwo<-slownik[sapply(Zadluzenie.gmin$Kod.Teryt,substr,start=0,stop=2)]
```

Tak jak na poprzednich zajęciach, jeśli uważasz, że w danych występują obserwacje odstające, to usuń je przed dalszą analizą.

Zadanie 3. Czy na podstawie danych możemy stwierdzić, że średnie zadłużenie gminy w województwie mazowieckim jest mniejsze niż 25%? Wykorzystaj jednopróbkowy test t Studenta przy hipotezach $H_0 : \mu = 25$, $H_1 : \mu < 25$. Oblicz p-wartość samodzielnie, a następnie porównaj swój wynik z funkcją `t.test`. Dystrybuantę rozkładu t Studenta możesz obliczyć korzystając z funkcji `pt`. Pamiętaj, że funkcja `var` wykorzystuje nieobciążony estymator wariancji.

Wskazówka. Mając nazwy województw w kolumnie `Wojewodztwo`, dane dotyczące województwa mazowieckiego możesz wybrać komendą `Zadluzenie.gmin[Zadluzenie.gmin$Wojewodztwo == 'Mazowieckie', ]`.

```
data<-Zadluzenie.gmin[Zadluzenie.gmin$Wojewodztwo=="Mazowieckie",]
n<-length(data$Zadłużenie.gmin)
sn<-var(data$Zadłużenie.gmin)*(n-1)/n
```

```
M<-mean(data$Zadłużenie.gmin)
X<-sqrt(n)*(M-25)/sqrt(sn)
myp<-pt(X,n-1)
print(c(X,myp))
```

```
## [1] -0.6834946  0.2474000
```

Ponieważ p-wartość jest duża, to nie mamy podstaw żeby uznać że średnie zadłużenie jest mniejsze niż 25%. *Co to oznacza?* W tym zadaniu założyliśmy, że zadłużenie gminy to zmienna losowa o pewnym rozkładzie prawdopodobieństwa, a w danych obserwujemy realizacje tej zmiennej losowej. Po pewnym czasie możemy obliczyć zadłużenie gmin ponownie. Jeśli nasze "warunki eksperymentalne" się nie zmienią, to otrzymamy nowy zestaw realizacji zmiennych losowych z tego samego rozkładu.
Wynik testu wskazuje, że rozrzut zadłużenia gmin jest na tyle duży, że po wykonaniu pomiaru ponownie średnie zadłużenie może okazać się wyższe lub równe 25%.

Zwróć uwagę, że jest to nieco inna interpretacja niż w przypadku badanych wcześniej chrząszczy. Tam mieliśmy do czynienia z próbą wybraną z pewnej populacji, i zastanawialiśmy się, czy te losowo wybrane chrząszcze pozwalają nam wyciągnąć wnioski na temat całej populacji. To, że wśród 110 zbadanych chrząszczy 66,7% mieszkało w składzie drewna nie musi oznaczać, że dokładnie 66,7% wszystkich chrząszczy z Bieszczad mieszka w składach drewna. Możemy jednak z dużą dozą pewności uznać, że w takich składach żyje więcej niż połowa wszystkich osobników *R. longicorn*.

W tym zadaniu mamy do dyspozycji całą "populację" gmin, więc średnie zadłużenie gmin na Mazowszu w roku 2015 wyniosło po prostu 24,38% i kropka.
To, co tutaj badamy, to własności ukrytego procesu losowego który wygenerował nam nasze dane. Interesuje nas na przykład to, na ile to zadłużenie może się zmieniać w czasie.
Takie podejście jest przydatne na przykład w sytuacji, gdy chcemy przekazać środki finansowe z województw o małym zadłużeniu do województw o wysokim zadłużeniu. Chcielibyśmy wówczas być pewni, że wybrane "bogate" województwa rzeczywiście mają wystarczająco małe zadłużenie. Przeprowadzenie testów i skonstruowanie przedziałów ufności w takiej sytuacji pozwala nam w większym stopniu kontrolować naturalną zmienność występującą w naszych danych i zmniejszyć ryzyko związane z podejmowaniem decyzji.

Choć formalizm matematyczny w obu przypadkach jest identyczny, to warto zdawać sobie sprawę z innego znaczenia wykorzystywanych wzorów.

Zadanie 4. Wybierz dane dotyczące województw łódzkiego i pomorskiego. Przy założeniu, że rozkład zadłużenia jest normalny, przetestuj hipotezę, że wariancja zadłużenia w każdej z tych gmin jest równa $\sigma_0^2 = 226$. Wzór na statystykę testową znajdziesz na slajdach do Wykładu 2. Hipoteza alternatywna to $H_1 : \sigma^2 \neq 15$.

W tym teście obliczanie p-wartości jest nieco bardziej skomplikowane. Mamy do czynienia z alternatywą dwustronną, i nie wiemy, czy statystyka testowa będzie przyjmować wartości nietypowo niskie, czy nietypowo wysokie. Jednym ze sposobów, w jaki można sobie z tym poradzić, to obliczenie p-wartości jako

$$p = 2 \min\{\mathbb{P}(X < x),\ \mathbb{P}(X > x)\},$$

gdzie X to statystyka testowa, x to jej zaobserwowana wartość, a prawdopodobieństwo obliczamy przy założeniu hipotezy zerowej.
Taki sposób obliczania p-wartości wykorzystuje na przykład funkcja `varTest` z pakietu `EnvStats`.
Innym sposobem jest ustalenie z góry *poziomu istotności*, czyli progu p-wartości poniżej którego odrzucamy H_0. Możemy wówczas wykorzystać wzory na obszar krytyczny podane na slajdach do Wykładu 2.

Wynikiem powinny być bardzo wysokie p-wartości, wskazujące na to, że możemy przyjąć że wariancja w obu województwach jest równa 226. Możemy to interpretować tak, jak w poprzednim zadaniu – jako wynik dotyczący naturalnej zmienności naszych danych. Inna interpretacja jest taka, że przy takiej zmienności zadłużenia jaką obserwujemy w danych, przyjęcie, że wariancja jest równa 226, zapewni nam dobre przybliżenie.

Powyższy wniosek uzasadnia dodatkowo wykorzystanie w następnym zadaniu testu t Studenta dla prób o równych wariancjach.

Zadanie 5. Czy na podstawie danych możemy stwierdzić, że przeciętna gmina z Pomorskiego jest zadłużona bardziej niż z Łódzkiego?
Wykorzystaj test t Studenta dla populacji o różnych licznościach, ale równych wariancjach (Wykład 2).
Porównaj swoje wyniki z otrzymanymi za pomocą funkcji `t.test`; Zwróć uwagę na parametr `var.equal`.

W tym przypadku p-wartość możemy obliczyć łatwiej, niż w poprzednim zadaniu. Ponieważ interesują nas zarówno nietypowo wysokie jak i niskie wartości statystyki testowej, a rozkład tej ostatniej jest symetryczny względem zera, robimy to następująco:

$$p = \mathbb{P}(|X| > |x|) = \mathbb{P}(X > |x|) + \mathbb{P}(X < -|x|) = 2\mathbb{P}(X < -|x|).$$

Zadania dodatkowe (nieobowiązkowe).

Zadanie 1. Wybierz dwa województwa. Zweryfiuj hipotezę o tym, że średnie zadłużenie gmin z tych województw jest równe, za pomocą dwupróbkowego testu t Studenta dla prób o różnych rozmiarach i wariancjach (tzw. test Welscha). Hipoteza alternatywna w tym przypadku jest taka, że średnie zadłużenia są różne (tzw. *hipoteza dwustronna*).
Oblicz p-wartość samodzielnie, korzystając ze wzoru na statystykę testową (mającą rozkład t Studenta) i jej liczbę stopni swobody podaną w podlinkowanym artykule. Porównaj ją z wynikiem funkcji `t.test`.

Zadanie 2. Liczba obserwacji w macierzy kontyngencji, którą obliczyliśmy w Zadaniu 1, jest raczej niska. To oznacza że test chi-kwadrat może nie być odpowiedni.
Wykonaj dokładny test Fishera i porównaj wyniki. Po wykonaniu własnego testu, zweryfikuj wyniki za pomocą `fisher.test`.