Intro to R for Epidemiologists

Lab 5 (2/12/15)

Many of these questions go beyond the information provided in the lecture. Therefore, you may need to use R help files and the internet to search for answers. Feel free to ask questions of the instructor, the TAs, or your classmates, but try to work through as much as you can independently.

For the lab, you are expected to create an R script (.R file in the R editor) with your code corresponding to each question. Begin each question with a commented line of code indicating the question. As an example:

```
# Jenna Krall

# Question 1.
head(iris)
```

Part 1. Statistical Tests for Continuous Data

We will use the `chickwts` dataset to look at statistical tests for continous variables. This dataset contains the weight (in grams) of 6 week old chicks fed one of six different diets. For all of these questions, assume α=.05. (Source: Biometrika, vol. 35)

1. Test the hypothesis that the mean chick weight is 265 grams (ignoring feed groups). (Hint: `?t.test`)
2. Use the `tapply` statement to get the mean weight by feed.
3. Find the difference in mean weight between linseed diet and soybean diet.
4. Test whether the mean weight in the linseed diet group is the same as the mean weight in the soybean diet group.
    - Extract the confidence interval for the difference in mean weights.
    - Extract the p-value for this test
5. Test whether the mean weight in the linseed diet group is the less than the mean weight in the soybean diet group (Hint: check the alternative argument in `t.test`)

```
# Part 1
head(chickwts)
```

```
##   weight      feed
## 1    179 horsebean
## 2    160 horsebean
## 3    136 horsebean
## 4    227 horsebean
## 5    217 horsebean
## 6    168 horsebean
```

```
# Test whether mean is equal to 260
t_weight <- t.test(x = chickwts$weight, mu = 260)
t_weight
```

```
##
##  One Sample t-test
##
## data:  chickwts$weight
## t = 0.1414, df = 70, p-value = 0.888
## alternative hypothesis: true mean is not equal to 260
## 95 percent confidence interval:
##  242.8301 279.7896
## sample estimates:
## mean of x
##  261.3099
```

```
# 2. Find means by species
means <- tapply(chickwts$weight, chickwts$feed, mean)[c("linseed", "soybean")]
means
```

```
##  linseed  soybean
## 218.7500 246.4286
```

```
# 3. Find mean difference
meandiff <- means["linseed"] - means["soybean"]

# 4. First get vectors of mean weight by diet
linseed <- chickwts$weight[chickwts$feed == "linseed"]
soybean <- chickwts$weight[chickwts$feed == "soybean"]
# Apply ttest
ttest_weight <- t.test(linseed, soybean)
# see R objects in t.test
names(ttest_weight)
```

```
## [1] "statistic"   "parameter"   "p.value"     "conf.int"    "estimate"
## [6] "null.value"  "alternative" "method"      "data.name"
```

```
# Extract out confidence interval and p-value
ttest_weight$conf.int
```

```
## [1] -70.84262  15.48547
## attr(,"conf.level")
## [1] 0.95
```

```
ttest_weight$conf.int[1:2]
```

```
## [1] -70.84262  15.48547
```

```
ttest_weight$p.value
```

```
## [1] 0.1979871
```

```r
# Using equal variances, test whether equal
ttest_weight <- t.test(linseed, soybean, alternative = "less")
ttest_weight
```

```
##
##  Welch Two Sample t-test
##
## data:  linseed and soybean
## t = -1.3246, df = 23.63, p-value = 0.09899
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##     -Inf 8.09536
## sample estimates:
## mean of x mean of y
##  218.7500  246.4286
```

Part 2. Statistical Tests for Categorical Data

Suppose we have cross-sectionally collected data on smoking status and cancer status on 206 individuals. Our data looks like this:

	Cancer	No cancer
Smoker	94	27
Non-smoker	22	63

1. Is the proportion of smokers with lung cancer statistically significantly different from the proportion of non-smokers with lung cancer?
    - Create your table `sc_table <- matrix(c(94, 22, 27, 63), ncol = 2)`
    - Name your columns and rows
    - Use the function `prop.test`
2. Use a chi-squared test to determine whether smoking and lung cancer are independent (Hint: `chisq.test`)
3. Using the epitools package, extract the odds ratio and its 95% confidence interval for the association between smoking and lung cancer (Hint: `?epitab`).
4. Using the epitools package, find the risk ratio and its 95% confidence interval for the association between smoking and lung cancer (Hint: `?epitab`. Look at the method argument).

```r
# Part 2 Create matrix
sc_table <- matrix(c(94, 22, 27, 63), ncol = 2)
# Get names
rownames(sc_table) <- c("Smoker", "Nonsmoker")
colnames(sc_table) <- c("Cancer", "No_cancer")
# Test the proportions
prop.test(sc_table)
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  sc_table
## X-squared = 52.3763, df = 1, p-value = 4.582e-13
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.3889706 0.6471013
## sample estimates:
##    prop 1    prop 2
## 0.7768595 0.2588235
```

```
# Chi-squared test
chisq.test(sc_table)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  sc_table
## X-squared = 52.3763, df = 1, p-value = 4.582e-13
```

```
# Odds ratio NOTE: EPITAB Uses first row and first column as reference
# groups Disease should be in columns
library(epitools)

# the rev argument will rearrange the table for us
table1 <- epitab(sc_table, rev = "both")

# Or rearrange data ourselves
sc_table <- (matrix(c(63, 27, 22, 94), ncol = 2))
# Get names
rownames(sc_table) <- c("Nonsmoker", "Smoker")
colnames(sc_table) <- c("No_cancer", "Cancer")
# get odds ratio
table1 <- epitab(sc_table)
table1$tab[2, c("oddsratio", "lower", "upper")]
```

```
## oddsratio     lower     upper
##  9.969697  5.219787 19.041938
```

```
# Relative risk
table1 <- epitab(sc_table, method = "riskratio")
table1$tab[2, c("riskratio", "lower", "upper")]
```

```
## riskratio     lower     upper
##  3.001503  2.068673  4.354975
```