Biometry. Lecture 13

Alexey Shipunov

Minot State University

March 24, 2014

Outline

1. Questions and answers

2. One-dimensional statistics
   - Tests for proportions

3. Two-dimensional statistics
   - Hypotheses and tests
   - Tests for the independence of two variables

Outline

1 Questions and answers

2 One-dimensional statistics

- Tests for proportions

3 Two-dimensional statistics

- Hypotheses and tests
- Tests for the independence of two variables

Outline

1. Questions and answers

2. One-dimensional statistics
   - Tests for proportions

3. Two-dimensional statistics
   - Hypotheses and tests
   - Tests for the independence of two variables

Starting...

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

or

“`Change dir`”

in menu!

Previous final question: the answer

Why we need the "`for`" construction in R?

Previous final question: the answer

Why we need the “`for`” construction in R?

- To repeat stuff

One-dimensional statistics

Tests for proportions

Why we need to test proportions

- Proportions are secondary data
- The main question is: how well the proportion calculated from sample represents the population proportion?
- Null is that proportion of sample does not differ significantly from population proportion

Smokers and non-smokers example

- In hospital, among lung cancer patients, 356 from 476 are smokers ($\approx 75\%$)
- However, among all patients this proportion is lightly lower.
- How well our sample (lung cancer group) represents the whole hospital? In other words, is the deviation we see accidental?

Exact binomial test

```
> binom.test(x=356, n=476, p=0.7, alternative="two.sided")
```

`"two.sided"` means that the deviation may be to the both possible sides. It was possible to write `"greater"` instead; in this case we would test if the proportion in our sample is bigger. One-sided tests are normally more powerful but you should **newer** use two and one-sided tests together (this is not far from falsification of results)!

Proportion test

Proportion tests are more universal than binomial, but return very similar results:

```
> prop.test(x=356, n=476, p=0.7, alternative="two.sided")
```

Voters example

In the exit poll, 262 persons were questioned. 136 ($\approx$ 53%) said they voted for the candidate A. Check if candidate A won.

```
> prop.test(x=136, n=262, p=.5, alt="greater")

1-sample proportions test with continuity correction

data:  136 out of 262, null probability 0.5
X-squared = 0.3092, df = 1, p-value = 0.2891
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.4664802 1.0000000
sample estimates:
       p
0.519084
```

Two-dimensional statistics

Hypotheses and tests

Hypotheses are cornerstones of science

- The inferential science is based on hypotheses construction and calculation of their probability.
- The simplest approach is to establish null and alternative hypotheses.

Statistical errors

- Type I error is a false alarm: we accept alternative when null is true
- Type II error is a carelessness: we accept null when alternative is true

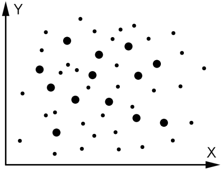
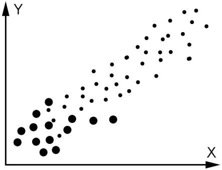
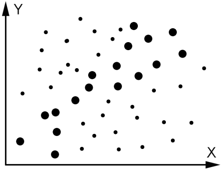
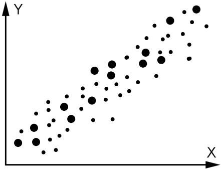
Level of significance

- The probability to make Type I error is a p-value
- We may ignore this probability if it is too low, in other words, below the level of significance
- The level of significance is a matter of experience and agreement, it could be 0.05, but sometimes also 0.1 and 0.01

Power

- Probability NOT to make a Type II error is a *power*
- The significance level for the power is normally around 0.8, tests with lesser power should be considered as weak

Type I and II errors for two variables

Population Sample	Null true	Alternative true
Accept null		
Accept alternative		

Two-dimensional statistics

Tests for the independence of two variables

What is tested?

- Null: difference equal to 0 $\approx$ similar $\approx$ related $\approx$ samples came from same population
- Alternative: difference not equal to 0 $\approx$ different $\approx$ non-related $\approx$ samples came from different populations

Tests are based on central values

```
> a <- 51:59
> b <- 1:9
> x <- rep(5, 9)
> t.test(a, b)
> t.test(b, x)
```

Homoscedasticity, similarity of variance (like in a and b but not like in b an x) is an important assumption of all two variable tests. In R, the Welch correction for **non-homogeneity of variance** is by default applied inside `t.test()`

Paired and non-paired

- Paired: came from one set of objects (e.g., measurements done at different time)
- Non-paired: do not belong to one set of objects

Tests are based on central values

```
> set.seed(1);  t.test(a,  (a+rnorm(9)), paired=T)
```

We introduced here a random noise (`rnorm()` function)

Parametric and non-parametric

- Parametric: Student’s, or t-test (in R, with Welch correction for **non-homogeneity of variance**)
- Non-parametric: Wilcoxon tests

Leaves example

```
> leaves <- read.table(
+ "http://ashipunov.info/data/leaves.txt", h=T)
> Normality3 <- function(df, p=.05)
+ {
+ sapply(df, function(.x)
+ ifelse(shapiro.test(.x)$p.value > p,
+ "NORMAL", "NOT NORMAL"))
+ }
> Normality3(leaves) # all normal!
> t.test(leaves[,1], leaves[,2], paired=T)
> wilcox.test(leaves[,1], leaves[,2], paired=T)
> t.test(leaves[,1], leaves[,3])
> wilcox.test(leaves[,1], leaves[,3])
```

“Classical” sleep data and model formula

```
> str(sleep)
> boxplot(extra ~ group, data=sleep)
> t.test(extra ~ group, data=sleep)
```

`sleep` is a data in so-called long format, `extra ~ group` is a **model formula** of `response ~ factor` form.

For t-test, “`group`” should have exactly 2 levels!

Model formula for leaves data

```
> leaves12 <- stack(leaves[,1:2])
> leaves12
> t.test(values ~ ind, data=leaves12, paired=T)
```

`stack()` converts from short to long form

Differences in leaf growth

```
> difc <- leaves[,2] - leaves[,1]
> dife <- leaves[,4] - leaves[,3]
> difce <- stack(data.frame(difc, dife))
> boxplot(values ~ ind, data=difce)
> t.test(values ~ ind, data=difce)
```

Air quality data in May and August

```
> str(airquality)
> air15 <- unstack(airquality[,c(1,5)])
> Normality3(air15)
> boxplot(Ozone ~ Month, data=airquality,
+ subset=Month %in% c(5,8))
> wilcox.test(Ozone ~ Month, data=airquality,
+ subset=Month %in% c(5,8))
```

`unstack()` converts from long to short form

`%in%` is a selection operator

Two main questions

- Normal?
- Paired?

Finishing...

```
> savehistory("20140324.r")
```

Final question (10 points!)

Final question (10 points!)

These are points from the first and second exam in one small class:
63, 72, 77, 76, 67, 56, 55, 51, 77, 64
and
87, 86, 76, 79, 54, 60, 97, 80, 73, 97
Both exams were equivalent. Provide a statistical support for the hypothesis that second exam went better. Report commands and all values which support your conclusion.

Summary: most important commands

- `binom.test()` and `prop.test()`—tests for the equality of proportions
- `t.test()`—paired and non-paired two-sample parametric test
- `wilcox.test()`—paired and non-paired two-sample non-parametric test
- `response ~ factor`—if factor has exactly two levels, this is a model formula for two-sample test
- `%in%`—selection operator

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.]