R-cheat sheet

1 Генериране на редици

rep(5, times = 8); rep(c(1,2), times=5) // 1,2,1,2...
~~rep(5, each~~ = rep(c(1,2), each = 5) // 1,1,1,1,1,2,2,2,2,2
rep(c(1,2), length.out = 7) ; x<-5:12 ; y=c(10:1)
seq(from=1, to=10, by=2), seq(from=10, to=1, by=-2)
&

2. Матрици

M=rbind(c(5,3,5,6), c(8,3,7,4)) ; M[2,3] // 7 M[,3]
M=cbind(c(5,3,5,6), c(8,3,7,4)) ; M=matrix(c(1:12), nrow=3, nc
M= matrix(c(1:12), nrow=3, ncol=4, by row = TRUE)

3 Data frame

df = data.frame(x,y) ; df$x ; df$y df[5,]
df$x[df$y <=5] ; df[df$z <=5, c("x","z")]

4. Полезни функции

read.table(file); replace(x, list, values);
ifelse(test, yes, no); any(...); all(...); unique(...);
duplicated(...); is.element(x,y); x %in% y
tabulate(...); substr(x, start, stop)
sample(c(1:10), 5, replace=T); sample(c(1:7), 2, replace=T
any(diff==0); anyDuplicated(x)>0; all(c(1,2,3) %in% x)
x=c(1,2,2,3); unique(x) // 1,2,3; duplicated(x) // F F T F
is.element(x,2) // F T T F; anyDuplicated(x) // 3

5. Условна вероятност

prob.bw = function(Nrep){
res = replicate(Nrep, sim.bw())
sum(rs[2,]=="ww")/sum(rs[1,]=="w")
}
sim.bw = function(){ ..... c(up, card) }

6. Биномно разпределение: $X \sim Bi(n,p)$; $q = 1-p$

$EX = n.p$; $Var(X) = n.p.q$

$dbinom(k, n, p) = P(X = k) = \binom{n}{k} p^k . q^{n-k}$

$pbinom(k, n, p) = P(X \leq k)$

$rbinom(N, n, p)$ генерира N случайни числа от биномното разпределение с параметри n, p

7. Геометрично разпределение: p; $q = 1-p$; X - броя опити до първия успех (включително): $X \sim Ge(p)$

$EX = \frac{1}{p}$, $DX = \frac{q}{p^2}$

$dgeom(k-1, p) = P(X = k) = q^{k-1} p$

$pgeom(k-1, p) = P(X \leq k)$

$rgeom(N, p) + 1$ - генерира N случайни числа от Ge(p)

Y = брой неуспехи преди първия успех; $X = Y + 1$

$EY = \frac{q}{p}$ $DY = \frac{q}{p^2}$

$dgeom(k, p) = P(Y = k) = q^k . p$

$pgeom(k, p) = P(Y \leq k)$

$rgeom(N, p)$

8. Отрицателно биномно разпределение $X \sim BN(r, p)$

X - броя опити до r-ти успех (включително)

$EX = \frac{r}{p}$ $Var(X) = \frac{rq}{p^2}$

$dnbinom(k-r, r, p) = P(X = k) = \binom{k-1}{r-1} p^r q^{k-r}$

$pnbinom(k-r, r, p) = P(X \leq k)$

$rnbinom(N, r, p) + r$

Y = брой неуспехи пред r-ти успех ; $X = Y + r$

$EY = \frac{rq}{p}$ $Var(Y) = \frac{rq}{p^2}$

$dnbinom(k, r, p) = P(Y = k) = \binom{r+k-1}{r-1} p^r . q^k$

$pnbinom(k, r, p) = P(Y \leq k)$

$rnbinom(N, r, p)$

9. Хипергеометрично разпределение $HG(N, M, n)$

M - бели, $N-M$ - черни топки. Вадим n без да връщам

X - броя на извадени бели топки

$dhyper(k, M, N-M, n) = P(X=k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$

$phyper(k, M, N-M, n) = P(X \leq k$

$rhyper(R, M, N-M, n)$

10. Поасоново разпределение $X \sim Po(\lambda)$

$EX = \lambda$, $DX = \lambda$

$dpois(x, \lambda) = P(X=k) = e^{-\lambda} \frac{\lambda^k}{k!}$

$ppois(k, \lambda) = P(X \leq k)$

$rpois(N, \lambda)$

11. Равномерно разпределение: $X \sim U(a, b)$

$dunif(x, a, b) = f(x)$

$punif(q, a, b) = P(X \leq q) = F(q)$

$qunif(p, a, b) = Q(p) = F^{-1}(p)$

$runif(N, a, b)$

$f(x) = \begin{cases} \frac{1}{b-a}, & x \in (a,b) \\ 0, & x \in (a,b) \end{cases}$

$EX = \frac{a+b}{2}$, $Var(X) = \frac{(b-a)^2}{12}$

12. Експоненциално разпределение: $X \sim Exp(\lambda)$

$dexp(x, \lambda) = f(x)$

$pexp(q, \lambda) = P(X \leq q) = F(q)$

$qexp(p, \lambda) = Q(p) = F^{-1}(p)$

$rexp(N, \lambda)$

$f(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$

$EX = \frac{1}{\lambda}$, $Var(X) = \frac{1}{\lambda^2}$

13. Нормално разпределение: $X \sim N(\mu, \sigma^2)$

$dnorm(x, \mu, \sigma) = f(x)$

$pnorm(q, \mu, \sigma) = P(X \leq q) = F(q)$

$qnorm(p, \mu, \sigma) = Q(p) = F^{-1}(p)$

$rnorm(N, \mu, \sigma)$

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma}$

$EX = \mu$

$Var(X) = \sigma^2$

14. Данни. Таблици

barplot(sort(table(raws)/length(raws)), decreasing=T);
pie(table(raws)); wait_grp = cut(widit, breaks=seq(0,12,2));
table(wait_grp), hist(wait, breaks=seq(0,12,3))

15. Числови характеристики на данните

$\bar{x}$ = mean(x) - средна стойност; Me = median(x) - медиана
p-квантил = quantile(x, p); IQR = IQR(x) - интерквартилен размах
s = sd(x) - стандартно отклонение; $[\bar{x}-3s, \bar{x}+3s]$ - Неравенство на Чебишев
boxplot(x, horizontal=T)

16. Многомерни данни

prop.table(table(Smoke, W.Hand)); prop.table(table(Smoke, W.Hand), 1)
boxplot(Pulse ~ Smoke); cor(x, y, use="complet.obs")

17. Доверителни интервали

* t.test(x, conf.level=0.95)$conf.int [1:2] - 95-процентен доверителен интервал за средно при неизвестна дисперсия
* prop.test(x=2700, n=25000, conf.level=0.95, correct=F) - доверителен интервал за пропорция (вероятност за успех)

18. Проверка на хипотези при една извадка

* t.test(x, mu=5.2) - t-тест за средно (One sample t-test)
Ако p.value ≤ 0.05, отхвърляме H_0 в полза на H_1
Ако p.value ≥ 0.05, няма достатъчно основание да отхвърлим H_0
* prop.test(x=2700, n=25000, p=0.1, correct=F) - z-тест за пропорция (1-sample proportions test without ...)

19 Проверка на хипотези при 2 извадки

* t-test за разлика на средни: Независими извадки
t.test(x, y, alternative="less") // Welch Two Sample ... mean of x ...
* t-test при зависими извадки. t-test за разлика на средни зависими извадки
t.test(x, y, alternative="greater", paired=T) // Paired t-test ... mean of the differences

* z-тест за разлики на пропорции
prop.test(c(71,58), c(220,210), correct=F)
// 2-sample test for equality of proportions.. prop1 prop2

20. Хи-квадрат тестове за съгласуваност
chisq.test(c(10,20,5), p=c(0.3, 0.6, 0.1))
// Chi-squared test for given probabilities

21 Хи-квадрат тест за независимост
chisq.test(table) // Pearson's Chi-squared test

21. Линейни модели
$y = \beta_0 + \beta_1 X + \varepsilon$, β_0 и β_1 са константи; lm(y~x)
X- предиктор, Y- отклик (response)
Проверка на хипотеза за β_1 - $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$

22. Линеен модел с няколко предиктора
$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n + \varepsilon$; $\beta_1, \ldots, \beta_n$ са константи
(lm(y ~ x_1 + x_2)