Biometry. Lecture 12

Alexey Shipunov

Minot State University

February 22, 2012

Outline

Outline

Starting...

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

or

“Change dir”

in menu!

Previous final question: the answer

What is a main practical difference between mean and median?

Previous final question: the answer

What is a main practical difference between mean and median?

- Median is more robust

Lab 5: conclusion

There were two variants:

- Mean and confidence interval
- Median and IQR—this is always preferable!

Therefore, the best answer could be: “If character1 is from median$-$IQR to median$+$IQR...”, or even simpler: “median$\pm$IQR”.

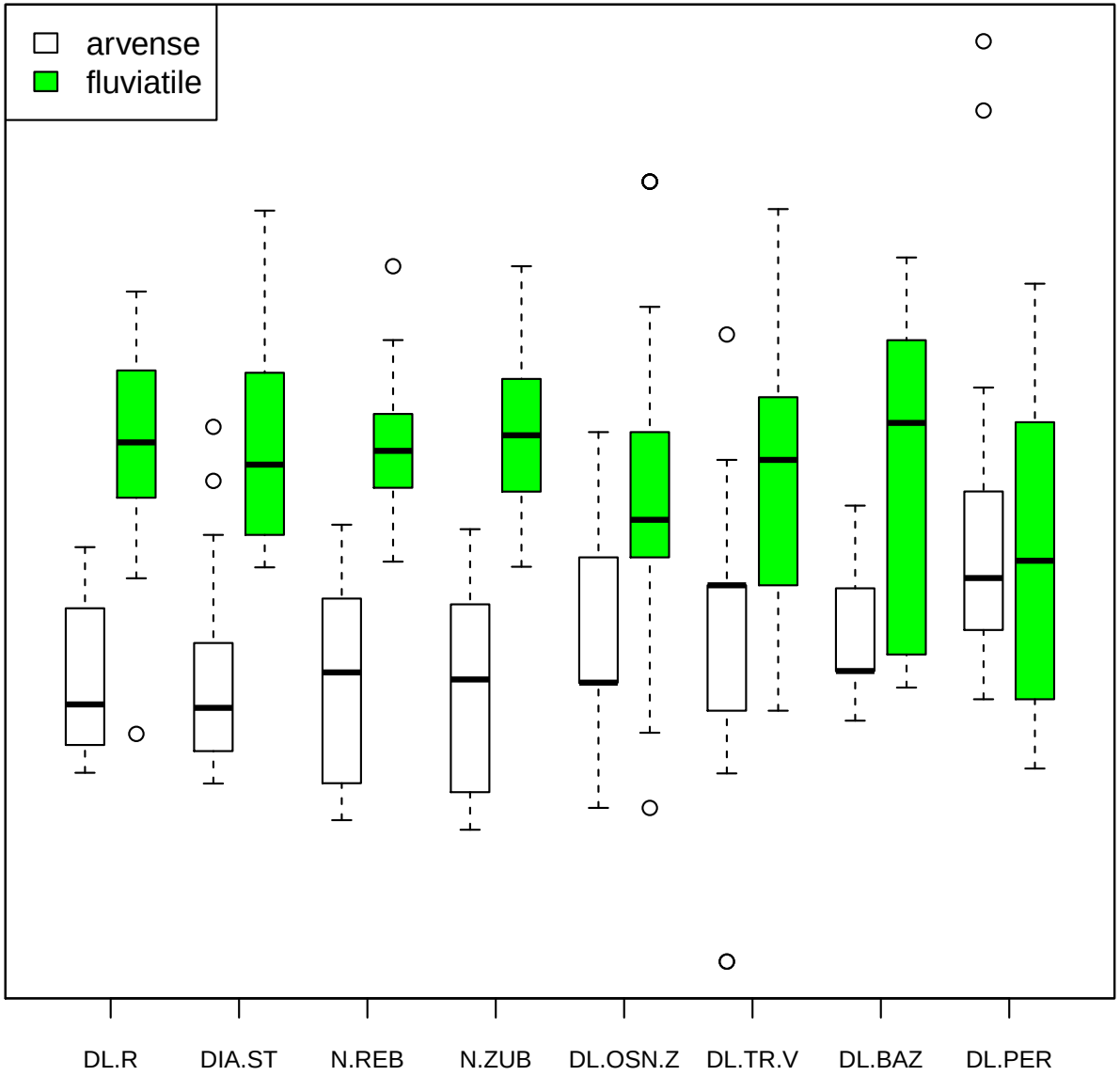
Comparative boxplot

```
> eq <- read.table("http://ashipunov.info/data/eq.txt",
+ h=TRUE)
> 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

Minot State UNIVERSITY

Horsetails, scaled and paired characters



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) tales into account the normality of sample whereas Wilcoxon test do not consider the distribution, it is non-parametric
- Both give a *confidence interval*

Minot State UNIVERSITY

t-test for one variable

```
> salary <- c(21, 19, 27, 11, 102, 25, 21)
> t.test(salary, mu=mean(salary))
One Sample t-test

data:  salary
t = 0, df = 6, p-value = 1
alternative hypothesis: true mean is not equal to 32.28571
95 percent confidence interval:
  3.468127 61.103302
sample estimates:
mean of x
 32.28571
```

Understanding the test output: theory

- Alternative hypothesis (“something”) and null hypothesis (“nothing”)
- Type I error (false alarm), p-value (probability to issue the false alarm) and significance level (matter of agreement)

Understanding the test output: quick and dirty

- Which hypothesis is null?
- Does p-value less than 0.05?
  1. No: accept the null hypothesis
  2. Yes: reject the null hypothesis

Wilcoxon test for one variable

```
> wilcox.test(salary, mu=median(salary), conf.int=TRUE)

Wilcoxon signed rank test with continuity correction

data:  salary
V = 10, p-value = 0.5896
alternative hypothesis: true location is not equal to 21
80 percent confidence interval:
 17.99999 63.50002
sample estimates:
(pseudo)median
      24.99994
```

This will test median, not mean! Wilcoxon test is more universal but less traditional.

How to understand which test to use? Normality.

- Normality tests will check if we can accept the normal distribution of our sample
- It is widely accepted that the strict normality testing is not generally required, it is enough, for example, to test normality graphically

Shapiro test for normality

```
> shapiro.test(salary) # What is a null hypothesis?!
> shapiro.test(rnorm(1000)) # Null is normality!
```

Quantile-quantile plot for normality

```
> qqnorm(salary); qqline(salary) # Bad!
> # Good:
> set.seed(1); qqnorm(rnorm(100)); qqline(rnorm(100))
```

`set.seed()` helps to maintain the same set of random numbers in the session.

One-dimensional data

Normality and R functions

- `shapiro.test()` is good but it is hard to apply if for data frames, and output is not very helpful.
- We will create the user function which run Shapiro-Wilks test with better output.

What is function

```
> Sum <- function(a, b)
+ {
+ a + b
+ }
> Sum(2, 3)
```

Function is a piece of code which may run independently. All R commands are functions. Please note that every functions requires two parts: **arguments** in *round brackets* and **body** in *curly brackets*. It is too boring to enter functions line-by-line. Instead, it is better to copy function from external editor. If function contains mistake(s), one may use `fix()` command.

Normality function, version 1

```
> Normality <- function(data.f)
+ {
+ result <- data.frame(var=names(data.f),
+ p.value=rep(0, ncol(data.f)),
+ normality=is.numeric(names(data.f)))
+ for (i in 1:ncol(data.f))
+     {
+     data.sh <- shapiro.test(data.f[, i])$p.value
+     result[i, 2] <- round(data.sh, 5)
+     result[i, 3] <- (data.sh > .05)
+     }
+ return(result)
+ }
Normality(trees)
```

Complicated stuff! Please listen to explanations carefully.

Minot State UNIVERSITY

Finishing...

```
> savehistory("20120222.r")
```

Final question (3 points)

Final question (3 points)

Please explain the difference between null and alternative hypotheses.

Summary: most important commands

- `shapiro.test()`—Shapiro-Wilks test for normality
- `function(){}`—creates a function

For Further Reading

A. Shipunov.
Biometry [Electronic resource].
2012—onwards.
Mode of access: `http://ashipunov.info/shipunov/school/biol_299`

P. Dalgaard
Introductory Statistics with R. 2nd edition.
Springer, 2008.
Chapters 4, 5:1–2.

Minot State UNIVERSITY