Biometry. Lecture 11

Alexey Shipunov

Minot State University

March 2, 2015

Outline

Starting...

`> setwd("<working folder>")`

or

"`Change dir`"

in menu!

On Mac, be sure that startup option is working: `getwd()`
(`getwd()` checks if R is in working folder, `dir()` checks the folder content)

One-dimensional data

Central tendency

How to calculate means for all columns

```
> sapply(trees, mean)
```

Commands of `*apply()` family (`sapply()`, `apply()`, `lapply()`, `mapply`, `tapply()`) are most powerful in R

One-dimensional data

Range

Standard deviation, variance and IQR

- Variance is a sum of square differences between each value and mean divided by number of degrees of freedom (so-called “Bessel’s correction”)
- Standard deviation is a square root from variance
- IQR (inter-quartile range) is simply a difference between fourth and second quartiles. It is more robust than standard deviation.

Standard deviation, variation and IQR

```
> sd(salary); var(salary); IQR(salary)
```

Coefficient of variation

Coefficient of variation (CV) is a standardized (by mean) standard deviation

```
> cv.trees <- 100*sapply(trees, sd)/colMeans(trees)
> cv.trees
```

“Volume” variable variates most.

Boxplots

Boxplots (invented by John Tukey) are one of the best representations of data central tendency and range.

```
> boxplot(salary)
> boxplot(trees)
> eq <- read.table("http://ashipunov.info/data/eq.txt",
+ h=TRUE)
> boxplot(DL.R ~ SPECIES, data=eq)
```

Boxplots do not show mean and standard deviation.

Stacked boxplots

```
> seq <- scale(eq[,2:9]) # scale all except SPECIES
> boxplot(seq[eq$SPECIES=="arvense",],
+ at=1:8-0.2, boxwex=0.3, # shift and narrow
+ xaxt="n", yaxt="n", # no labels
+ main="Horsetails, scaled and paired characters")
> boxplot(seq[eq$SPECIES=="fluviatile",],
+ at=1:8+0.2, boxwex=0.3,
+ xaxt="n", yaxt="n",
+ add=T, col="green") # overlay and colorize
> axis(1, 1:8, names(eq)[2:9], cex.axis=.75) # labels
> legend("topleft", c("arvense","fluviatile"),
+ fill=c("white","green"))
```

This stuff is better to remember in separate *.r file

Histograms

Histograms show the frequency of every data interval:

```
> hist(salary)
> hist(trees[,1])
```

Density plots

Density plot smooths the histogram:

```
> plot(density(trees[,3]))
> plot(density(rnorm(1000))) # 1000000 is even better!
```

Density plots looks prettier but may lead to wrong colnclusions especially if sample is small.

summary()

summary() is a “smart” (generic) function which gives the most appropriate description of data. In many cases, it will give quantiles + mean:

```
> summary(salary)
> summary(trees)
> summary(sex)
```

`summary()` and “bad” data

summary() is very useful when one needs to check a reliability of data:

```
> err <- read.table("http://ashipunov.info/data/errors.txt",
+ h=TRUE, sep="\t")
> str(err)
> summary(err)
```

`AGE` became a factor (erroneous “a”), empty name (instead of `NA`), and impossible minimal height.

One-dimensional data

One-dimensional tests

t-test and Wilcoxon test for one-dimensional data

- Statistical tests allow to check how well the general characteristic (central tendency or range) calculated from *sample* represents a *population*
- t-test (Student's) takes into account the normality of sample whereas Wilcoxon test do not consider the distribution, it is non-parametric
- Both give a *confidence interval*

Finishing...

Save your commands!

(`savehistory(<todaysdate>.r)` or `File -> Save as...` on Mac)

Summary: most important commands

- `IQR()`—returns robust range
- `boxplot()`— draws a Tukey's boxplot

For Further Reading

A. Shipunov.
Biometry [Electronic resource].
2012—onwards.
Mode of access:
`http://ashipunov.info/shipunov/school/biol_240`

A. Shipunov, and others.
Visual statistics. Use R!
Ongoing translation from Russian.