Вероятности и статистика с R

Асен Чорбаджиев

November 16, 2017

1 Математическо очакване и Дисперсия на експериментални данни

Когато се работи с експериментални данни $X_1, ..., X_n$, очакването $E(X) = \overline{X}$, дисперсията Var(x) и стандартното отклонение s:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{1}$$

се изчисляват с mean(), var(), sd(). Например:
x = c(74, 122, 235, 111, 292, 111, 211, 133, 156, 79)
var(x)
sd(x)

2 Централна Гранична Теорема

Когато емпиричните данни S_n са придобити от независими опити с биномно разпределение с $B(n, p)$ тогава стандартизираното средно:

$$Z = \frac{S_n - np}{\sqrt{np(1-p)}} \tag{2}$$

асимптотически е със Стандартно Нормално разпределение N(0,1).
Когато емпиричните данни X_i са нормално разпределени с $N(\mu, \sigma^2)$ тогава стандартизираното средно:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{3}$$

е със Стандартно Нормално разпределение N(0,1). Когато n е достатъчно голямо (m » 30), то $\overline{X} \sim N((\mu, \sigma^2/n))$

3 Разпределения производни на нормалното

3.1 Хи-квадрат

Ако Y_i е нормално разпределено с очакване 0 и дисперсия 1, то

$$\chi^2(r) = \sum_{i=1}^{r} Y_i^2 \tag{4}$$

е Хи-квадрат разпределено с r степени на свобода. Тогава вероятността е равна на:

$$P_r(x) = \frac{x^{r/1-1} e^{-x/2}}{\Gamma(r/2) 2^{r/2}} \tag{5}$$

, където с $\Gamma(r/2)$ е означена гамма функцията. Функциите в R са chisq(), със съответните представки d,p,q,r за вероятност, разпределение, квантил и генератор на случайни числа.

3.2 t-разпределение

Нека имаме N независими измервания x_i с очакване μ, извадково средно $\overline{x}$, оценка за стандартно отклонение s и неизвестно σ. Тогава най-добрата оценка на неизвестно σ се дава с:

$$t(n) = \frac{\overline{x} - \mu}{s/\sqrt{N}} \tag{6}$$

и се нарича t-разпределение с n=N-1 степени на свобода. Функциите в R са t(), със съответните представки d,p,q,r за вероятност, разпределение, квантил и генератор на случайни числа.

3.3 F-разпределение

Непрекъснато статистическо разпределение получено от следното отношение на две Хи-квадрат разпределения $\chi^2(m)$, $\chi^2(n)$:

$$F_{n,m} = \frac{\chi^2(n)/n}{\chi^2(m)/m} \tag{7}$$

Употребява се за тестване дали две извадки имат иднаква дисперсия. Функциите в R са f(), със съответните представки d,p,q,r за вероятност, разпределение, квантил и генератор на случайни числа.

4 Математическо очакване и Дисперсия на изучените разпределения

Математическото очакване и дисперсия за изучените вероятностни разпределения са равни на:

Table 1: Mean, median and Variations

Разпределение	E(X)	Median	Var(X)	Описание
Равномерно	$\frac{1}{2}a+b$	$\frac{1}{2}a+b$	$\frac{1}{12}(b-a)^2$	[a,b]
Биномно	np		np(1-p)	B(n,p)
Отрицателно Биномно	$\frac{pr}{(1-p)}$		$\frac{pr}{(1-p)^2}$	NB(r,p)
Геометрично	$\frac{1-p}{p}$		$\frac{1-p}{p^2}$	Ge(p)
Хипергеометрично	$\frac{nN}{m+n}$		$\frac{mnN(m+n-N)}{(m+n)^2(m+n-1)}$	HG(m,n,N)
Поасоново	λ		λ	$Po(\lambda)$
Експоненциално	μ	$\mu ln(2)$	μ^2	$Exp(1/\mu)$
Хи-квадрат	r		$2r$	$\chi^2(r)$
t-разпределение	0	0	n/(n-2)	t(n); n>=2
F-разпределение	m/(m-2)		$\frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$	F(n,m); m>=4

5 Измервателна грешка

Често измерванията са съпроводени с несигурност изразяващи се с наличието на единични стойности, наречени outliers, които са твърде различни от общата извадка. За тяхната идентификация трябва да бъде оценена количествено достоверността. За тази цел могат да бъдат използвани използват графични и аналитични методи.

5.1 Достоверност

Когато в статистическо изследване някой желае да оцени неизвестен параметър θ на популацията с наблюдавани измервания, резултатът е с вероятност на достоверност $\gamma = 1 - \alpha = P(-z^\star < \theta < z^\star)$ и γ се нарича се ниво на

достоверност. Нивото на достоверност се задава според търсената точност и често е равна на квантили със стойности 0.90, 0.95, and 0.99. В реализацията на R намирането на $z^\star$ за нормално разпределение е симетрично и се пресмята с qnorm(1-α/2).

5.2 Графични методи

За графическо откриване на outliers се използа boxplot. Тя работи с определянето на граници съответните $Q_3 = 75\%$ и $Q_1 = 25\%$ квантили и съответните разстояния според размаха на $IQR = Q_3 - Q_1$, както е показано в следния код:

```
q1 <- qnorm(0.25)
q2 <- qnorm(0.5)
q3 <- qnorm(0.75)
lower <- q1 - 1.5*(q3-q1)
upper <- q3 + 1.5*(q3-q1)
tmp.list <- list( stats=rbind(lower, q1, q2, q3, upper), out=numeric(0), group=numeric(0),
names='')
bxp( tmp.list )
```

Функцията с име boxplot() рисува автоматично boxplot от данни. Важно е да си запомни, че този метод работи само с нормално разпределени данни и достатъчно голями по обем, т.е. повече от 30.

5.3 Доверителни интервали

Доверителен интервал дава оценка на размаха на стойности от стойности пресметнат от налични данни, които е възможно да включват неизвестен параметер в извадката.
В бинарен случай, когато имаме очаквана честота на успех p и такава измерена в извадка $\hat{p}$, интервалната оценка за p, наречена интервал на Wald е равна на:

$$[\hat{p} - z^\star SE, \hat{p} + z^\star SE], \tag{8}$$

където SE е равно на:

$$\sqrt{\frac{1}{n}\hat{p}(1-\hat{p})} \tag{9}$$

Когато трябва да се оценява разликата,$\theta = p_1 - p_2$ с размери n и m има γ доверителен интервал със SE=$\sqrt{\frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}}$ равен на:

$$[\theta - z^\star SE, \theta + z^\star SE] \tag{10}$$

В случай, когато $\theta = \mu$ е неизвестно, за нормално разпределенни данни доверителните интервали са равни на:

$$[\overline{x} - z^\star \frac{\sigma}{\sqrt{(n)}}, \overline{x} + z^\star \frac{\sigma}{\sqrt{(n)}}] \tag{11}$$

В случай, когато $\theta = \mu, \sigma$ са неизвестни, за нормално разпределенни данни доверителните интервали са равни на:

$$[\overline{x} - t^\star \frac{s}{\sqrt{(n)}}, \overline{x} + t^\star \frac{s}{\sqrt{(n)}}] \tag{12}$$

където $t^\star$ е квантилът на t-разпределение с n-1 степени на свобода.

6 Упражнения:

- **Зад. 1.**: Да се демонстрира графично ЦГТ със 100 случайно генерирани с разпределения x=Bi(100,1/2), Bi(100,1/3). Да се начертаят хистограмите използвайки параметъра prob=T. Да се начертаят с curve съответните вероятности за 100 случайно генерирани числа. Изберете подходящото разпределение.

- **Зад. 2.**: Да се изследва за наличие на otliers генерираните опити от зад. 1. Да се начертае графика.

- **Зад. 3.**: Студент прави 6 измервания на температурата на завиране на течност. Стойностите са (градуси Celsius) 102.5, 101.7, 103.1, 100.9, 100.5, and 102.2. Той знае, че стандартното отклонение σ на уреда е 1.2 градуса. Какъв е 95% доверителния интервал за експеримента и има ли грешки в измерванията?

- **Зад. 4.**: Да се провери графически разпределението на експеримента на Michaelson-Morley, наличен в таблицата morley. Да се изследва графически и с 95% 99.7% доверителни интервали за измервателна грешка.Да се начертаят графики за доверителните интервали.