Biometry. Lecture 15

Alexey Shipunov

Minot State University

March 31, 2014

Outline

Outline

Starting...

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

or

“Change dir”

in menu!

Previous final question: the answer

What is a null hypothesis for the chi-squared test?

Previous final question: the answer

What is a null hypothesis for the chi-squared test?

- Variables are distributed independently (no association)

Two-dimensional statistics
Multiple comparisons

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)
```

Toxicity with correction

```
> tox <- read.table("http://ashipunov.info/data/tox.txt",
+ h=TRUE)
> answer <- data.frame(FOOD=NA, CHISQ.P=NA)
> for (m in 2:ncol(tox))
+ {
+ tmp <- chisq.test(tox$ILL, tox[,m])
+ answer[m-1,] <- c(names(tox)[m], tmp$p.value)
+ }
> answer[,2] <- p.adjust(answer[,2])
> answer
```

Two-dimensional statistics
Concordance and Cohen kappa

- Concordance is a measure of “agreement” between two expert answer sheets
- The most common application are psychological tests
- Cohen kappa test is frequently used for understanding the degree of concordance; the null hypothesis for Cohen kappa is that two answer sheets are non-concordant

Cohen kappa and island flora

```
> download.file("http://ashipunov.info/data/concordance.r",
+ "concordance.r")
> source("concordance.r")
> isl <- read.table(
+ "http://ashipunov.info/data/pokorm_03.dat",
+ h=TRUE, sep=";")
> str(isl); head(isl)
> cohen.kappa(as.matrix(isl))
```

Two-dimensional statistics

Correlation

Covariance and correlation

- It is always interesting to know, **how much** are two random variables change together. Covariance show that but it is not easy to interpret.
- **Correlation coefficient** is a normalized version of covariance and therefore widely used as a measure of correlation. If correlation is close to 1 or -1, it is high.
- Therefore, correlation coefficient will show the strength of relation

Features of correlation coefficient

- Correlation is a measure of **linear** relation. If relation is non-linear, correlation could be small or even zero. To check the linearity, it is recommended to make a `plot()` of two variables (scatterplot).
- Correlation may be positive or negative (from -1 to 1). If you need a sign-less measure, you may use determination coefficient $=$ correlation coefficient^2
- Correlation will only show that relation exists and has some strength, it will not show any other details about relation. For example, if correlation between A and B is high, it could mean that:
  - A depends on B
  - B depends on A
  - A and B depends on each other
  - A and B both independently depend on C and have nothing in common

Calculation of correlation coefficient

```
> cor(5:15, 7:17)
> cor(5:15, c(7:16, 23)
> cor(5:15, c(7:16, 2)
> cor(5:15, 17:7)
> cor(trees)
```

`cor()` function works with vectors or tables (matrices and data frames). If `NA`s are present, one may use option `use="complete.obs"` (better) or `use="pairwise.complete.obs"`

Non-parametric correlation

By default, cor() calculates parametric Pearson's correlation coefficient, it is possible to specify non-parametric (Spearman or Kendall) coefficients.

```
> cor(5:15, 7:17, method="spearman")
```

Visualization of correlation

```
> cor(longley)
> symnum(cor(longley))
> install.packages("ellipse")
> library(ellipse)
> plotcorr(cor(longley), type="lower")
```

Correlation tests

- The alternative 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, Volume))
> with(trees, cor.test(Girth, Height, 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_0 and b_1 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)

Finishing...

```
> savehistory("20140331.r")
```

Final question (2 points)

Final question (2 points)

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

Summary: most important commands

- `cor()`—calculates correlation coefficients
- `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_240

A. Shipunov, and others.
Visual statistics. Use R!
DMK Press, 2012. [Under translation from Russian.]