Biometry. Lecture 9

Alexey Shipunov

Minot State University

March 3, 2014

Outline

1. Questions and answers

2. Inside R
   - Data frames (tables)

3. One-dimensional data
   - Central tendency
   - Range

Outline

1. Questions and answers

2. Inside R
   - Data frames (tables)

3. One-dimensional data
   - Central tendency
   - Range

Outline

1. Questions and answers

2. Inside R
   - Data frames (tables)

3. One-dimensional data
   - Central tendency
   - Range

Starting...

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

or

“`Change dir`”

in menu!

Previous final question: the answer

How to select from data frame `eq` column which name is `NUM.Z`?

Previous final question: the answer

How to select from data frame `eq` column which name is `NUM.Z`?

- `eq[,"NUM.Z"]`
- `eq$NUM.Z`

Inside R

Data frames (tables)

My last example

```
> b <- 1:8 # vector
> dim(b) <- c(4,2) # two columns, four rows
> b <- data.frame(b) # convert to data frame
> b[2, 2] <- "string" # replace one number with characters
> b[!is.na(as.numeric(b[,2])),] # remove all strings with chars
```

Selection by condition

```
> d[d$sex=="f",] # will select only women
> d[d$sex!="f",] # will select all other genders ;)
```

`==` is “equal?”, `&` “and”, `|` “or” and `!` is “not”

Sorting and ordering

```
> sort(x) # ascending
> rev(sort(x)) # descending
> d[order(d$sex, d$height), ] # sort by sex then by height
```

One-dimensional data

Central tendency

Mean and median

- These are two most frequently used characteristics of the central tendency.
- Median is more robust than mean.

Mean and median

```
> salary <- c(21, 19, 27, 11, 102, 25, 21)
> mean(salary); median(salary)
> median(1:3); median(1:4)
```

When number of elements is odd, median is a central value; if even—median is the average between two centrals.

Median is the third quartile

Quartiles take out 0% (minimum, `min()`), 25% (lower hinge) , 50%, 75% (upper hinge) and 100% (maximum, `max()`) of ordered data. Median is simply a 50% (third) quartile.

```
> fivenum(salary)
```

Mode

Mode is the most frequent value:

```
> sex <- c("m", "f", "m", "m", "f", "m", "m")
> t.sex <- table(sex)
> mode <- t.sex[which.max(t.sex)]
> mode
```

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)
```

Boxplots do not show mean and standard deviaiton.

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)
```

Finishing...

```
>savehistory("20140303.r")
```

Final question (2 points)

Final question (2 points)

What is a main practical difference between mean and median?

Summary: most important commands

- `median()`—returns a median value
- `IQR()`—returns robust range
- `boxplot()`— draws a 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!
DMK Press, 2012. [Translating from Russian.]