Biometry. Lecture 19

Alexey Shipunov

Minot State University

March 28, 2012

Outline

Outline

Minot State UNIVERSITY

Starting...

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

or

“Change dir”

in menu!

Previous final question: the answer

In the embedded data `USArrests`, there are numbers of murders and rapes per 100,000 for every state.
Are murders and rapes correlated?

```
> with(USArrests, cor(Murder, Rape))
[1] 0.5635788
```

Two-dimensional statistics

Chi-squared test

Multiple comparisons

- If we apply multiple tests to one component (e.g., test several samples against one), probability to obtain incorrect answer will grow
- To keep the error rate low, one should apply the so-called *Bonferroni correction*, in other words—increase p-values to avoid the growing error

Seedling example

Which fungus terminates the germination?

```
> pr <- read.table(
+ "http://ashipunov.info/data/seedlings.txt", h=TRUE)
> head(pr)
> unique(pr$CID)
# correct only for one comparison!
> chisq.test(table(subset(pr, CID==c(0,105))))
> p105 <- chisq.test(table(subset(pr,
+ CID==c(0,105))))$p.value
> p63 <- chisq.test(table(subset(pr,
+ CID==c(0,63))))$p.value
> p80 <- chisq.test(table(subset(pr,
+ CID==c(0,80))))$p.value
> all.p <- c(p105, p63, p80)
> p.adjust(all.p)
```

Two-dimensional statistics

Correlation

Correlation tests

- The null hypotheses for these tests is that correlation differs from zero
- There are both parametric and non-parametric tests

Correlation tests

```
> with(trees, cor.test(Girth, Height))
> cor.test(first, second, method="spearman")
```

Two-dimensional statistics

Regression

Idea of regression

- The basic regression analysis will apply the linear model for data
- It will study not only if variables are associates and not only the strength of association but also the form of association (law of association)

Regression formula

- The simplest is a linear regression,

$$m = b_0 + b_1 \times x,$$

  where m is a **predicted value**, x is an **independent variable** and b_1 and b_2 are coefficients (so-called **intersect** and **slope**).
- In other terms, linear regression is
  `response = intersect + slope * influence`
- In R model formula language, it is simply
  `response ~ influence`

Analysis of regression model

- If y is a real response, then error of model

$$E = y - m$$

- If σ^2 are dispersions of m and y, then

$$R^2 = 1 - \sigma_m^2/\sigma_y^2,$$

- In a background, R^2 is similar to coefficient of determination

Test of regression

- To test if regression model is correct, the Fisher test is normally applied
- Null hypothesis for Fisher test is that a model is not reliable

Regression example: `women` data

```
> lm.women <- lm(weight ~ height, data = women)
> plot(weight ~ height, data = women, main="",
+ xlab="Height (feet)", ylab="Weight (pounds)")
> grid()
> abline(lm.women, col="red")
```

Analysis of regression

```
> summary(lm.women)
```

Analysis of analysis

- Resulted model: `weight = -87.51667 + 3.45 * height`
- Maximum deviations from model are -1.7333 and 3.1167 pounds
- Almost half of residuals are between first and third quartiles
- All coefficients are significant
- Adjusted R-squared is close to 1 (very high!)
- The overall p-value is much less than 0.05 therefore the model is reliable
- There are 1 and 13 degrees of freedom (for columns and for rows)

Minot State UNIVERSITY

Finishing...

```
> savehistory("20120328.r")
```

Final question (3 points)

Final question (3 points)

What is a difference between regression and correlation?

Summary: most important commands

- `cor.test()`—run correlation tests
- `lm()`—estimate the linear regression

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.
Chapter 6.