# Biometry. Lecture 11

Alexey Shipunov

Minot State University

February 17, 2012

# Outline



Minot State UNIVERSITY

# Outline



Minot State UNIVERSITY

# 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`

# 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()`, `by()`, `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)
```

# `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="")
> 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) 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

```
> 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—sit and relax
  2. Yes: reject the null hypothesis—jump and call the police

# Finishing...

```
> savehistory("20120217.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
- `t.test()`— checks the reliability of mean (assuming that data distribution is normal)

# 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*.