Practical R: Lecture 4

Eugen Buehler

September 30, 2016

Summary Statistics

Frequently we use summary statistics to briefly describe large data sets. Among the summary statistics available in R are: mean, median, max, and min. Most of these functions will return NA if there are any missing values in the data. To ignore missing values, you need to supply the option "na.rm=TRUE".

```
> mean(c(1,56,82,NA,105,7))
```

```
[1] NA
```

```
> mean(c(1,56,82,NA,105,7), na.rm=TRUE)
```

```
[1] 50.2
```

Summary, the command

The summary command can be used to print a brief description of many things in R. When summary is called with a data frame as its argument, R responds with summary statistics for each column. Note the difference between factor columns and number columns.

```
> summary(esoph[,c(1,2,4)])
```

```
   agegp          alcgp        ncases
 25-34:15   0-39g/day:23   Min.   : 0.000
 35-44:15   40-79    :23   1st Qu.: 0.000
 45-54:16   80-119   :21   Median : 1.000
 55-64:16   120+     :21   Mean   : 2.273
 65-74:15                  3rd Qu.: 4.000
 75+  :11                  Max.   :17.000
```

Hypothesis Testing

- Hypotheses are only disproven, never proven.
- Sometimes there is only one possible alternative hypotheis.
- A p-value indicates the likelihhood of our data if the null hypothesis was true.

Parametric versus non-parametric statistics

- Parametric statistics assume an underliying distribution to the data (usually normal).
- Non-parametric statistics work off of the rank order of the data, and make no assumptions.
- From the standpoint of publication, non-parametric statistics are always safer.
- Non-parametric are inherintly less powerful.

Tests on a single variable: Normality

We can evaluate the probability that a set of data is normally distributed using the Shapiro-Wilk's Test.

```
> shapiro.test(rnorm(100, mean = 5, sd = 3))
```

```

    Shapiro-Wilk normality test

data:  rnorm(100, mean = 5, sd = 3)
W = 0.97408, p-value = 0.04574
```

```
> shapiro.test(runif(100, min = 2, max = 4))
```

```

    Shapiro-Wilk normality test

data:  runif(100, min = 2, max = 4)
W = 0.97459, p-value = 0.05012
```

Tests on a single variable: Normality

- We can't prove data is normal, only that it is not normal.
- Journals and reviewers will often ask if you tested your data for normality.
- Any sufficiently large set of data will invariably generate a small p-value for the Shapiro-Wilk Test.
- If you want to use parametric statistics and you have more than 30 or so data points, use the test to determine if it can normality can be disproven.
- If it can be disproven, you will have to use non-parametric statistics.

Tests on a single vairable: Mean

For normally distributed data, we can use the one sample t-test to evaluate the probability that the mean is equal to some number:

```
> t.test(1:10, mu=1)
```

```

    One Sample t-test

data:  1:10
t = 4.7001, df = 9, p-value = 0.00112
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 3.334149 7.665851
sample estimates:
mean of x
      5.5
```

A word on one-sided statistical tests: Dont.

Many statistical tests will give you the option of using a one-tailed or one sided null hypothesis. For example, we could allow our null hypothesis to be "the mean is greater than or equal to zero" instead of "the mean is zero". This will double the significance of our result (half the p-value) but requires a strong justification and will be deeply suspicious to any reviewer that understands statistics. Best to stick with the "two-sided" null hypothesis.

Tests on a single variable: Median

A non-parametric equivalent of the t-test is the Wilcox test, which evaluates the null hypothesis of the variable having a specifc median.

```
> wilcox.test(1:10, mu=1)
```

```
Warning in wilcox.test.default(1:10, mu = 1): cannot
compute exact p-value with zeroes


    Wilcoxon signed rank test with continuity
    correction

data:  1:10
V = 45, p-value = 0.009152
alternative hypothesis: true location is not equal to 1
```

Tests on a single variable: Proportion

If we get 53 heads our of 100 coin tosses, how likely is it that our coin is unbiased (probability of heads or tails is 0.5)?

```
> prop.test(53,100, 0.5)
```

```

    1-sample proportions test with continuity
    correction

data:  53 out of 100, null probability 0.5
X-squared = 0.25, df = 1, p-value = 0.6171
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4280225 0.6296465
sample estimates:
   p
0.53
```

Statistical tests for two variables

- Two categorical variables
- A categorical variable and a continuous variable
- Two continuous variables

Two categorical variables

Before testing whether two categorical variables are correlated, we will need to get them into a contingency table.

```
> esoph[1:2,]
```

```
  agegp     alcgp    tobgp ncases ncontrols
1 25-34 0-39g/day 0-9g/day      0        40
2 25-34 0-39g/day    10-19      0        10
```

```
> table(esoph[,1:2])
```

```
       alcgp
agegp   0-39g/day 40-79 80-119 120+
  25-34         4     4      3    4
  35-44         4     4      4    3
  45-54         4     4      4    4
  55-64         4     4      4    4
  65-74         4     3      4    4
  75+           3     4      2    2
```

Chi-Square Test

The null hypothesis is that there is no relationship between the two categorical variables.

```
> chisq.test(table(esoph[,1:2]))
```

```
Warning in chisq.test(table(esoph[, 1:2])): Chi-squared
approximation may be incorrect


    Pearson's Chi-squared test

data:  table(esoph[, 1:2])
X-squared = 1.4189, df = 15, p-value = 1
```

Fisher's Exact Test

Fisher's exact test also associates a p-value with the null hypothesis that the two categroical variables are unrelated. However, it is not limited by the total numbers of counts or the counts in any cell. Preferable to use instead of Chi-Square, but reviewers may be less familiar with it. Fisher's is also sometimes refered to as the "hyper-geometric test", because of the distribution expected under the null hypothesis.

```
> fisher.test(table(esoph[,1:2]))
```

```

    Fisher's Exact Test for Count Data

data:  table(esoph[, 1:2])
p-value = 1
alternative hypothesis: two.sided
```

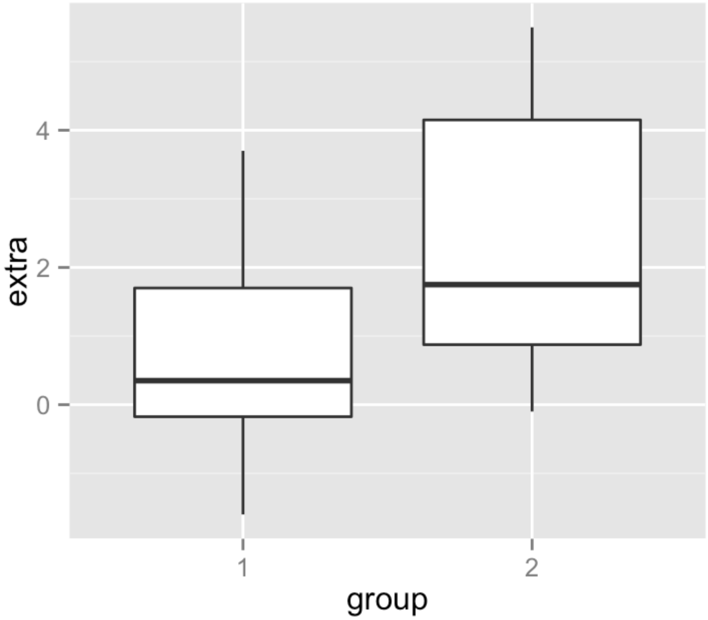
Tests for relationship between a two-value categorical variable and a continuous variable.

- T-test (paired and unpaired)
- Wilcox Test
- KS-Test

Student's Sleep Data

"Data which show the effect of two soporific drugs (increase in hours of sleep compared to control) on 10 patients."

```
> library(ggplot2)
> ggplot(sleep, aes(x=group, y=extra)) + geom_boxplot()
```



T-test

T-tests are parametric (only valid on normal distributions) and can be paired or unpaired. It will generally be more powerful to use a paired t-test on data where there is a natural pairing between continuous valued data points. For example, if you had measurements of LDL cholesterol before and after treatment, it would be logical to use a paired t-test instead of an unpaired test.

T-test, the long way

```
> t.test(sleep$extra[sleep$group == 1], sleep$extra[sleep$group == 2])
```

```

    Welch Two Sample t-test

data:  sleep$extra[sleep$group == 1] and sleep$extra[sleep$group == 2]
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean of x mean of y 
     0.75      2.33 
```

T-test, the short way

Using the formula interface, which we will learn more about in later lectures. For the t-test, we place the continuous vairable to the left of the "~" and the factor (or categorical variable) the the right.

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

```

	Welch Two Sample t-test

data:  extra by group
t = -1.8608, df = 17.776, p-value = 0.07939
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.3654832  0.2054832
sample estimates:
mean in group 1 mean in group 2 
           0.75            2.33 
```

Wilcox Test

The non-parametric test if two groups have different medians. We can use the same formula interface we used for the t-test.

```
> wilcox.test(extra ~ group, data = sleep)
```

```
Warning in wilcox.test.default(x = c(0.7, -1.6, -0.2,
-1.2, -0.1, 3.4, 3.7, : cannot compute exact p-value
with ties


    Wilcoxon rank sum test with continuity
    correction

data:  extra by group
W = 25.5, p-value = 0.06933
alternative hypothesis: true location shift is not equal to 0
```

KS-test

A non-parametric test that two variables come from the same distribution. Formula notation not allowed.

```
> ks.test(sleep$extra[sleep$group == 1], sleep$extra[sleep$group == 2])
```

```
Warning in ks.test(sleep$extra[sleep$group == 1], sleep
$extra[sleep$group == : cannot compute exact p-value
with ties


    Two-sample Kolmogorov-Smirnov test

data:  sleep$extra[sleep$group == 1] and sleep$extra[sleep$group == 2]
D = 0.4, p-value = 0.4005
alternative hypothesis: two-sided
```

Correlation of Continuous Variables

- Correlation (pearson and spearman)
- Tests of correlation

Lawyers' Ratings of State Judges in the US Superior Court

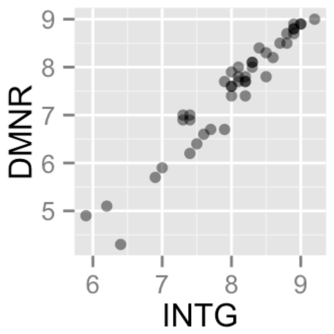
USJudgeRatings is a data frame in the datasets package. It includes "Lawyers' ratings of state judges in the US Superior Court."

- CONT Number of contacts of lawyer with judge.
- INTG Judicial integrity.
- DMNR Demeanor.
- DILG Diligence.
- CFMG Case flow managing.
- DECI Prompt decisions.
- PREP Preparation for trial.

Correlation Metrics: Pearson

Pearson correlation (aka r). 1 indicates perfect correlation, 0 indicates no correlation, -1 indicates perfect inverse correlation.

```
> ggplot(USJudgeRatings, aes(x=INTG, y=DMNR)) + geom_point(alpha=0.5)
```



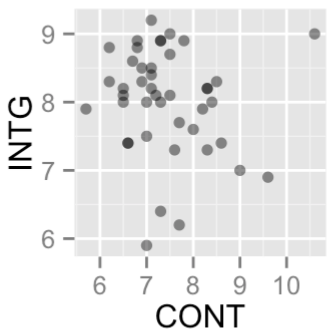
```
> cor(USJudgeRatings$INTG, USJudgeRatings$DMNR, method="pearson")
```

```
[1] 0.9646153
```

Correlation Metrics: Spearman

Rho: Same boundaries as pearson (-1,1), but non-parametric.

```
> ggplot(USJudgeRatings, aes(x=CONT, y=INTG)) + geom_point(alpha=0.5)
```



```
> cor(USJudgeRatings$CONT, USJudgeRatings$INTG, method="spearman")
```

```
[1] -0.1764773
```

Use

Use: An optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings "everything", "all.obs", "complete.obs", "na.or.complete", or "pairwise.complete.obs".

In general, I would recommend using "complete.obs".

```
> cor(USJudgeRatings$CONT, USJudgeRatings$INTG, use="complete.obs")
```

```
[1] -0.1331909
```

Testing Correlation

We can test the null hypothesis that r or rho are 0 (variables are uncorrelated).

```
> cor.test(USJudgeRatings$CONT, USJudgeRatings$INTG, method="pearson")
```

```

    Pearson's product-moment correlation

data:  USJudgeRatings$CONT and USJudgeRatings$INTG
t = -0.8605, df = 41, p-value = 0.3945
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4168591  0.1741182
sample estimates:
       cor
-0.1331909
```

Testing Correlation

```
> cor.test(USJudgeRatings$CONT, USJudgeRatings$INTG, method="spearman")
```

```
Warning in cor.test.default(USJudgeRatings$CONT,
USJudgeRatings$INTG, method = "spearman"): Cannot
compute exact p-value with ties


    Spearman's rank correlation rho

data:  USJudgeRatings$CONT and USJudgeRatings$INTG
S = 15581, p-value = 0.2576
alternative hypothesis: true rho is not equal to 0
sample estimates:
       rho
-0.1764773
```