Biometry. Lecture 22

Alexey Shipunov

Minot State University

April 30, 2014

Outline

Outline

1. Questions and answers
   - Previous final question
2. Two-dimensional statistics
   - Logistic regression
   - ANalysis Of VAriation (ANOVA)

Starting...

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

or

“Change dir”

in menu!

Questions and answers

Previous final question

Previous final question: the answer

What is a logistic regression?

Previous final question: the answer

What is a logistic regression?

- Analysis of categorical (binary) response from numerical input

Two-dimensional statistics

Logistic regression

Numeric influence but categorical response

- What if response is binary?
- It is possible to convert success/failure to the **probability of success** and then apply a **generalized linear model**

Analysis of logistic regression

```
> lo <- read.table("http://ashipunov.info/data/logit.txt")
> head(lo); str(lo)
> lo.logit <- glm(formula=V2 ~ V1, family=binomial,
+ data=lo)
> summary(lo.logit)
```

Visualizing logistic regression

```
> new.points <- seq(min(lo$V1), max(lo$V1), length.out=14)
> predicted.points <- predict(lo.logit,
+ list(V1=new.points), type="response")
> success <- as.numeric(lo$V2) - 1
> plot(success ~ V1, data=lo)
> lines(new.points, predicted.points)
```

Logistic regression example II: poisoning

- Caesar or tomatoes?
- High significance of both terms could be a result of coincidence (people often took these things together)
- If we construct a logistic model and then update it (taking out one of two terms), AIC will show which model is better.

Poisoning analysis

```
> tox <- read.table("http://ashipunov.info/data/tox.txt",
+ h=TRUE)
> tox.logit <- glm(formula=I(2-ILL) ~ CAESAR + TOMATO,
+ family=binomial, data=tox)
> tox.logit2 <- update(tox.logit, . ~ . - TOMATO)
> tox.logit3 <- update(tox.logit, . ~ . - CAESAR)
> tox.logit$aic
> tox.logit2$aic # lowest!
> tox.logit3$aic
> summary(tox.logit2) # highly significant!
```

Caesar!

Two-dimensional statistics
ANalysis Of VAriation (ANOVA)

Categorical influence and numerical response

- What if there are three species in horsetail data? How to compare the diameter of stem of them all?
- Paired comparisons are “temptation” for p-value, and Bonferroni correction is sometimes of no help.
- There is a solution: analysis of variation (ANOVA) and its non-parametric twin, Kruskal-Wallis test.

ANOVA null and alternative hypotheses and assumptions

- Null is that all groups are not different, alternative is that **at least one group is different from all others**.
- All variables **should be normally distributed**. Small deviations from normality are typically accepted but Kruskal-Wallis test is preferable for all “non-normal” data.
- Actually, t-test is just an ANOVA for only two groups. Wilcoxon test, in turn, is a special case of Kruskal-Wallis test for two groups only.

Introductory example: eight horsetails

```
> eq8 <- read.table("http://ashipunov.info/data/eq8.txt",
+ h=T)
> str(eq8); head(eq8)
> plot(DIA.ST ~ SPECIES, data=eq8)
> eq8.anova <- lm(DIA.ST ~ SPECIES, data=eq8)
> anova(eq8.anova)
# If variables are not normal:
> kruskal.test(DIA.ST ~ SPECIES, data=eq8)
```

Weight, height and hair color

```
> hwc <- read.table("http://ashipunov.info/data/hwc.txt", h=T)
> str(hwc)
> boxplot(WEIGHT ~ COLOR, data=hwc)
> anova(lm(WEIGHT ~ COLOR, data=hwc))
> kruskal.test(WEIGHT ~ COLOR, data=hwc)
```

`kruskal.test()` in a non-parametric alternative for `anova()`

Post-hoc tests and graphs

```
> pairwise.t.test(hwc$HEIGHT, hwc$COLOR)
> pairwise.wilcox.test(hwc$HEIGHT, hwc$COLOR)
> w.c <- aov(lm(WEIGHT ~ COLOR, data=hwc))
> (w.c.hsd <- TukeyHSD(w.c))
> plot(w.c.hsd) # note confidence intervals, dashed line is 0
```

`pairwise.wilcox.test()` is a non-parametric post-hoc test

`TukeyHSD()` runs “Tukey Honest Significant Differences test”, one of the best post-hoc tests

Are people with different types of hair color differ also by their height?

```
> hwc <- read.table("http://ashipunov.info/data/hwc.txt",
+ h=T)
> str(hwc)
> boxplot(HEIGHT ~ COLOR, data=hwc)
> anova(lm(HEIGHT ~ COLOR, data=hwc))
> h.c <- aov(lm(HEIGHT ~ COLOR, data=hwc))
> (h.c.hsd <- TukeyHSD(h.c))
> plot(w.c.hsd)
```

Finishing...

```
> savehistory("20140430.r")
```

Final question (3 points)

Final question (3 points)

What are the null and alternative hypotheses for ANOVA?

Summary

- `glm()`—estimates the logistic regression model and many others
- `anova()`—checks analysis of variation model
- `kruskal.test()`—non-parametric alternative to ANOVA

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.]