Testy statystyczne w R.
Testy zgodności

Agnieszka Goroncy

UNIWERSYTET
MIKOŁAJA KOPERNIKA
W TORUNIU
Wydział Matematyki
i Informatyki

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, ale 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 rozkładu empirycznego wyłącznie z rozkładem normalnym**, np. test Shapiro-Wilka.

Test chi-kwadrat zgodności

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

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

Uwaga: W każdej klasie powinno być co najmniej 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`=TRUE, jeśli składowe wektora z prawdopodobieństwami `p` nie sumują się do 1 i należy je przeskalować (domyślnie FALSE).

Test chi-kwadrat zgodności: przykład

Przykład: Przeprowadzimy test zgodności chi-kwadrat z rozkładem Poissona dla zmiennej z pliku `dane.csv`.

Zaczynamy od wczytania pliku do R:

```
> dane=read.table("dane.csv", sep=";")
```

Rozkład Poissona jest rodziną rozkładów indeksowanych dodatnim parametrem λ. I tu jest pierwszy problem: test zgodności chi-kwadrat sprawdza zgodność z **konkretnym rozkładem**, a nie z rodziną. Żeby skonkretyzować rozkład Poissona, musimy wybrać jakieś λ. Wartość oczekiwana dla rozkładu Poissona wynosi akurat λ. Zatem dobrym estymatorem parametru λ jest **średnia z próby**.

```
> l=mean(dane$x) # parametr λ przybliżamy średnią
```

Teraz zwróćmy uwagę, jakie wartości przyjmuje badana zmienna:

```
> t=table(dane)
```

I tu mamy drugi problem: zmienna losowa o rozkładzie Poissona może przyjąć każdą wartość nieujemną całkowitą, a nasza zmienna przyjmuje tylko sześć wartości 0,1,2,3,4,5. Co mamy robić?

Test chi-kwadrat zgodności: przykład - c.d.

Popatrzmy, jakie są prawdopodobieństwa przyjęcia wartości 0,1,2,3, 4,5 dla zmiennej o rozkładzie Poissona z parametrem $\lambda = 1.7$:

```
> prob=c() # tworzymy wektor prawdopodobieństw rozkładu Poissona
> for (i in 0:5) {
+  prob[i+1]=dpois(i,l)
+ }
> sum(prob)
```

Widzimy, że suma tych sześciu prawdopodobieństw wynosi 0.9920006, czyli jest prawie jedynką. Zatem można w przybliżeniu uznać, że reszta wartości $(6, 7, 8, \ldots)$ nie jest przyjmowana. Stosujemy więc test chi-kwadrat zgodności:

```
> chisq.test(t, p=prob, rescale.p=TRUE)
```

Otrzymujemy *p*-wartość równą 0.644 (duża!), co interpretujemy na korzyść hipotezy zerowej o 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: Porównajmy zgodność rozkładu próbki losowej wygenerowanej uprzednio z rozkładu normalnego $N(1, 1)$ z rozkładem jednostajnym czy też normalnym:

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

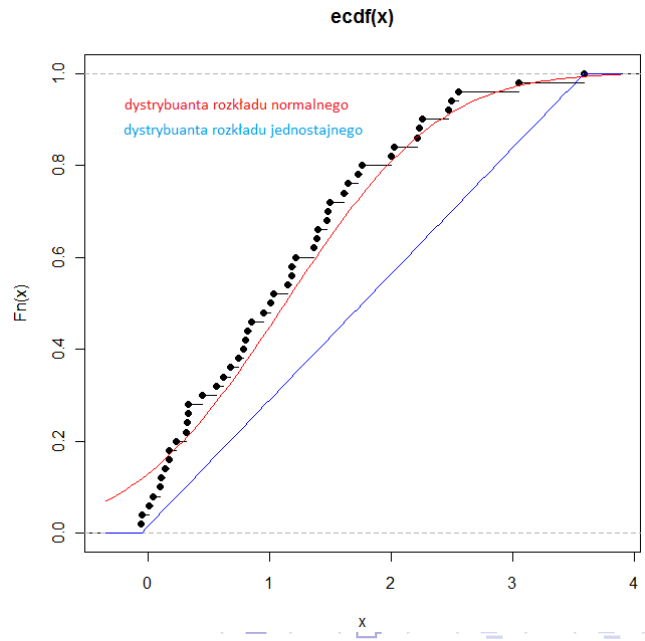
Uwaga: W pierwszym przypadku precyzujemy testowany rozkład jednostajny jako rozkład na przedziale (min(probka), max(probka)); końce przedziału są parametrami rozkładu jednostajnego i wybrane wartości służą oszacowaniami tych końców. W drugim przypadku precyzujemy testowany rozkład normalny poprzez wybranie średniej z próby jako oszacowania wartości oczekiwanej (przecież zakładamy, że nie znamy rozkładu, z którego pochodzi próbka).

Wynik porównania rozkładu próby z rozkładem jednostajnym: p-wartość jest raczej „mała" i należy odrzucić hipotezę zerową o zgodności z rozkładem jednostajnym, choć dla niektórych poziomów istotności hipoteza ta może nie zostać odrzucona (co wynika z niedużej liczby obserwacji w próbie). Natomiast w drugim przypadku wynik jest bardziej zdecydowany: p-wartość jest na tyle „duża", że nie należy odrzucać hipotezy zerowej o zgodności z rozkładem normalnym.

Test Kołmogorowa-Smirnowa jednopróbkowy: wykresy dystrybuant

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

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



Wniosek jest oczywisty: różnica między dystrybuantą empiryczną próbki a dystrybuantą rozkładu jednostajnego jest duża, ale już z dystrybuantą rozkładu normalnego różnica jest mała.

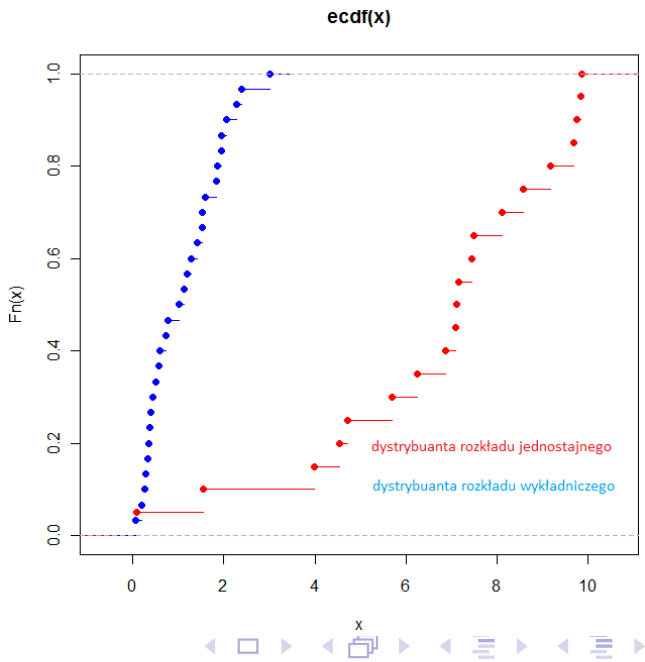
Test Kołmogorowa-Smirnowa dwupróbkowy: przykład

Przykład testu dwupróbkowego: Porównajmy zgodność rozkładów próbek losowych wygenerowanych uprzednio z rozkładu jednostajnego $U(0, 10)$ oraz z rozkładu standardowego wykładniczego $E(1)$:

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

Bardzo mała p-wartość wskazuje, że próby nie pochodzą z tych samych rozkładów (odrzucamy hipotezę zerową o równości rozkładów próbek), 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

W poprzednim przykładzie można zauważyć, że dystrybuanta empiryczna próby x leży poniżej dystrybuanty empirycznej próby y. W takim przypadku możemy przeprowadzić test Kołmogorowa-Smirnowa **z jednostronnymi alternatywami**.

Przetestujmy najpierw hipotezę zerową o równości rozkładów x i y 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 zerowej o równości rozkładów x i y.

Sprawdźmy zatem hipotezę zerową o równości rozkładów x i y wobec alternatywnej mówiącej, 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, że dystrybuanta rozkładu x jest mniejsza niż dystrybuanta rozkładu y.

Test Shapiro-Wilka

Test Shapiro-Wilka jest **testem zgodności wyłącznie z rozkładem normalnym**.

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

Przykłady:

```
> 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 nam odrzucić hipotezy zerowej, zatem przyjmujemy, że rozkład próby y jest normalny.

Inne 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()`
  (jest to test Kołmogorowa-Smirnowa z poprawką Lillieforsa),
- test normalności chi-kwadrat Pearsona: funkcja `pearson.test()`,
- test Shapiro-Francia: funkcja `sf.test()`.

Jak sprawdzać normalność rozkładu?

Ogólne rekomendacje sprawdzenia normalności rozkładu zmiennej są następujące.

- Wykonujemy podstawową analizę statystyczną: liczymy skośność (powinna być bliska 0), kurtozę (powinna być bliska 3), rysujemy histogram, wykres skrzynkowy, wykres kwantyl-kwantyl.

- Jeśli obserwacji nie jest bardzo dużo (poniżej 5000), stosujemy test Shapiro-Wilka (lub test Andersona-Darlinga). Jeżeli obserwacji jest powyżej 5000, stosujemy test Kołmogorowa-Smirnowa z poprawką Lillieforsa (lub test Andersona-Darlinga).

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 o normalności.

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

```
> set.seed(100)
> x=rbinom(15,5,0.6)
```

Wygenerowaliśmy próbkę z rozkładu dwumianowego, który jest **rozkladem dyskretnym**, i w żaden sposób nawet nie jest bliski do rozkladu normalnego. Przeprowadzimy test na normalność:

```
> shapiro.test(x)
```

Wynik wskazuje, że nie mamy podstaw do odrzucenia hipotezy zerowej o normalności rozkładu próbki!

Uwagi dotyczące testowania normalności - c. d.

Jeszcze jeden przyklad:

```
> x=rlnorm(20,0,0.4)
```

Wygenerowaliśmy próbkę z rozkładu lognormalnego, który nie jest rozkładem normalnym. Test normalności

```
> shapiro.test(x)
```

znowu wskazuje, że nie mamy podstaw do odrzucenia hipotezy zerowej o normalności rozkładu próbki.

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

```
> install.packages("nortest")
> library(nortest)
```

Wygenerujemy bardzo dużą próbkę rozmiaru 500000 z rozkładu Studenta o 200 stopni swobody:

```
> x=rt(500000,200)
```

Uwagi dotyczące testowania normalności - c. d. (2)

Rozkład Studenta o takiej liczbie stopni swobody jest nieodróżnialny od rozkładu normalnego. W potwierdzenie tych słów można przytoczyć histogram, czy też wykres kwantyl-kwantyl:

```
> hist(x)
> qqnorm(x)
```

Popatrzmy jednak, jaki jest wynik testu na normalność próbki (stosujemy np. test Andersona-Darlinga, bo rozmiar próby jest bardzo duży):

```
> ad.test(x)
```

Widzimy, że mała p-wartość sugeruje jednak odrzucenie hipotezy zerowej o normalności rozkładu próbki.

Literatura

- Vito Ricci, **Fitting distributions with R**
- 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
- Joseph Adler, **R in a Nutshell**, O'Reilly Media, 2009