Data analysis with R

Lecture 8

Traditional testing

Jouni Junnila

Population parameters

- In data analysis we often try estimate different kind of parameters.
- Of course we would like to know the value of the population parameter (eg. population mean)
  - However we are practically always dealing with only samples and that's why we have to settle to sample parameters (eg. sample mean).
- Because we are just estimating the population parameters, we are also interested in the accuracy of our estimation.

Different statistics

- There are several different statistics we can calculate from our data.
- Often in data analysis we start with a set of descriptive statistics to give the reader a picture what kind of data we are dealing with.
- The set of statistics can include mean, median, quartiles, standard deviation, variance, standard error and the correlation coefficient.
  - Each of these can be used as an estimate of the corresponding population parameter.

Standard errors

- For assessing the accuracy of our estimation of some parameter, we most commonly use standard errors.
- For example the standard error of mean tells us in what extent is the sample mean expected to vary from one sample to another (if we would have several samples).
- When the sample size gets bigger, we are able to estimate the mean more accurately.
  - This means smaller SEs.

Confidence intervals

- Often we want an interval that most often, when samples are taken in the way that our sample has been taken, will include the population mean.
- There are two common choices for the proportion of similar samples that should contain the population mean – 95 % and 99 %.
- Confidence intervals can be used as the basis for tests of hypotheses.
  - If CI for the population mean doesn't contain zero, we will reject the hypothesis, that it would be zero.

p-value

- A p-value gives us information about probability.
- It tells us that on what probability would the estimated difference occur by chance.
- If the p-value is 0.05 that means, that there is a 5 % chance that the estimated difference would occur by chance. i.e the difference is quite markable.
- On the other han a p-value of 0.95 would mean that there is a 95 % chance, that the estimated difference would occur by chance, i.e there is clearly no real difference between groups.

What is small p-value

- At what point is a *p*-value small enough to be convincing?
- Most commonly, $p = 0.05$ (= 5 %) is used as the cutoff.
- In some special cases we might use other cutoff-values, eg. $p = 0.01$ or $p = 0.10$. In these cases we give statement that why we have selected this kind of cutoff-value.

Simple t-test

- If we want to compare means between two groups, the simplest way to go, is to perform a two-sample t-test.
- T-test is designed for comparing means. There are three different kinds of t-tests available: one-sample t-test, two-sample t-test and paired samples t-test.
- We can conduct the tests either considering equal variances or not equal variances.
- Let’s consider examples of these simple test statistics.

One sample t-test

```
>t.test(ToothGrowth$len,mu=15)

        One Sample t-test

data:  ToothGrowth$len
t = 3.8615, df = 59, p-value = 0.0002823
alternative hypothesis: true mean is not equal
  to 15
95 percent confidence interval:
 16.83731 20.78936
sample estimates:
mean of x
 18.81333
```

Two sample t-test

- `heated <- c(254,252,239,240,250,256,267,249,259,269)`
- `unheated <- c(233,252,237,246,255,244,248,242,217,257)`
- `t.test(heated,unheated, var.equal=T)`

Two sample t-test; result

```
 Two Sample t-test

data:  heated and unheated
t = 1.973, df = 19, p-value = 0.06322
alternative hypothesis: true difference in
  means is not equal to 0
95 percent confidence interval:
 -0.5723273 19.3905091
sample estimates:
mean of x mean of y
 253.5000  244.0909
```

Homogeneity of variance

- We can look at the homogeneity of variance with different plots as shown before. There are also formal tests for figuring out are the variances between groups homogenious or not.
- Most common one is called Levene's test, available for example in *car-package*.
  - Function is called *leveneTest*.

Homogeneity of variance; result

```
leveneTest(conformity~partner.status,
  data=Moore)

Levene's Test for Homogeneity of
  Variance (center = median)
       Df F value Pr(>F)
group  1  0.0611  0.806
```

Wilcoxon test

- A nonparametric option for t-test is called Wilcoxon test. The two sample version of that is called Mann-Whitney's U-test. The same example again:

```
>wilcox.test(heated,unheated)
Wilcoxon rank sum test with continuity correction
data:  heated and unheated
W = 79, p-value = 0.09774
alternative hypothesis: true location shift is
  not equal to 0
Warning message:
In wilcox.test.default(heated, unheated) :
  cannot compute exact p-value with ties
```

Contingency tables

- With contingency tables (or in other words cross-tables) we can do simple comparisons with categorical variables.
- The question we are asking from the data:
  - Is variable *a* related to variable *b*.
- We can either do just the cross-tabulation and the more formal analysis with some other method or add a test of connection with the table as well.

Cross-tables

- The simplest (but also the most stripped) way to do a cross-table in R is with the *table*-function

```
table(nsw74psid3$trt, nsw74psid3$nodeg)
```

```
     0   1
0   63  65
1   54 131
```

Cross-tables

- Nicer output is produced for example with the function *CrossTable (gmodels)*. Where we can easily add also row/column percentage, different tests (Chi-square, Fisher, McNemar) etc.
- If we want our cross-tabulation to be printed out for example in HTML or LateX we can use function *xtable* from the *xtable*-library.

CrossTable

- CrossTable(nsw74psid3$trt, nsw74psid3$nodeg, prop.chisq=F,prop.t=F)
- Total Observations in Table: 313

	nsw74psid3$nodeg		
nsw74psid3$trt	0	1	Row Total
0	63	65	128
	0.492	0.508	0.409
	0.538	0.332	
1	54	131	185
	0.292	0.708	0.591
	0.462	0.668	
Column Total	117	196	313
	0.374	0.626	

Pearson’s Chi-square test

- A Chi-square test answers to the question: ”Are these two categorical variables related to each other?” with a test statistic and a p-value.
- `chisq.test(table(nsw74psid3$trt, nsw74psid3$nodeg))`

- `Pearson's Chi-squared test with Yates' continuity correction`
- `data:table(nsw74psid3$trt,nsw74psid3$nodeg)`
- `X-squared = 12.125, df = 1, p-value = 0.0004975`

Chi-square test (2)

- Chi-square test does not give ANY information about the causality of the relation.
  - So we cannot say because variable *a* variable *b* seems to behave in some way etc. Only that the variables are related.
  - If we want to say something about the causality as well we need to form a statistical model.
- If the expected cell values are very low (say below 5) the Chi-square doesn't perform well. We have other choices there.

Fisher’s exact test

- When Pearson’s chi-square test fails (so with small samples) Fisher’s exact test gives us the correct answer.
- Fisher’s test would work with larger samples as well, but it requires a lot of memory, so when the sample size is big, it’s frustrating or impossible to use it.
- It answers the same question as does the Chi-square.

Fisher’s test; example

- `fisher.test(table(nsw74psid3$trt, nsw74psid3$nodeg))`

- `        Fisher's Exact Test for Count Data`
- `data:  table(nsw74psid3$trt, nsw74psid3$nodeg)`
- `p-value = 0.0003674`
- `alternative hypothesis: true odds ratio is not equal to 1`
- `95 percent confidence interval:`
- `  1.430486 3.863972`
- `sample estimates:`
- `odds ratio`
- `   2.344587`