Scientific programming in R – pandemic edition

Dr. Henry Löffler-Wirth
Maria Schmidt

Interdisciplinary Center for Bioinformatics - IZBI,
University Leipzig
Descriptive statistics: discrete data

Discrete variables realize a defined set of values

Example: a random genome – repeated manifestation of one of the 4 bases "A", "C", "G" or "T"

```r
> rand.gen <- sample( c( "A", "C", "G", "T" ), 50, replace=TRUE )
> rand.gen
[1] "G" "G" "C" "C" "C" "G" "A" "A" "A" "T" "C" "G" "C" "C" "A" "G" "C" "G" "C" "C" "C" "C" "C"
> table( rand.gen )
rand.gen
 A  C  G  T
11 18 10 11
> prop.table( table( rand.gen ) )
rand.gen
 A  C  G  T
0.22 0.36 0.20 0.22
```

*sample* *(set, n, replace)* randomly draws n values from the given set with/without replacement.

... and corresponding probability distribution (%)
Descriptive statistics: discrete data

Discrete variables realize a defined set of values

Basic visualization of the frequencies can be generated using, e.g. `pie()` and `barplot()` functions:

```r
> pie( table( rand.gen ) )
```

pie chart of frequencies

```r
> barplot( table( rand.gen ) )
```

bar chart of frequencies
Descriptive statistics: continuous data

Continuous variables realize a virtually infinite set of values (e.g. distances, size, genomic activity)

Probability tables are inappropriate, here we utilize localization and variability measures:

```r
> data(trees)
```

Loads a built-in data frame on trees.

```r
> mean( trees$Height )
[1] 76
```

Calculates the average value (arithmetic mean):

\[
\bar{x}_{\text{arithm}} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}
\]

Quantiles: x %-quantile divides data, such that x % of data points are below the quantile and 100-x % above

```r
> median( trees$Height )
[1] 76
```

median = 50% quantile

mean = 2.3
median = 1

Median value is more robust than mean:
The outlier value influences mean dramatically, but not median.
Descriptive statistics: continuous data

Continuous variables realize a virtually infinite set of values (e.g. distances, size, genomic activity)

Probability tables are inappropriate, here we utilize localization and variability measures:

```r
> quantile(trees$Height, 0.25)
25%    72
> quantile(trees$Height, 0.75)
75%    80
> summary(trees$Height)
     Min.  1st Qu.   Median      Mean  3rd Qu.     Max.
     63.0   72.0   76.0     76.0   80.0   87.0
```

First & third quartile are the 25% & 75% quantiles; `summary()` shows extrema, mean, median and quartiles
Descriptive statistics: continuous data

Continuous variables realize a virtually infinite set of values (e.g. distances, size, genomic activity)

Probability tables are inappropriate, here we utilize localization and variability measures:

```r
> var( trees$Height )
[1] 40.6
```

Calculates the variance (i.e. quadratic deviation from mean)

\[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \]

```r
> sd( trees$Height )
[1] 6.371813
> sqrt( var( trees$Height ) )
[1] 6.371813
```

Calculates the standard deviation (i.e. square root of variance): \( s = \sqrt{s^2} \)

```r
> range( trees$Height )
[1] 63 87
```

Data range (minimum and maximum)
Descriptive statistics: continuous data

A closer look to boxplots

> boxplot(trees$Height)

\[ \text{IQR} = 3\text{rd} - 1\text{st quartile} \]
Descriptive statistics: continuous data

Histogram: frequency distribution of data binned into homogeneous intervals

> hist( trees$Height )
Descriptive statistics: continuous data

Scatterplot showing two (equally long) vectors in 2-dimensional coordinate system

```r
> plot( trees$Height, trees$Volume )
```

Simple plotting: `plot( x-coordinates, y-coordinates )`
Inductive statistics: a very short view

Very common question: Is there difference between group A and group B?

→ Statistical testing .... but: how can we quantify a difference?

• fold change (FC): ratio of mean values of two groups

• statistical scores (e.g. t-statistic): combines FC with reliability adjustment

• p-value [0, 1]: probability to obtain a given score by chance → small p-values are good!

• q-value (aka fdr – false discovery rate) [0, 1]: probability to find at least one false positive
Inductive statistics: a very short view

Implementations for two-group comparisons in R:

- **Wilcoxon signed-rank test**
  - no normal distribution
  - paired
  - \[ \text{wilcoxon.test}(x, y, \text{paired}=\text{TRUE}) \]

- **Dependent t-test**
  - normal distribution
  - \[ \text{t.test}(x, y, \text{paired}=\text{TRUE}, \text{var.equal}=\text{TRUE}) \]

- **Wilcoxon rank-sum test**
  - no normal distribution
  - unpaired
  - \[ \text{wilcoxon.test}(x, y, \text{paired}=\text{FALSE}) \]

- **Welch test**
  - differing variances
  - \[ \text{t.test}(x, y, \text{paired}=\text{FALSE}, \text{var.equal}=\text{FALSE}) \]

- **Independent t-test**
  - normal distribution
  - equal variances
  - \[ \text{t.test}(x, y, \text{paired}=\text{FALSE}, \text{var.equal}=\text{TRUE}) \]

- **x** and **y** are vectors containing the values to test, e.g. expression values of a gene

- length of **x** and **y** may differ (un-paired test only)

- tests return complex objects: see \texttt{?t.test} and \texttt{?wilcoxon.test} for details
Programming task V

Organization for economic cooperation and development (OECD) provided data of member countries:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mode</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td>factor</td>
<td>30 countries with available data</td>
</tr>
<tr>
<td>Continent</td>
<td>factor</td>
<td>continent of the country</td>
</tr>
<tr>
<td>Europe</td>
<td>logical</td>
<td>does the country belong to Europe</td>
</tr>
<tr>
<td>Income</td>
<td>numeric</td>
<td>average income in $</td>
</tr>
<tr>
<td>Poverty</td>
<td>numeric</td>
<td>percent children living in poor families</td>
</tr>
<tr>
<td>LackOfEducation</td>
<td>numeric</td>
<td>percent children without access to (books, computer, ...)</td>
</tr>
<tr>
<td>LackOfRoom</td>
<td>numeric</td>
<td>percent children living in very small flats</td>
</tr>
<tr>
<td>Pollution</td>
<td>numeric</td>
<td>percent children suffering from pollution</td>
</tr>
<tr>
<td>Reading</td>
<td>numeric</td>
<td>average PISA reading score</td>
</tr>
<tr>
<td>LowBirthWeight</td>
<td>numeric</td>
<td>percent children with less than 2.5kg at birth</td>
</tr>
<tr>
<td>InfantMortality</td>
<td>numeric</td>
<td>mortality of babies (&lt;1 year), x in 1,000</td>
</tr>
<tr>
<td>TeenMortality</td>
<td>numeric</td>
<td>mortality of children (1-19 years), x in 100,000</td>
</tr>
<tr>
<td>Suicides</td>
<td>numeric</td>
<td>adolescent suicides (15-19 years), x in 100,000</td>
</tr>
<tr>
<td>Exercise</td>
<td>numeric</td>
<td>percent adolescents (11-15 years) with regular phys. ex.</td>
</tr>
<tr>
<td>Smoking</td>
<td>numeric</td>
<td>percent adolescents (15 years) smoking at least one a week</td>
</tr>
<tr>
<td>Alcohol</td>
<td>numeric</td>
<td>percent adolescents (13-15 y) that have been drunk twice</td>
</tr>
<tr>
<td>Bullying</td>
<td>numeric</td>
<td>percent children threatened at school</td>
</tr>
<tr>
<td>LikeSchool</td>
<td>numeric</td>
<td>percent children who like going to school</td>
</tr>
</tbody>
</table>
Programming task V

Download the data: > load( url("http://izbifs.izbi.uni-leipzig.de/~wirth/oecd.RData") )

Solve the following questions / tasks using descriptive and inductive statistics, and basic visualization:

a) Which country has highest adolescent alcohol experience. What is this percentage?

b) Which country has lowest infant mortality? How high is it? Visualize infant mortality over all countries.

c) In which countries have adolescents less physical exercise than average?

d) In which countries suffer particularly many children from bullying? "many" is defined here as more than one standard deviation above median bullying percentage.

e) How frequent are continents represented? Visualize it!

f) Generate scatterplots to investigate following relations: Does lack of basic education influence PISA reading score? Do poverty and teen mortality associate?

g) Generate boxplots of variable LackOfEducation in Europe and Non-Europe. Is there a significant difference? Hint: Use \texttt{wilcox.test()} and corresponding help page.