Testy statystyczne w R
testy zgodności

Agnieszka Goroncy

Zakład Statystyki Matematycznej i Analizy Danych
Wydział Matematyki i Informatyki UMK

Testy zgodności

Popularne testy sprawdzające **zgodność rozkładu empirycznego z dowolnym rozkładem teoretycznym** w R to

- test chi-kwadrat, który może być używany do testowania zarówno rozkładów dyskretnych jak i absolutnie ciągłych, przy czym lepiej nadaje się do testowania rozkładów dyskretnych,
- test Kołmogorowa-Smirnowa, który służy do testowania rozkładów absolutnie ciągłych.

Ponadto w R mamy do dyspozycji testy które służą do testowania **zgodności z rozkładem normalnym**, np. test Shapiro-Wilka.

Test chi-kwadrat zgodności

Funkcja **chisq.test()** pozwala przeprowadzić testy oparte na statystyce chi-kwadrat.

Aby wykonać test chi-kwadrat zgodności, najpierw należy odpowiednio przygotować dane. Wartości obserwacji z próbki muszą być odpowiednio pogrupowane w klasy (szereg rozdzielczy punktowy bądź przedziałowy), aby liczebności klas mogły zostać porównane z oczekiwaną liczbą obserwacji (wyznaczoną przy założeniu, że dane pochodzą z danego rozkładu prawdopodobieństwa).

UWAGA: W każdej klasie powinno być conajmniej 10 obserwacji.

Test chi-kwadrat zgodności, c.d.

W przypadku testu zgodności pierwszym argumentem funkcji powinien być wektor lub szereg rozdzielczy uzyskany w wyniku użycia funkcji `table()`. Jeżeli nie podamy żadnego innego argumentu, funkcja domyślnie testuje hipotezę zerową zakładającą jednostajny rozkład prawdopodobieństwa (równe oczekiwane liczebności każdej klasy). Jeżeli nie chcemy testować równych proporcji, należy je podać jako kolejny argument

- `p` - wektor z prawdopodobieństwami (tego samego rozmiaru jak dane). Domyślnie jest to wektor prawdopodobieństw rozkładu jednostajnego, tzn. każdej obserwacji w próbie długości n przypisuje jednakowe prawdopodobieństwo równe $\frac{1}{n}$,
- `rescale.p` - jeżeli TRUE, to wektor z prawdopodobieństwami p jest skalowany tak, aby jego składowe sumowały się do 1 (domyślnie FALSE).

Test chi-kwadrat zgodności: przykład

Przykład: Wczytamy do `R` dane znajdującą się w pliku `dane.csv` a następnie przeprowadzimy na nich test zgodności z rozkładem Poissona.

Test chi-kwadrat zgodności: przykład

Przykład: Wczytamy do R dane znajdującą się w pliku `dane.csv` a następnie przeprowadzimy na nich test zgodności z rozkładem Poissona.

```
> dane<-read.table("dane.csv", sep=";")
> l<-mean(dane) # parametr λ rozkładu Poissona przybliżony
średnią,
> # generujemy wektor p prawdopodobieństw rozkładu Poissona
> prob<-c()
> for (i in 0:5) {
+  prob[i+1]=dpois(i,l)
+ }
> chisq.test(table(dane), p=prob, rescale.p=T)
```

Test chi-kwadrat zgodności: przykład

Przykład: Wczytamy do R dane znajdującą się w pliku `dane.csv` a następnie przeprowadzimy na nich test zgodności z rozkładem Poissona.

```
> dane<-read.table("dane.csv", sep=";")
> l<-mean(dane) # parametr λ rozkładu Poissona przybliżony
średnią,
> # generujemy wektor p prawdopodobieństw rozkładu Poissona
> prob<-c()
> for (i in 0:5) {
+  prob[i+1]=dpois(i,l)
+ }
> chisq.test(table(dane), p=prob, rescale.p=T)
```

W wyniku otrzymujemy *p*-wartość równą 0.644 (bardzo duża!), co zinterpretujemy na korzyść hipotezy zgodności danych z rozkładem Poissona.

Test chi-kwadrat zgodności: przykład

Przykład: Wczytamy do R dane znajdującą się w pliku dane.csv a następnie przeprowadzimy na nich test zgodności z rozkładem Poissona.

```
> dane<-read.table("dane.csv", sep=";")
> l<-mean(dane) # parametr λ rozkładu Poissona przybliżony
średnią,
> # generujemy wektor p prawdopodobieństw rozkładu Poissona
> prob<-c()
> for (i in 0:5) {
+  prob[i+1]=dpois(i,l)
+ }
> chisq.test(table(dane), p=prob, rescale.p=T)
```

W wyniku otrzymujemy p-wartość równą 0.644 (bardzo duża!), co zinterpretujemy na korzyść hipotezy zgodności danych z rozkładem Poissona.

UWAGA: Wektor prawdopodobieństw musi być wyraźnie wskazany poprzez przyrównanie (p=), w przeciwnym przypadku R źle go zinterpretuje.

Test Kołmogorowa-Smirnowa

Funkcja **ks.test()** pozwala przeprowadzić jedno- lub dwupróbkowy test Kołmogorowa-Smirnowa zgodności rozkładów.

Pierwszym argumentem funkcji jest wektor zawierający próbkę, zaś kolejne argumenty są następujące:

- wektor z danymi (jeżeli testujemy zgodność rozkładów dwóch prób) lub ciąg znaków określający nazwę funkcji (własną lub zaimplementowaną w `R`) definiującą dystrybuantę absolutnie ciągłego rozkładu teoretycznego,
- `alternative`: two.sided / less / greater - określa hipotezę alternatywną (domyślnie: dwustronna),
- `exact` - wartość logiczna określająca, czy ma być obliczana dokładna p-wartość testu (opcja niedostępna w przypadku jednostronnego testu dwupróbkowego bądź gdy występują tzw. **węzły**, czyli identyczne wartości obserwacji w próbie).

UWAGA: W przypadku rozkładów absolutnie ciągłych obecność węzłów może być niepokojąca (często wynika ona z dokładności zaokrąglania liczb) i może mieć istotny wpływ na wynik testu!

Test Kołmogorowa-Smirnowa jednopróbkowy: przykład

Przykład testu jednopróbkowego: Porównamy zgodność rozkładu próbki losowej wygenerowanej uprzednio z rozkładu normalnego $N(1, 1)$ z rozkładem jednostajnym oraz normalnym:

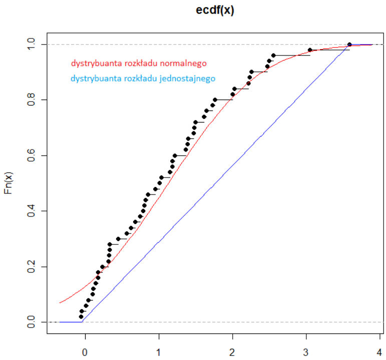
```
> probka=rnorm(50, mean=1)
> ks.test(probka, "punif", min(probka), max(probka))
> ks.test(probka, "pnorm", mean(probka))
```

Wynik porównania rozkładu próby z rozkładem jednostajnym nie jest jednoznaczny, mimo, że p-wartość jest raczej „mała”, to dla niektórych poziomów istotności hipoteza zerowa nie będzie mogła zostać odrzucona (może to wynikać z małej liczby obserwacji w próbie). Natomiast w drugim przypadku wątpliwości nie ma. P-wartość jest na tyle duża, że nie daje szans na odrzucenie hipotezy o zgodności z rozkładem normalnym.

Test Kołmogorowa-Smirnowa jednopróbkowy: wykresy dystrybuant

Test bazuje na statystyce opartej na odległości między dystrybuantą empiryczną próbki a dystrybuantą teoretyczną. Sprawdźmy, jak wyglądają odpowiednie wykresy dystrybuant:

```
> plot.exdf(probka)
> curve(pnorm(x,mean(probka)), add=T, col="red",)
> curve(punif(x,min(probka),max(probka)), add=T, col="blue")
```



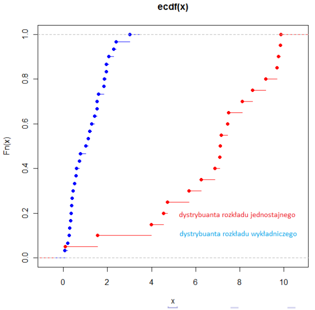
Test Kołmogorowa-Smirnowa dwupróbkowy: przykład

Przykład testu dwupróbkowego: Porównamy zgodność rozkładów próbki losowej wygenerowanej uprzednio z rozkładu jednostajnego $U(0, 10)$ oraz standardowego wykładniczego $E(1)$

```
> x=runif(20, min=0, max=10)
> y=rexp(30)
> ks.test(x,y)
```

Bardzo mała p-wartość (rzędu $e - 08$) wskazuje, że próby nie pochodzą z tych samych rozkładów, czego się zresztą spodziewaliśmy. Spójrzmy na całkiem inne wykresy dystrybuant empirycznych:

```
> plot.ecdf(x, col="red")
> plot.ecdf(y, col="blue", add=T)
```



Test Kołmogorowa-Smirnowa dwupróbkowy: przykład alternatyw jednostronnych

Jak widać, dystrybuanta empiryczna próby x leży poniżej dystrybuanty empirycznej próby y. W takim razie przeprowadźmy test Kołmogorowa-Smirnowa **z jednostronnymi alternatywami**.

Przetestujmy najpierw hipotezę zerową wobec alternatywy mówiącej, że dystrybuanta rozkładu x jest **nie mniejsza** (leży **powyżej**) dystrybuanty y (x jest stochastycznie mniejsze niż y):

```
> ks.test(x,y, alternative="greater")
```

Otrzymaliśmy p-wartość na tyle dużą, że nie możemy odrzucić hipotezy o równości rozkładów x i y wobec hipotezy, że dystrybuanta x leży powyżej dystrybuanty y.

Sprawdźmy zatem hipotezę alternatywną mówiącą, że dystrybuanta rozkładu x jest **mniejsza** (leży **poniżej**) dystrybuanty y (x jest stochastycznie większe od y):

```
> ks.test(x,y, alternative="less")
```

Otrzymana p-wartość jest na tyle mała, że pozwala nam odrzucić hipotezę o równości rozkładów na rzecz tej, która mówi o tym, że prawdziwa dystrybuanta rozkładu x jest mniejsza niż dystrybuanta y.

Test Shapiro-Wilka

Test Shapiro-Wilka to test **zgodności z rozkładem normalnym**.

Funkcja **shapiro.test()** pozwala przeprowadzić test Shapiro-Wilka.

Przykład:

```
> x<-rgamma(50,3,3)
> shapiro.test(x)
```

Tak jak się spodziewaliśmy, p-wartość jest na tyle mała, że odrzucamy hipotezę mówiącą o tym, że próba x pochodzi z rozkładu normalnego.

```
> y<-rnorm(20,3,2)
> shapiro.test(y)
```

W tym przypadku nie mamy wątpliwości - wynik testu nie pozwala odrzucić nam hipotezy zerowej, zatem przyjmujemy, że rozkład próby y jest normalny.

Testy normalności - pakiet `nortest`

W R dostępnych jest wiele innych testów badających **zgodność z rozkładem normalnym** (niektóre z nich to modyfikacje testu Kołmogorowa-Smirnowa). Są one dostępne w pakiecie `nortest`:

- test Andersona-Darlinga: funkcja `ad.test()`,
- test Cramera-Von Misesa: funkcja `cvm.test()`,
- test Lillieforsa: funkcja `lillie.test()`,
- test normalności chi-kwadrat Pearsona: funkcja `pearson.test()`,
- test Shapiro-Francia: funkcja `sf.test()`.

Jak testować normalność rozkładu?

- Wykonujemy podstawową analizę statystyczną: skośność (powinna być bliska 0), kurtoza (powinna być bliska 0 - w IBM SPSS lub 3- w *R*), histogram, wykres skrzynkowy, wykresy kwantyl-kwantyl.
- Jeśli obserwacji nie jest dużo (poniżej 2000), stosujemy test Shapiro - Wilka (lub test Andersona - Darlinga). Jeżeli obserwacji jest powyżej 2000, stosujemy test Kołmogorowa z poprawką Lillieforsa.

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

```
> set.seed(100)
> x <- rbinom(15,5,.6)
> shapiro.test(x)
```

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

$>$ set.seed(100)

$>$ x $<-$ rbinom(15,5,.6)

$>$ shapiro.test(x) - *rozkład normalny? NIE!*

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

```
> set.seed(100)
> x <- rbinom(15,5,.6)
> shapiro.test(x)        - rozkład normalny? NIE!
> x <- rlnorm(20,0,.4)
> shapiro.test(x)
```

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

> set.seed(100)

> x < − rbinom(15,5,.6)

> shapiro.test(x) - *rozkład normalny? NIE!*

> x < − rlnorm(20,0,.4)

> shapiro.test(x) - *rozkład normalny? NIE!*

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

```
> set.seed(100)
> x <- rbinom(15,5,.6)
> shapiro.test(x)        - rozkład normalny? NIE!
> x <- rlnorm(20,0,.4)
> shapiro.test(x)        - rozkład normalny? NIE!
```

Przykład: W przypadku **dużych** prób, nawet małe odchylenie od normalności może prowadzić do odrzucenia hipotezy zerowej.

```
> library(nortest)
> x <- rt(500000,200)
> ad.test(x)
```

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

```
> set.seed(100)
> x <- rbinom(15,5,.6)
> shapiro.test(x)        - rozkład normalny? NIE!
> x <- rlnorm(20,0,.4)
> shapiro.test(x)        - rozkład normalny? NIE!
```

Przykład: W przypadku **dużych** prób, nawet małe odchylenie od normalności może prowadzić do odrzucenia hipotezy zerowej.

```
> library(nortest)
> x <- rt(500000,200)
> ad.test(x)        - rozkład nie jest normalny?
```

UWAGI dotyczące testowania normalności

Żaden test nie stwierdzi wprost, że dane pochodzą z rozkładu normalnego! Test jest tylko w stanie wskazać kiedy dane są wystarczająco niezgodne z rozkładem normalnym i wówczas należy odrzucić hipotezę zerową.

Przykład: Gdy próbka jest **mała**, nawet duże odchylenia od normalności mogą nie zostać wykryte.

```
> set.seed(100)
> x <- rbinom(15,5,.6)
> shapiro.test(x)
```
- rozkład normalny? NIE!
```
> x <- rlnorm(20,0,.4)
> shapiro.test(x)
```
- rozkład normalny? NIE!

Przykład: W przypadku **dużych** prób, nawet małe odchylenie od normalności może prowadzić do odrzucenia hipotezy zerowej.

```
> library(nortest)
> x <- rt(500000,200)
> ad.test(x)
```
- rozkład nie jest normalny?
```
> hist(x)
> qqnorm(x)
```

Literatura

- **Testu chi-kwadrat zgodności**, data dostępu: 08.01.2017
- Vito Ricci, **Fitting distributions with R**, data dostępu: 08.01.2017
- Jacek Koronacki, Jan Mielniczuk, **Statystyka dla studentów kierunków technicznych i przyrodniczych**, WNT, Warszawa, 2001
- Przemysław Biecek, **Przewodnik po pakiecie** R, Oficyna Wydawnicza GiS, Wrocław, 2011
- Łukasz Komsta, **Wprowadzenie do środowiska** R, data dostępu: 13.10.2011
- Joseph Adler, **R in a Nutshell**, O'Reilly Media, 2009