Данни

и какво можем да
правим с тях в R?

Извадка и популация

- Популация - всички данни изобщо, които са от наш интерес.
- Извадка - измерванията, които сме направили.

Дескриптивна и инференчна статистика

- Дескриптивна (описателна) статистика: да пресметнем характеристики (средно, медиана, дисперсия) за **извадката**, която сме събрали (или за цялата популация, ако имаме данни за нея).
- Инференчна статистика: да направим изводи (тестване на хипотези, построяване на доверителни интервали) за параметрите на цялата **популация**.

Данни

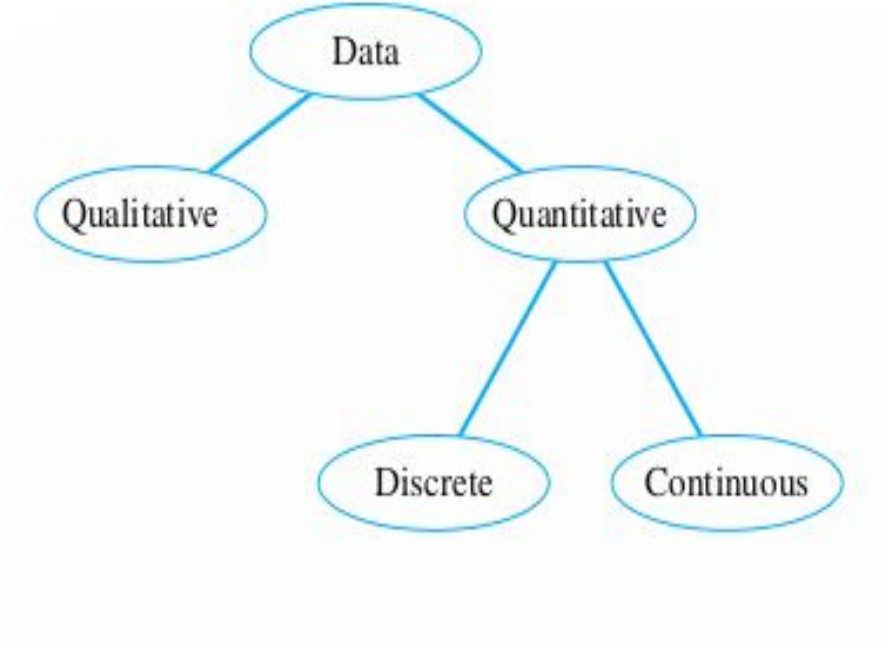
- Променлива
- Константа
- Univariate data – данните, които сме получили, когато сме наблюдавали една променлива.
- Bivariate data – когато сме наблюдавали две променливи.
- Multivariate data – множество променливи.

Класифициране на данните/променливите

- количествените измерват количество
- качествените измерват качество...

Още: качествени = категорни

Височината на човек: може да ги опишем с думи– висок, среден, нисък – така са качествени данни (категорни). Но можем и с числа. (така са количествени данни)


Data
Qualitative
Quantitative
Discrete
Continuous

Задача: Определете типовете данни:

1. Вратата, която избира мишката при експеримент – A, B, C.

2. Времето, за което Юсейн Болт бяга 100м.

3. Броят на студентите във втори курс.

За различните видове данни съществуват различни начини да ги опишем – можем да пресметнем различни описателни статистики и да построим различни графики.

Категорни данни

- таблици
- bar chart (правоъгълна/стълбчеста диаграма)
- pie chart (кръгова диаграма)

Таблици

- честота
- относителна честота

```
x=c("Yes","No","No","Yes","Yes")
table(x)
```

Factors

- удобни за категорни данни
- вектор от категорни данни
- изброима колекция, която отзад се представя с числа (като ENUM в c++)
- Различните стойности, които категорните данни могат да заемат, се наричат нива - levels

```
x=c("Yes","No","No","Yes","Yes")
factor(x)
```

```
chocolate = scan()

# 3 1 1 3 3 3 1 3 2 1 2 1 2 3 2 3 1 1 1 1 3 1 2

chocolateTable = table(chocolate)

names(chocolateTable) = c("white","black","milk") # Дава имена на колоните

barplot(chocolateTable)

# Това не е вярно!

barplot(chocolate)

barplot(chocolateTable/length(chocolate)) # Относителни честоти

pie(chocolateTable,col=c("white","black","brown")) # rainbow
```

Количествени данни

- ако приема изброимо много стойности – дискретна
- ако приема неизброимо – непрекъсната

Да си припомним:

mean, var, sd, median

Нови:

fivenum - # min, lower hinge, Median, upper hinge, max

summary # - Min. 1st Qu. Median Mean 3rd Qu. Max.

Квантили

quantile(данни, вектор с всички квантили, които искаме | един квантил)

quantile(data,.25)

quantile(data,c(.25,.75))

Средното и стандартното отклонение “страдат”, когато има outlier-и, не дават вярна представа за данните. Може да се предпазим със следните робастни (устойчиви откъм силно отличаващи се наблюдения) статистики:

- Медиана
- Може да се ползва “изрязано” средно, trimmed mean, където първо отрязваме определен процент от данните и после намираме средното. mean(sals,trim=1/10) # trim 1/10 off top and bottom.
- Median Average Deviation – MAD

$$\text{median}|X_i - \text{median}(X)|(1.4826)$$

Stem-and-leaf chart - диаграма клони с листа

```
scores = scan()
# 2 3 16 23 14 12 4 13 2 0 0 0 6 28 31 14 4 8 2 5
stem(scores)
```

Числови ↠ Категорни

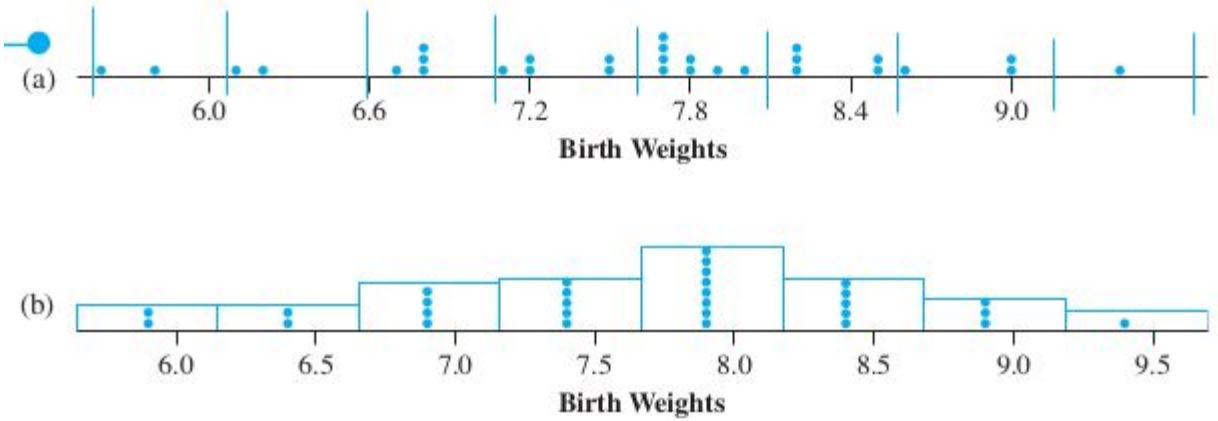
12 .4 5 2 50 8 3 1 4 .25
[0, 1],(1, 5],(5, 50]
breaks – казваме къде да дели! ,breaks=c(min(sals),1,5, max(sals))
x = scan()
cut(x, breaks = c(min(x), 1, 5, max(x))) – разделя на интервали и за всяко наблюдение казва в кой от интервалите е то. Ако няма такъв интервал, връща <NA>.
cutted = cut(x, breaks = c(0, 1, 5, 50))
table(cutted) - след това table преброява всеки интервал по колко пъти се среща.
levels(cutted) = c("poor","rich","rolling in it") # Дава имена на нивата!!!

Хистограма

Barplot vs Histogram

Birth Weights of 30 Full-Term Newborn Babies

7.2	7.8	6.8	6.2	8.2
8.0	8.2	5.6	8.6	7.1
8.2	7.7	7.5	7.2	7.7
5.8	6.8	6.8	8.5	7.5
6.1	7.9	9.4	9.0	7.8
8.5	9.0	7.7	6.7	7.7



lines.density - изгладен вариант на хистограмата, лицето под кривата е 1.

хистограмата с параметър prob=T - лицето на правоъгълника в интервала е вероятността случайно избрано наблюдение да е в този интервал. Сумата от лицата на всички правоъгълници е 1.

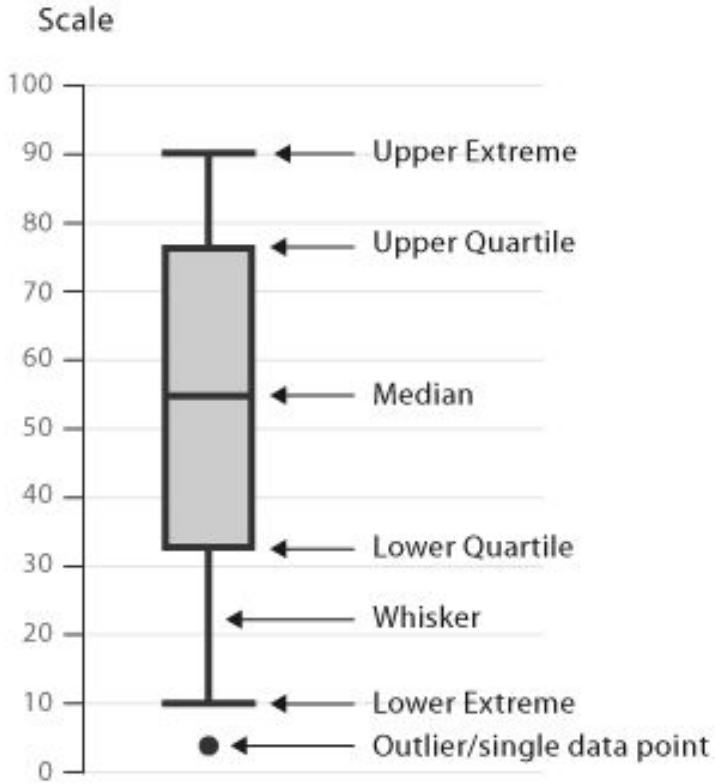
След като вече съществува хистограма, може да добавим следните:

- плътност на хистограмата - lines(density(faithful$eruptions))
- да отбележим къде са измерванията - rug(jitter(x))
  - rug - keyword side се указва дали да ги чертае долу, или горе.
  - jitter(r) → тогава r = r + runif(n, -a, a)

Frequency polygon

```
x = c(.314,.289,.282,.279,.275,.267,.266,.265,.256,.250,.249,.211,.161)

tmp = hist(x) # store the results

lines(c(min(tmp$breaks), tmp$mids, max(tmp$breaks)), c(0, tmp$counts,0),
type="l")

# install.packages("UsingR")

library(UsingR)

    simple.freqpoly(x)
```

Boxplot


Scale
100
90
80
70
60
50
40
30
20
10
0
Upper Extreme
Upper Quartile
Median
Lower Quartile
Whisker
Lower Extreme
Outlier/single data point

Packages & Sets

To list all available packages Use the command library().
To list all available datasets Use the command data() without any arguments
To list all data sets in a given package Use data(package=’package name’) for example data(package=ts)