Bivariate data

Но преди това задача от предния път

3.6 The number of O-ring failures for the first 23 flights of the US space shuttle Challenger were

```
0 1 0 NA 0 0 0 0 0 1 1 1 0 0 3 0 0 0 0 0 2 0 1
```

(NA means not available – the equipment was lost). Make a table of the possible categories. Try to find the mean. (You might need to try `mean(x,na.rm=TRUE)` to avoid the value `NA`, or look at `x[!is.na(x)]`.)

https://cran.r-project.org/doc/contrib/Verzani-SimpleR.pdf

23 страница

Bivariate data

Получаваме, когато сме наблюдавали две характеристики (данните може да са от един тип или от два различни типа).
Тогава е интересна и връзката между тях – Възникват въпроси като: зависят ли помежду си? Линейна ли е зависимостта?

Категорни данни

table – използвахме я за едномерни (univariate) данни, но се използва и при двумерни (bivariate) данни.
Пример: зависимост между пушене и часове учене.

Person	Smokes	amount of Studying
1	Y	less than 5 hours
2	N	5 - 10 hours
3	N	5 - 10 hours
4	Y	more than 10 hours
5	N	more than 10 hours
6	Y	less than 5 hours
7	Y	5 - 10 hours
8	Y	less than 5 hours
9	N	more than 5 hours
10	Y	5 - 10 hours

Категорни данни

- таблица
  - маргинални разпределения
- barplot
  - stacked
  - beside

```
smokes = c("Y","N","N","Y","N","Y","Y","Y","N","Y")
amount = c(1,2,2,3,3,1,2,1,3,2)
table(smokes,amount)
```

Ако сега искаме да разгледаме:
каква част от пушачите учат по... часа?
каква част от учещите по ... часа са пушачи?
Наричат се маргинални разпределения.

Маргинални разпределения

Честотна таблица: **prop.table(от table обект)** – всяка клетка е стойността на клетката от table/бройката от всички.

$g(x) = \Sigma_y f(x,y)$

$h(y) = \Sigma_x f(x,y)$

Т.е. сборът на ред/стълб

Ползваме го, когато ни интересува само едната променлива.

Условни разпределения

Условни разпределения: сумата от пропорциите по ред/стълб е 1
prop.table(от table обект, 1) - всяка клетка е стойността на клетката от table/бройката от тези на реда.
Все едно задраскваме другите редове. Гледаме само този, който ни интересува.

prop.table(от table обект, 2) - всяка клетка е стойността на клетката от table/бройката от тези на стълба.
Все едно задраскваме другите стълбове. Гледаме само този, който ни интересува.

Закръгляне на числата

options – глобални опции

options(“digits”) ← getter # или getOption('digits')
options(digits=3) ← setter
old.digits = options("digits") # запазваме преди колко е било
options(digits=3) # променяме го

... работим ...

options(digits=old.digits) # връщаме го

Barplot

barplot(таблица)

Взима колоните на таблицата да са имената на стълбовете в диаграмата. После всеки ред го наслагва с различен цвят. (това се нарича stacked barplot)

- с ключовата дума beside му казваме да не наслагва стълбовете, а да ги слага един до друг. (т.е. не stacked)
- с legent.text = TRUE да показва легендата. (взима данните от factor обекта)

```
barplot(table(amount,smokes),main="table(amount,smokes)",
beside=TRUE,
legend.text=c("less than 5","5-10","more than 10"))
```

Apply

apply – прилага функция към всеки ред или всяка колона на матрица/таблица.

apply(структурата от данни, 1 за редове; 2 за колони, функция, която приема ред/стълб)

така
mult2 = function(x) x*2
apply(x, 1, mult2)

prop = function(x) x/sum(x)

prop.table(x, 2) е като
apply(x, 2, prop)

prop.table(x, 1) е като
apply(x, 1, prop)

t(x) - транспонира матрицата x

Две количествени

- 2 boxplot-а един до друг
- линейна регресия
- stripchart
- violin chart
- коефициент на корелация
  - Pearson
  - Spearman
- Scatterplot
  - plot
  - curve

Количествени характеристики

```
x = c(5, 5, 5, 13, 7, 11, 11, 9, 8, 9) # резултати на експериментална група 1
y = c(11, 8, 4, 5, 9, 5, 10, 5, 4, 10) # резултати на експериментална група 2
boxplot(x,y)
```

Същите данни, но...

Ако първоначално сме имали всички резултати на едно място, но имаме и данни кои резултати към коя група са:

```
 amount = scan()
# 5 5 5 13 7 11 11 9 8 9 11 8 4 5 9 5 10 5 4 10
category = scan()
 # 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2
boxplot(amount ~ category)
```

Така имаме **количествена и качествена** променлива.

Формули

В статистиката моделът е начин да опишем данни, а това става с формула.
В R формули се създават с ~
y = x1 + x2 + ... + xn съответства на
y ~ x1 + x2 + ... + xn

Сега да проверим функцията boxplot и какви са ѝ аргументите:

boxplot

formula - a formula, such as y ~ grp, where y is a numeric vector of data values to be split into groups according to the grouping variable grp (usually a factor).

boxplot(y ~ x) разделя стойностите на y в групи в зависимост от стойностите на x и прави толкова кутии с мустаци (boxplots), колкото различни стойности заема x.

Stripchart

Stripchart - 1d графика, която нанася измерванията, които сме получили върху числовата ос. Добра алтернатива на boxplot е, когато имаме малко измервания. Подаваме му dataframe или list.

- Една променлива
- Няколко stripcharts

stripchart(new)

stripchart(new, method = “jitter”)

Scale

scale – функция, която скалира (центрира и нормира данните - средно 0 и стандартно отклонение 1)

scale(x) – Първо центрира, после дели на стандартното отклонение.

Центриране: ако x е вектор, намира средното на x и после от всеки елемент на x вади средното.

Violin plot

Да се разгледат заедно две разпределения. Начертани са вертикално разпределенията и огледалните им образи.

simple.violinplot(old, new)

simple.violinplot(scale(old),scale(new))

Plot

plot(x, y)

abline(lm(y~x))

- plot(x, y)
- lm(y~x) намира линеен модел по данните (т.е. коефициентите a и b)
- abline(коефициенти) чертае права линия, като приеме коефициенти, като трябва да има съществуваща вече графика, защото чертае отгоре.

Тези две стъпки са еквивалентни на следната: simple.lm(x, y)

Коефициентите може да ги достъпим по два начина:

```
model = simple.lm(x, y)
coef(model) # нещо като официален getter
model$coef # както казахме, че се достъпват атрибути на обекти
```

Корелация

- Коефициент на Пиърсън – проверка за линейна корелация, коефициентът е [-1, 1]

cor(x, y)

$$R = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - X)^2 \sum (Y_i - Y)^2}}$$

Ковариация и корелация

$$\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$$

$$\rho_{XY} = E[(X - \mu_X)(Y - \mu_Y)]/(\sigma_X \sigma_Y)$$

Корелация на Спиърман

Това е корелация между ранговете на данните.

rank(x) за всяко наблюдение показва ранга му (на кое място се намира във вариационния ред)
На равните се взима средно-аритметичното.

Спиърман корелацията се дефинира по следния начин: cor(rank(x),rank(y))
Не следи за линейна връзка, а просто дали у расте или намалява с нарастване по х.

identify

Как да идентифицираме точките на графиката?
функцията identify намира индекса на най-близкото измерване, до което сме кликнали. Работи за scatter plot
identify(x, y, n)
n – брой точки, които да идентифицираме
x, y – данните за scatter plot.
Забележка: Трябва да имаме вече съществуваща графика!

Вече можем да идентифицираме силно отличаващи се наблюдения (outliers) и да ги премахнем.

curve – чертае крива

curve(expr, from = NULL, to = NULL, n = 101, add = FALSE,

type = "l", xname = "x", xlab = xname, ylab = NULL,

log = NULL, xlim = NULL, ...)

expr име на функция на x, или израз, който съдържа x. Така резултатът от извикването на този израз с x, са стойностите, които се чертаят на y-координатата. Т.е. чертае f(x) = y, като му казваме from =... ,to =..., с keyword arguments, които по подразбиране са от 0 до 1.
add – ако е TRUE, добавя към съществуваща графика.

Еквивалентни изрази

```
z = seq(0, 4, by=.1)
plot(z, z^2, type="l")

curve(x^2, 0, 4)
curve(x ^2)
```