Basic Course on R: Hypothesis Testing and Confidence Intervals 2

Elizabeth Ribble*

14-18 May 2018

Contents

*emcclel3@msudenver.edu

Most of the following examples use the data "R_data_January2015.csv" which contains variables on mothers whose babies are either intellectually disabled or developmentally normal.

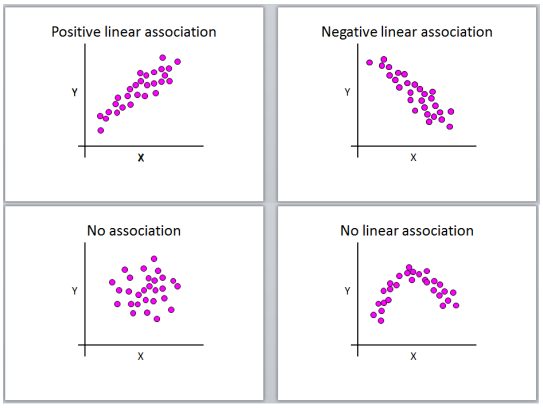
```
babies <- read.csv("R_data_January2015.csv",header=T,row.names=1)
names(babies)

##  [1] "Status"                "iodine_deficiency"
##  [3] "BMI"                   "educational_level"
##  [5] "alcohol"               "smoking"
##  [7] "medication"            "birthweight"
##  [9] "pregnancy_length_weeks" "pregnancy_length_days"
## [11] "SAM"                   "SAH"
## [13] "homocysteine"          "cholesterol"
## [15] "HDL"                   "triglycerides"
## [17] "vitaminB12"            "folicacid_serum"
## [19] "folicacid_erys"

attach(babies)
```

1 Correlations

If we want to know the degree of a linear association between 2 variables, we can calculate correlation. Correlation does not make any *a priori* assumptions about whether one variable is dependent on the other and is not concerned with the relationship between the variables. We have 4 general types of association to consider:



Pearson's correlation coefficient r is defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

and indicates to what extent two continuous variables have a linear relationship.

Pearson's r can only lie in the interval [-1,1] (inclusive), where

- $r = 0$, no linear correlation
- $r > 0$, positive linear correlation
- $r < 0$, negative linear correlation
- $r = 1$, perfect positive linear correlation
- $r = -1$, perfect negative linear correlation

Note that correlation does *not* imply causation. If two variables are highly correlated you cannot infer that one is causing the other; they could both be varying along with a third, possibly unknown confounding factor (either causal or not).

We assume that the two variables x and y

- show a linear relationship
- are continuous random variables
- are normally distributed
- are independent of each other

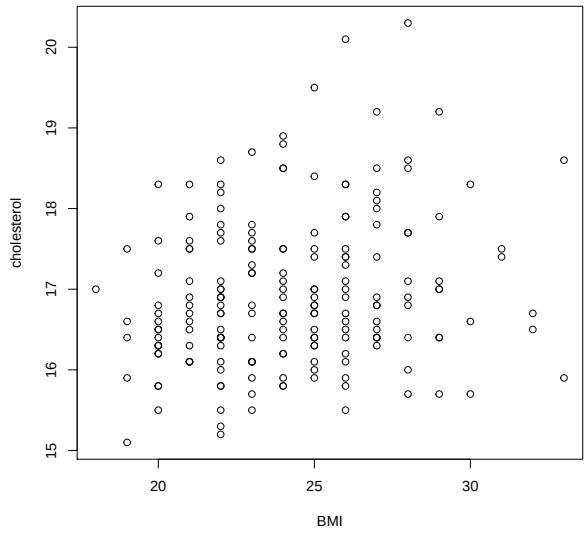
and when regressing one variable on another (e.g. linear regression, more on this later), r^2 is called the goodness of fit, which is the proportion of variance in y that can be explained by the regression on x.

If data are not normally distributed, the degree of association can be determined by the ranked correlation coefficient, Spearman's ρ, which replaces the x's and y's in the Pearson formula with their ranks.

R provides p-values and confidence intervals for both Pearson and Spearman correlations.

Let's look at an example of how BMI and cholesterol are associated:

```
plot(BMI,cholesterol)
```




```
cor(BMI, cholesterol)

## [1] 0.2004703

cor(BMI, cholesterol, method="spearman")

## [1] 0.1876358

cor.test(BMI,cholesterol)

##
##  Pearson's product-moment correlation
##
## data:  BMI and cholesterol
## t = 2.8057, df = 188, p-value = 0.00555
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05982424 0.33331173
## sample estimates:
##       cor
## 0.2004703
```

```r
cor.test(BMI,cholesterol, method="spearman")

## Warning in cor.test.default(BMI, cholesterol, method = "spearman"): Cannot
compute exact p-value with ties

##
##  Spearman's rank correlation rho
##
## data:  BMI and cholesterol
## S = 928640, p-value = 0.009531
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.1876358
```

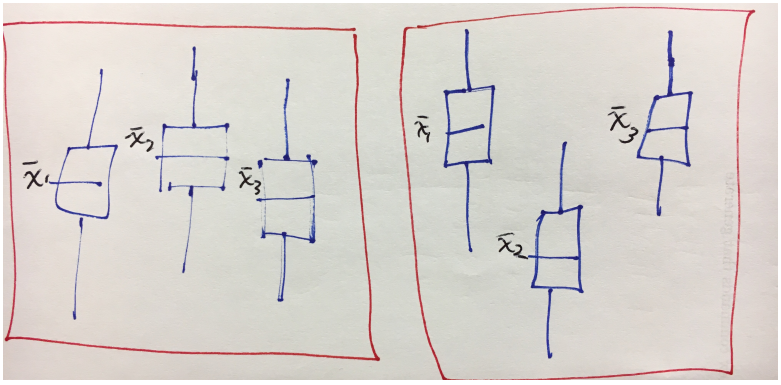
Note the *warning* message (which is not an error!) that indicates if your data have tied values (e.g. 1, 1, 3, 5) then the p-value is approximated and is not exact. It's nothing to worry about (especially if you're not a statistician...).

2 ANOVA

A two-sample t-test is used to test hypotheses about the means of two normal populations using two datasets which were sampled independently of one another from the two respective populations. An analysis of variance (ANOVA) allows for comparing the means of one variable among more than two populations (each of which is normally distributed), again under the assumption that the samples are independent.

- If the variance between groups is higher than the variance within groups, then there is evidence for a difference in means between groups.
- The null hypothesis is that the means of all groups are equal; the alternative is that at least one group has a different mean from the others.
- ANOVA does not provide which group is different nor in what direction; visualization and/or post-hoc pairwise t-tests can provide this information.

Here is a small illustration of the analysis of variance:



In the left image there is more variablility within each of the three groups (boxplots) than between the three group means ($\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$). In the right image there is more variability between the three group means than within each of the three groups.

2.1 One-way ANOVA

Suppose we have k groups and n total samples. We first calculate between- and within-group variations (sums of squares or SS), and for this we need the mean of all data points, which is $\bar{X} = \sum x/n$. The SS (between and within) for each group i are:

$$SSB = \sum n_i(\bar{x}_i - \bar{X})^2$$
$$SSW = \sum (n-k)s_i^2$$

where s_i^2 is the variance within group i. It can be shown that $SSB + SSW = SST$, the total sum of squares $\sum (x - \bar{X})^2$ (the total variance times $n-1$). Next we calculate the mean squared (MS) errors, which are:

- MSB for between: $SSB/(k-1)$
- MSW for within: $SSW/(n-k)$.

The test-statistic for an ANOVA is MSB/MSW, which, to get a p-value, is compared to an F distribution on $k-1$ and $n-k$ degrees of freedom.

The above statistics are generally presented in an ANOVA table:

	df	SS	MS	F
between	k-1	SSB	$SSB/(k-1)$	MSB/MSW
within	n-k	SSW	$SSW/(n-k)$	
total	n-1	$SSW + SSB$		

As with the previous testing procedures we've seen, we compare the p-value to the pre-selected significance level. If $p \leq \alpha$ we reject the null and conclude we have evidence that at least one population (or group) mean is different from the others, while if $p > \alpha$, we fail to reject the null and conclude we do not have evidence that any of the means are different.

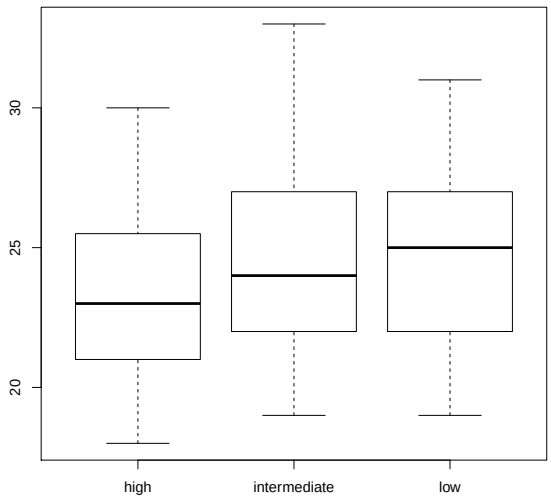
R will calculate all of the statistics for us! For example, let's test at the 5% significance level the alternative hypothesis that at least one of the mean BMIs for three different educational levels are different from the other two (versus the null of three equal means).

```
baby_aov1 <- aov(BMI ~ educational_level)
summary(baby_aov1)

##                     Df Sum Sq Mean Sq F value Pr(>F)
## educational_level    2   63.8   31.90   3.409 0.0351 *
## Residuals          187 1749.8    9.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here the F-statistic is large enough to get a p-value smaller than 0.05 so we conclude at least one population mean BMI is different from the other two. But which group is it? Let's look at a boxplot:

```
boxplot(BMI~educational_level)
```



This visual information is informative but we can actually make pairwise comparisons of all the means using Tukey's HSD. We initially used a signifance level of 0.05, but will now conduct several tests and use a multiple test procedure that adjusts the family-wise error rate to 5%:

```
TukeyHSD(baby_aov1)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = BMI ~ educational_level)
##
## $educational_level
##                         diff         lwr      upr     p adj
## intermediate-high 1.1397600 -0.07099027 2.350510 0.0697761
## low-high          1.3587302 -0.05180238 2.769263 0.0617385
## low-intermediate  0.2189702 -1.12174472 1.559685 0.9212512
```

Note that we use the multiple testing procedure to account for the fact that we are simultaneously performing more than one test or CI which compounds the error rates of each test. Since all of the CIs contain 0 and none of the p-values are significant at the 5% level, the initial test result is overturned! We do not actually have evidence that education level affects average BMI.

Also note you would not bother with these *posthoc* tests/CIs if your initial ANOVA results were not statistically significant.

2.2 Two-way ANOVA

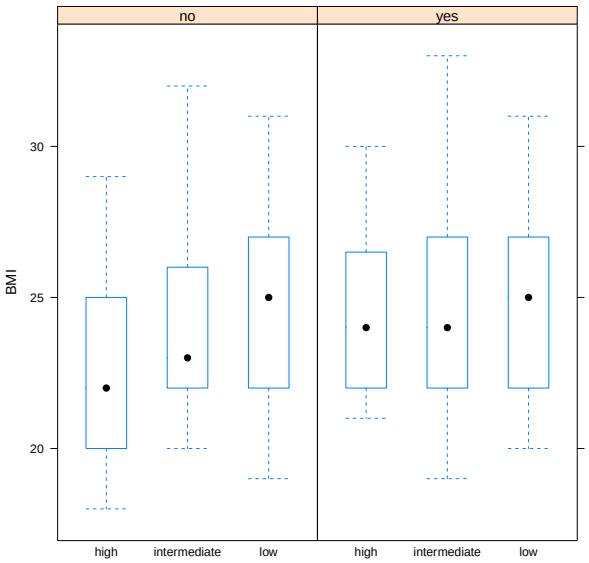
When we have two categorical explanatory variables and a continuous outcome, we can use a two-way ANOVA to determine whether the combinations of levels of the two variables have varying means in the outcome. Note that this is still a parametric test which requires normality. The ANOVA table is similar to the one-way, but an additional row is added for the second factor and an interaction between the two groups.

```
baby_aov2 <- aov(BMI ~ educational_level * iodine_deficiency)
summary(baby_aov2)

##                                      Df Sum Sq Mean Sq F value Pr(>F)
## educational_level                     2   63.8   31.90   3.441 0.0341 *
## iodine_deficiency                     1   27.4   27.40   2.955 0.0873 .
## educational_level:iodine_deficiency   2   16.6    8.32   0.898 0.4092
## Residuals                           184 1705.7    9.27
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(lattice)
```

```
bwplot(BMI ~ educational_level|iodine_deficiency)
```



Here the interaction is not significant so we take it out and re-run the ANOVA:

```
baby_aov2 <- aov(BMI ~ educational_level + iodine_deficiency)
summary(baby_aov2)

##                    Df Sum Sq Mean Sq F value Pr(>F)
## educational_level   2   63.8   31.90   3.445 0.0340 *
## iodine_deficiency   1   27.4   27.40   2.958 0.0871 .
## Residuals         186 1722.4    9.26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It appears that educational level and iodine deficiency may have an effect on BMI.

3 χ^2 Test

A χ^2 (read: chi-squared) test of independence tests the null hypothesis that rows and columns are independent in a $c \times r$ contingency table (r is number of rows, c columns); the alternative is that they are not independent. The counts must be independent and sampled randomly. We can calculate a chi-squared statistic as follows:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count for cell i in the table, and E_i is the expected count, calculated by multiplying the row and column totals for i divided by the overall total. The p-value for the calculated χ^2 statistic depends on the χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. R will provide us with the statistic and p-value.

Suppose we have the following table:

	event	no event	total
treatment	20	80	100
placebo	50	50	100
total	70	130	200

and we want to know whether our treatment prevents events, that is, does the occurence of an event depend on the type of treatment?

```
mytable <- matrix(c(20, 50, 80, 50), nrow=2)
cstest <- chisq.test(mytable)
cstest

##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mytable
## X-squared = 18.484, df = 1, p-value = 1.714e-05

cstest$p.value

## [1] 1.713801e-05
```