Вероятности и статистика с R

Асен Чорбаджиев

January 8, 2018

1 Дефиниции.

Когато се прави статистическо изследване, основното твърдение се нарича нулева хипотеза $H(0)$ и тя проверява вероятността за даден параметър θ да бъде равен на d. Всяко друго твърдение се нарича алтернативно твърдение. Тогава в частният случай когато $H(0) : \theta = 0$, $H(A) : \theta \neq 0$.
При изследване на дадена хипотеза, обаче, могат да бъдат два вида грешки в анализа на резултатите:
Грешка от първи род ("false positive") Грешно отхвърляне на вярна нулева хипотеза.
Грешка от втори род ("false negative") Грешно потвърждение на нулева хипотеза.
p-value Вероятностните стойности (p), които измерват вероятността за близост на измерваното разпределение до приетата за вярна нулевата хипотеза се наричат p-values. По-точно, това е вероятността за добиване на разлика в измерванията, или най-високата вероятносттна стойност за това, че няма разлика между контролното разпределение и наблюдаваното. Когато се изследват множество хипотези, p-values оценява силата на всяка хипотеза, затова най-доброто приближение е с най-висока верояност. Когато p-value е равно или по-малко от ниво (significance level) α, тогава нулевата хипотеза се отхвърля в полза на алтернативната. Когато, обаче, p-value е по-голямо от α, казваме само, че нулевата хипотеза не е отхвърлена.
За определяне на p за нулева хипотеза за средно μ се използват вероятностите на t-разпределението с n-1 степени на свобода.

2 Графически тестове за вид на разпределението

Quantile-quantile (q-q) plot е графически метод за определяне дали две извадки произхождат от множества с еднакви разпределения. Графиката съпоставя квантилите от първата извадка срещу тези на втората. Под квантил се разбира частта от точки под дадена стойност, т.е. квантил 0.3 е точка, където 30% от данните са под дадената стойност. В този тест размерите на извадките не е задължиелно да бъде равен. Функцията в R се нарича qqplot(). Много често се налага данни да бъдат сравнявани с Нормално разпределение. Това става в qqnorm().

Пример:

```
y <- rnorm(2000)*4
qqnorm(y); qqline(y, col = 2,lwd=2,lty=2)
y <- rnorm(2000)*4 - 4
qqnorm(y); qqline(y, col = 2,lwd=2,lty=2)
y <- rexp(2000,1)
qqnorm(y); qqline(y, col = 2,lwd=2,lty=2)
x <- rpois(2000,1)
y <- rbinom(2000,size=10,prob=1/10)
qqplot(x,y); qqline(y, col = 2,lwd=2,lty=2)
y <- rbinom(2000,size=10,prob=1/2)
qqplot(x,y); qqline(y, col = 2,lwd=2,lty=2)
```

3 Statistical Hypothesis Tests

3.1 Сравняване на пропорции

Проверява дали пропорцията в една нормално разпределена група отговаря на пропорцията в нормално разпределена друга или други сходни групи. Функцията в R е prop.test().

Пример:

```
sexsmoke<-matrix(c(70,120,65,140),ncol=2,byrow=T)
rownames(sexsmoke)<-c("male","female")
colnames(sexsmoke)<-c("smoke","nosmoke")
prop.test(sexsmoke)
prop.test(c(70,65),c(190,205)) # Идентично изпълнение.
prop.test(c(70,65),c(190,205),conf.level=0.99) # Смянана на α
prop.test(c(70,65),c(190,205),c(0.33,0.33)) # Предварително заложени пропорции.
```

3.2 t-test

За анализ на извадка $Z_n = N(\mu, \sigma^2)$, с извадкови средно и дисперсия $\overline{\mu_x}$, $\overline{d_x}$. Оценката

$$T_n = \sqrt{(n-1)}\frac{\overline{m_x} - \mu}{\sqrt{\overline{d_x}}} \tag{1}$$

е t(n-1) - разпределена. Тогава Т-статистиката за хипотеза μ_0

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{2}$$

дава p-values получени от t(n-1) - разпределението за следните хипотези:

$$p - val = \begin{cases} P\{T \leq t|H_0\} : \mu < \mu_0 \\ P\{T \geq t|H_0\} : \mu > \mu_0 \\ P\{|T - \mu_0| \geq |t - \mu_0|\} : \mu \neq \mu_0 \end{cases} \tag{3}$$

За пресмятане на теста в се използва t.test(x, mu, alt), като mu отговаря на стойността μ_0 в нулевата хипотеза.
Пример:
mpg= c(11.4,13.1,14.7,14.7,15.0, 15.5,15.6,15.9,16.0,16.8)
xbar=mean(mpg); s=sd(mpg); n=length(mpg)
SE=s/sqrt(n)
est=(xbar-17)/SE
pt(est,df=n-1, lower.tail=T) # Пресмятане на ръка.
t.test(mpg,mu=17,alt="less") # Автоматично

3.3 Тестове с медиани

Когато за оценка се налага да бъде използвана медиана, а не средно, се използват т.н. Wilcox test. Този тип тестове се използва също в повторни измервания, когато се сравняват данните поелементно. Така от N двойки се образуват N_r от тях, където са премахнати двойките за които разликата $x_{2,i} - x_{1,i}$ е различна от 0. Тези двойки се нареждат поред на големина, с най-малката стойност начело и им се поставя номер (rank) R_i. Тогава статистиката се образува от:

$$W = \sum_{i=1}^{N_r}(sign(x_{2,i} - x_{1,i})R_i) \tag{4}$$

Нулевата хипотеза се тества за $W = 0$. Това се прави за критически стойности от Z-статистиката на теста.
Функцията в R се нарича wilcox.test().

3.4 F-test

F-тестовете се прилагат когато дисперсиите на две извадки са значително различни. За нулева хипотеза за извадки със стандартни дисперсии σ_1^2, σ_2^2 се приема $H(0) : \sigma_1^2 = \sigma_2^2$. Възможните алтернативи са:

$$H(\alpha) : \begin{cases} \sigma_1^2 < \sigma_2^2;\ lower\ one - tailed \\ \sigma_1^2 > \sigma_2^2;\ upper\ one - tailed \\ \sigma_1^2 \neq \sigma_2^2;\ two - tailed \end{cases} \tag{5}$$

Значимостта на теста тогава се пресмята съгласно статистиката $F = \sigma_1^2/\sigma_2^2$ и $H(0)$ се счита за отхвърлена с критична стойност α когато:

$$p - val = \begin{cases} F > F_{\alpha,N_1-1,N_2-1}\ for\ upper\ one - tailed \\ F < F_{1-\alpha,N_1-1,N_2-1}\ for\ lower\ one - tailed \\ F > F_{\alpha/2,N_1-1,N_2-1}\ for\ two - tailed \end{cases} \tag{6}$$

където с F_{α,N_1-1,N_2-1} е означен α квантила на $F(N_1-1, N_2-1)$ разпределение. Функцията в R за F-test е var.test().
Пример:

Table 1: Two formulas of gentamicin

Formula	Patients (n)	Stand. Deviation (mg/l)
A	10	3.0
B	15	2.0

F = 3^2/2^2 # F - statistics
df(F,9,14) # p-value
qf(0.975,9,14)

4 Тестове за хомогенност и еднаквост

4.1 Хи-квадрат тест

Отговор на въпроса дали измервания с r редове и c брой колони са независими се дава с Хи-квадрат тест за независимост. Същият тест се използва също така и за оценка на Goddness of Fit между различни разпределения. За него нулевата хипотеза е, че променливите са независими. За такава оценка се използва следната статистика:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{7}$$

Където с O_i е означена текущата честота, а очакваната е $E_i = np_i$. Оценката се прави съгласно стойностите на на Хи-квдрат разпределението с n-1 за n параметъра, като нулевата хипотеза се отхвърля когато резултатът от статистиката е от критическата стойност на разпределението. Функцията за директно премятане в R е chisq.test().

Пример (Решение на задача 7):

```
# Data
nk=c(229,211,93,35,7,1)
n=sum(nk)
k=c(0,1,2,3,4,6)
r=sum(nk*k)/n # The rate of Poisson distribution
pk=dpois(0:4,r)*n # Expected values
pk[6]=n-sum(pk) # Complete distribution
# Аналитично решение
chi2=sum(((nk - pk) **2)/pk) #Chi-Square statistics
pchisq(chi2, df=5,lower.tail=F) # p-value # Решение с R
tbl=cbind(nk,pk)
chisq.test(tbl)
```

4.2 Колмогоров-Смирнов

Колмогоров-Смирнов тест може да бъде използван за:
- проверка за това дали дадена извадка има предполагаемо неприкъснато разпределение F(x)
- сравняване на две изваски. Предимството на този тест пред Хи-квадрат теста е, че се избягва дискретизацията. Слабостта му е, че точността се центрира в средата на извадката. Също така той може да бъде използван само за непрекъснати разпределения. За целта се следва следната процедура:

- Пресмята се статистиката $S_n(x)$ от наредените стойности $x_1 \leq x_2 \leq ... \leq x_n$ по следния начин:

$$S_n(x) = \begin{cases} 0 & x \leq x_1 \\ k/n & x_k \leq x < x_{k+1} \\ 1 & x \geq x_n \end{cases} \tag{8}$$

- Пресмята се D_n:

$$D_n = max_{0 \leq x \leq 1} |F(X \leq x) - S_n(x)| \tag{9}$$

- Повеже $S_n(x)$ зависи от n, D_n е случайна величина. Така теста D_n сравнява F(x), която е теоретичната кумулативна функция на търсеното разпределение. За намиране на критичните стойности $D_{n,\alpha}$ се използва разпределението на Колмогоров-Смирнов:

$$F(x) = \frac{\sqrt{1\pi}}{x} \sum_{k=1}^{\infty} e^{-(2k-1)^2\pi^2/(8x^2)} \tag{10}$$

Тогава $P(D_n \leq D_{n,\alpha}) = 1 - \alpha$.

За пресмятане в R се използва функцията ks.test. Параметърът за определяне се задава с име на съответната функция в R за пресмятане на кумулативното разпределение, като например *pnorm*.
Пример:
x=c(0.58, 0.42, 0.52, 0.33, 0.43, 0.23, 0.58, 0.76, 0.53, 0.64)
 # Test for Uniform distribution
ks.test(x, "punif")

4.3 Тест за нормалност

Когато се налага проверка за нормалност, много по-прост и често удобен тест е Shapiro-Wilk. Функцията в R е shapiro.test().
Пример:
x=rnorm(100,mean=1,sd=2)
shapiro.test(x)
ks.test(x, "pnorm", 1,2)

5 Упражнения:

- **Зад. 1.**: В таблицата **quine** от **MASS** е дадена статистика от малък австралийки град за децата по етническа принадлежност, пол, възраст, учебен статус и брой дни неприсъствие в клас. Групирайте данните по пол и етническа принадлежност. Да се намери 95% интервална оценка за разликата между пропорцията жени от аборигенската популация и останалите, всяка в своята етническа група.
- **Зад. 2.**: Дадени са 25 измервания от експеримент x=c(170, 167, 174, 179, 179, 156, 163, 156, 187, 156, 183, 179, 174, 179, 170, 156, 187, 179, 183, 174, 187, 167, 159, 170, 179). Да се тества нулевата хипотеза $H(0)$ за $\mu = 170$ срещу алтернативните $\mu > 170$ и $\mu < 170$.
- **Зад. 3.**: Производител твърди, че произвежда детайли с размер 7.5 инча. В контрола на качеството са взети 10 проби x=c(7.65, 7.60 ,7.65

,7.70, 7.55, 7.55, 7.40, 7.40, 7.50, 7.50). Да се провери нулевата хипотеза за това дали резмерът съвпада с предварително зададения размер.

- **Зад. 4.**: Да се провери нулевата хипотеза да медиана по-голяма от 5 за x = c(12.8,3.5,2.9,9.4,8.7,0.7,0.2,2.8,1.9,2.8,3.1,15.8).
- **Зад. 5.**: В таблицата **survey** от **MASS** е дадена статистика от студенти. В колоните Smoke са дадени данни за пушачите, а в Exer за спортните навици. Да се групират в таблица и да се провери дали съществува зависимост.
- **Зад. 6.**: Да се провери хипотезата за равенство на дисперсиите две извадки от по 100 елемента с N(0,1) и N(1,1). Какво ще стане ако се промени размера на едната извадка на 50, например?
- **Зад. 7.**: По време на бомбандировките в Лондон са разчистени 576 участъка. Те са групирани по брой намерени снаряди: Да се провери

Table 2: Сектори и брой снаряди

k	0	1	2	3	4	6
n_k	229	211	93	35	7	1

за зависимост с моделиране със съответен Поасонов закон с $\theta = 0.93$.

- **Зад. 8.**: Проверете за нормалност графически и с подходящ тест скоростта на светлината от таблица morley.