Wykład 7: Testowanie zgodności

SPRAWDZANIE NORMALNOŚCI ROZKŁADU

I. Metody graficzne sprawdzania normalności

1. Wykres skrzynkowy. Jeśli nie jest on symetryczny lub widać na nim dużo obserwacji odstających (dla rozkładu normalnego $\mathcal{N}(\mu, \sigma^2)$ średnio 7 obserwacji na 1000 znajduje się poza przedziałem $(Q_1 - 1.5IQR, Q_3 + 1.5IQR)$), to uznajemy, że rozkład cechy w próbie znacznie odbiega od rozkładu normalnego. W przeciwnym przypadku, dane mogą, ale nie muszą, pochodzić z rozkładu normalnego i trzeba to sprawdzać innymi metodami.

   ```
   > boxplot(dane)
   ```

   Niestety wykresy skrzynowe są wiarygodne jedynie dla dużej liczności prób (jeśli mamy mało danych, to nie warto na nie patrzeć podczas badania normalności rozkładu).

2. Histogram częstości i jądrowy estymator gęstości.

   ```
   > hist(dane,freq=F)
   > lines(density(dane))
   ```

   Pierwsza z powyższych komend rysuje w R histogram częstości (nie liczności); druga nanosi na poprzednio wykonany rysunek jądrowy estymator gęstości.

3. Wykres kwantylowy (wykres normalności) (Q-Q plot: quantile versus quantile plot).
   Niech $x_1, x_2, \ldots, x_n$ oznacza realizację próby losowej. Po jej uporządkowaniu (od obserwacji najmniejszej do największej) otrzymujemy tzw. *statystyki porządkowe* z próby, oznaczane $x_{1:n} \leqslant x_{2:n} \leqslant \ldots \leqslant x_{n:n}$. Wykres kwantylowy to zbiór punktów o współrzędnych $(u_{(i-0,5)/n}, x_{i:n})$, gdzie $i = 1, 2, \ldots, n$ zaś $u_{(i-0,5)/n}$ to kwantyl standardowego rozkładu normalnego rzędu $(i - 0, 5)/n$.

   Jeśli próba losowa pochodzi z rozkładu normalnego $\mathcal{N}(\mu, \sigma^2)$, to wykres kwantylowy jest zbiorem punktów leżących mniej-więcej na prostej $y = \sigma x + \mu$.

   ```
   > qqnorm(dane)
   > qqline(dane)
   ```

   Pierwsza z powyższych komend rysuje w R wykres kwantylowy; druga nanosi na ten wykres linię przechodzącą przez kwartyle.

II. Testy normalności

Zajmiemy się jedynie podstawowymi uniwersalnymi testami normalności. Weryfikujemy w nich następujące hipotezy:
H_0 : rozkład, z którego pochodzi badana próba losowa, jest normalny,
H_1 : rozkład, z którego pochodzi badana próba losowa, nie jest normalny.
Poniżej wymieniamy podstawowe uniwersalne testy normalności, poczynając od uznawanego za najlepszy i przechodząc stopniowo do uznawanych za coraz słabsze.

1. Test Shapiro-Wilka

   Zaproponowany w 1965 r. jest to dziś uznawany za najlepszy test uniwersalny normalności rozkładu. Konstrukcja tego testu opiera się na wykresie kwantylowym. Dokładniej, wyznacza się linię, która jest możliwie najlepiej dopasowana do punktów tego wykresu (mówiąc precyzyjniej, wyznacza się tzw. *prostą regresji*) i następnie bada się stopień dopasowania tych punktów do owej prostej.

   ```
   > shapiro.test(dane)
   ```

2. Test Andersona-Darlinga

```
> install.packages("nortest")
> library(nortest)
> ad.test(dane)
```

3. Test Craméra-von Misesa

```
> install.packages("nortest")
> library(nortest)
> cvm.test(dane)
```

W porównaniu z testem Craméra-von Misesa, test Andersona-Darlinga zwraca większą uwagę na ogony rozkładu.

4. Test Lilliefors'a

```
> install.packages("nortest")
> library(nortest)
> lillie.test(dane)
```

Test Lilliefors'a sprawuje się średnio gorzej niż test Andersona-Darlinga i test Craméra-von Misesa.

TESTOWANIE ZGODNOŚCI Z DOWOLNYM ROZKŁADEM

1. Test Kołmogorowa-Smirnowa

Test Kołmogorowa-Smirnowa jest przeznaczony do sprawdzania zgodności rozkładu, z którego pochodzi próba losowa, z dowolnym rozkładem ciągłym.

H_0 : badana próba losowa pochodzi z zadanego (ciągłego) rozkładu (lub rodziny takich rozkładów)
H_1 : badana próba losowa nie pochodzi z zadanego rozkładu (lub rodziny rozkładów)

```
>ks.test(x=dane,y="nazwa.dystrybuanty.rozkladu",liczby.opisujace.parametry.rozkladu)
```

Na przykład chcąc sprawdzić czy dane pochodzą z rozkładu wykładniczego o parametrze $\lambda = 2$ napiszemy:

```
> ks.test(x=dane,y="pexp",rate=2)
```

Niestety `ks.test` obsługuje test Kołmogorowa-Smirnowa jedynie dla prostej H_0. W przypadku złożonej H_0 (np. H_0: rozkład badanej cechy jest wykładniczy z nieznanym parametrem λ) pozostaje nam samemu wyznaczyć przybliżoną wartość *p-value* metodą symulacji.

2. Test zgodności χ^2-Pearsona

H_0 : badana próba losowa pochodzi z zadanego rozkładu (lub rodziny rozkładów)
H_1 : badana próba losowa nie pochodzi z zadanego rozkładu (lub rodziny rozkładów)

Statystyka testowa $\chi^2 = \sum_{j=1}^{k} \frac{(n_j - n\,p_j^0)^2}{n\,p_j^0}$, gdzie k -ilość klas; p_j^0 - prawdopodobieństwa teoretyczne wpadnięcia obserwacji do j-tej klasy przy założeniu prawdziwości H_0 (jeśli H_0 nie jest hipotezą prostą, to brakujące parametry rozkładu z H_0 wyznaczamy metodą największej wiarygodności), n_j - liczba obserwacji, które znalazły się w j-tej klasie, n - liczność próby.

Zbiór krytyczny $W = \left\langle \chi^2_{1-\alpha,k-1-r}; +\infty \right)$, gdzie r jest ilością parametrów szacowanych z próby, zaś $\chi^2_{1-\alpha,k-1-r}$ to kwantyl rzędu $1-\alpha$ rozkładu chi-kwadrat o $k-1-r$ stopniach swobody.

Jeżeli wyznaczona wartść statystyki χ^2 należy do zbioru krytycznego W, to, na poziomie istotności α, H_0 odrzucamy.

UWAGI dotyczące testu zgodności χ^2-Pearsona:

- Statystyka testowa χ^2 testu zgodności χ^2-Pearsona w sytuacji, gdy hipoteza H_0 jest prawdziwa, ma jedynie w przybliżeniu rozkład χ^2 o $k-1-r$ stopniach swobody. Przybliżenie to uznajemy za dopuszczalne, gdy wszystkie $np_j^0 \geqslant 5$, a za dobre, gdy wszystkie $np_j^0 \geqslant 10$. Gdyby np_j^0 nie były duże, to wtedy przybliżenie rozkładem χ^2 nie działa i pozostaje symulacyjne wyznaczanie rozkładu statystyki testowej.
- Gdy testem zgodności χ^2-Pearsona sprawdzamy zgodność z rozkładem ciągłym, to podczas jego dyskretyzacji, końce przedziałów klas wybieramy tak, by prawdopodobieństwa klas p_j^0 były przynajmniej w przybliżeniu równe i by był spełniony warunek, że wszystkie $np_j^0 \geqslant 5$.

Test zgodności χ^2-Pearsona jest zaimplementowany w R, niestety jedynie dla prostych hipotez H_0:

```
> chisq.test(x, p, simulate.p.value=FALSE, B=2000)
```

gdzie

- `x` to wektor z licznościami poszczególnych klas,
- `p` to wektor z prawdopodobieństwami teoretycznymi p_j^0 poszczególnych klas,
- `simulate.p.value` musimy ustawić na `TRUE` jeśli chcemy symulacyjnie wyznaczyć wartość p-value testu, wtedy zostanie przeprowadzonych domyślnie `B=2000` losowań realizacji próbki losowej.

Przykład 7.1. 100 losowo wybranych studentów zapytano iloma językami obcymi biegle władają. Otrzymane wyniki zapisano w poniższej tabeli:

liczba języków	liczba studentów
0	25
1	40
2	30
3	5

Czy na podstawie powyższych danych można uznać, że rozkład liczby języków obcych, którymi biegle posługują się studenci, jest (a) rozkładem Poissona o średniej równej 1, (b) rozkładem Poissona? Przyjąć poziom istotności 0,01.

Rozwiązanie przykładu 7.1:

(a) H_0 : *badana próba losowa pochodzi z rozkładu Poissona o średniej równej 1,*
H_1 : *badana próba losowa nie pochodzi z rozkładu Poissona o średniej równej 1.*

liczba języków	n_j *- liczba studentów*	p_j^0	np_j^0	$n_j - np_j^0$	$\frac{(n_j - np_j^0)^2}{np_j^0}$
0	*25*	*0,37*	*37*	*-12*	*3,89*
1	*40*	*0,37*	*37*	*3*	*0,24*
2	*30*	*0,18*	*18*	*12*	*8*
3 lub więcej	*5*	*0,08*	*8*	*-3*	*1,12*

Sposób I: rachunki przeprowadzamy sami, korzystając z tego, że jeśli X ma rozkład Poissona z parametrem λ, to $P(X=x) = e^{-\lambda}\frac{\lambda^x}{x!}$, $x = 0, 1, 2, \ldots$ i $EX = \lambda$.

W hipotezie H_0 jest rozkład Poissona o średniej równej 1. Stąd $\lambda = EX = 1$ i

$$
\begin{aligned}
p_0^0 &= P(X=0) = e^{-\lambda}\tfrac{\lambda^0}{0!} = e^{-\lambda} = e^{-1} \approx 0,37,\\
p_1^0 &= P(X=1) = e^{-\lambda}\tfrac{\lambda^1}{1!} = e^{-\lambda}\lambda = e^{-1} \approx 0,37,\\
p_2^0 &= P(X=2) = e^{-\lambda}\tfrac{\lambda^2}{2!} = e^{-\lambda}\tfrac{\lambda^2}{2} = \tfrac{1}{2}e^{-1} \approx 0,18,\\
p_3^0 &= P(X \geqslant 3) = 1 - P(X<3) = 1 - (P(X=0) + P(X=1) + P(X=2)) = 1 - \tfrac{5}{2}e^{-1} \approx 0,08.
\end{aligned}
$$

Możemy użyć testu zgodności χ^2-Pearsona, bo wszystkie $np_j^0 \geqslant 5$ i tylko jedno np_j^0 nie spełnia warunku $np_j^0 \geqslant 10$.

Statystyka testowa tego testu to $\chi^2 = \sum_{j=1}^{4} \frac{(n_j - n\,p_j^0)^2}{n\,p_j^0} \approx 3,89 + 0,24 + 8 + 1,12 = 13,25.$

Pozostało jeszcze wyznaczyć zbiór krytyczny $W = \left\langle \chi^2_{1-\alpha,k-1-r}; +\infty \right)$, *gdzie* $\alpha = 0,01$ *(bo to poziom istotności),* $k = 4$ *(k to liczba klas) i* $r = 0$ *(bo żadnego parametru nie szacowaliśmy z próby). Zatem szukamy kwantyla*

$$\chi^2_{1-\alpha,k-1-r} = \chi^2_{1-0,01;4-1-0} = \chi^2_{0,99;3}.$$

Znajdziemy go korzystając z R:
> `qchisq(0.99,3)`
Otrzymujemy $\chi^2_{0,99;3} \approx 11,34$, *co daje* $W \approx \langle 11,34; +\infty)$.
Widzimy, że

$$\chi^2 \approx 13,25 \in W \approx \langle 11,34; +\infty) \,,$$

więc odrzucamy H_0 *i stwierdzamy, że rozkład liczby języków obcych, którymi posługują się studenci, nie jest rozkładem Poissona o średniej równej 1.*

Sposób II: liczymy korzystając z R. *Zacznijmy od wektora prawdopodobieństw* $(p_0^0, p_1^0, p_2^0, p_3^0)$. *Przypomnijmy, że jeśli* X *ma rozkład Poissona z parametrem* λ, *to*

- `dpois(x=k,lambda=`λ) *podaje prawdopodobieństwo* $P(X = k)$;
- `ppois(x=k,lambda=`λ`,lower.tail=FALSE`) *podaje prawdopodobieństwo* $P(X > k)$.

Zatem szukany wektor prawdopodobieństw $(p_0^0, p_1^0, p_2^0, p_3^0)$, *gdzie*

$$p_0^0 = P(X = 0), p_1^0 = P(X = 1), p_2^0 = P(X = 2), p_3^0 = P(X \geqslant 3) = P(X > 2) \text{ i } X \sim Poiss(\lambda = 1),$$

otrzymamy pisząc
> `prawdop0 <- c(dpois(c(0,1,2),lambda=1),ppois(2,lambda=1,lower.tail=FALSE))`
Potrzebujemy także wektor z zaobserwowanymi licznościami
> `licznosci <- c(25,40,30,5)`
Przeprowadzamy test zgodności χ^2*-Pearsona (przypominjmy, że możemy to zrobić, bo jak już wcześniej zauważyliśmy wszystkie* $np_j^0 \geqslant 5$ *i tylko jedno* np_j^0 *nie spełnia warunku* $np_j^0 \geqslant 10$*)*
> `chisq.test(x=licznosci, p=prawdop0)`
Odczytujemy p wartość:

$$p - value = 0,005787 < \alpha = 0,01 \;\Rightarrow\; \text{odrzucamy } H_0,$$

gdzie $\alpha = 0,01$ *to poziom istotności testu. Wyciągamy więc wniosek, że rozkład liczby języków obcych, którymi posługują się studenci, nie jest rozkładem Poissona o średniej równej 1.*

(b) H_0 : *badana próba losowa pochodzi z rozkładu Poissona,*
H_1 : *badana próba losowa nie pochodzi z rozkładu Poissona.*

Teraz H_0 *jest hipotezą złożoną, więc musimy zacząc od oszacowania parametru* λ. *Można pokazać, że estymatorem największej wiarygodności parametru* λ *rozkładu Poissona* $Poiss(\lambda)$ *jest*

$$\hat{\lambda}_{NW} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Średnią $\bar{x}$ *możemy policzyć samemu:*

$$\bar{x} = \tfrac{1}{100}(0 \cdot 25 + 1 \cdot 40 + 2 \cdot 30 + 3 \cdot 5) = \tfrac{115}{100} = 1,15$$

lub używając R:
> `dane <- c(rep(0,25),rep(1,40),rep(2,30),rep(3,5))`
> `mean(dane)`

Uzupełniamy poniższą tabelkę, postępując analogicznie jak w pkt. (a), tylko teraz zamiast $\lambda = 1$ *wstawiamy* $\lambda = 1,15$*. W szczególności wektor prawdopodobieństw* $(p_0^0, p_1^0, p_2^0, p_3^0)$ *możemy policzyć w* R *pisząc*

```
> prawdop0.b <- c(dpois(c(0,1,2),lambda=1.15),ppois(2,lambda=1.15,lower.tail=FALSE))
```

liczba języków	n_j - liczba studentów	p_j^0	np_j^0
0	25	0,317	31,7
1	40	0,364	36,4
2	30	0.209	20,9
3 lub więcej	5	0.110	11

Widzimy, że wszystkie $np_j^0 \geqslant 10$*, więc możemy użyć testu zgodności* χ^2*-Pearsona. Wyliczamy statystykę testową tego testu:*

```
> sum((licznosci-n*prawdop0.b)^2/(n*prawdop0.b))
```

otrzymując

$$\chi^2 = \sum_{j=1}^{4} \frac{(n_j - n\,p_j^0)^2}{n\,p_j^0} \approx 8.94.$$

Następnie wyznaczamy zbiór krytyczny $W = \left\langle \chi^2_{1-\alpha,k-1-r}; +\infty \right)$*, gdzie* $\alpha = 0,01$*,* $k = 4$ *i* $r = 1$ *(bo szacowaliśmy jeden parametr czyli* λ*). Zatem szukamy kwantyla*

$$\chi^2_{1-\alpha,k-1-r} = \chi^2_{1-0,01;4-1-1} = \chi^2_{0,99;2},$$

i znajdujemy go korzystając z R:

```
> qchisq(0.99,2)
```

Otrzymujemy $\chi^2_{0,99;2} \approx 9.21$*, co daje* $W \approx \langle 9.21; +\infty)$*.*
Widzimy, że

$$\chi^2 \approx 8.94 \notin W \approx \langle 9.21; +\infty) \,,$$

więc nie mamy podstaw do odrzucenia H_0 *i stwierdzamy, że rozkład liczby języków obcych, którymi posługują się studenci, jest rozkładem Poissona.*

BARDZO WAŻNA UWAGA: W punkcie (b) nie możemy bezpośrednio skorzystać z funkcji `chisq.test(x=licznosci, p=prawdop0.b)` odczytując p-value, bo ono odpowiada prostej hipotezie H_0: badana próba losowa pochodzi z rozkładu Poissona o średniej równej 1,15. Aby przetestować złożoną hipotezę H_0: badana próba losowa pochodzi z rozkładu Poissona, możemy jedynie wykorzystać policzoną w ten sposób wartość statystyki testowej `X-squared = 8.9394`, ale zbiór krytyczny musimy wyznaczyć sami.