
BEES
School of Biological,
Earth & Environmental
Sciences at University
College Cork, Ireland

BL6024 - Quantitative Skills for Biologists using R

Lecture 3: Common statistical tests

Chi-square

- The chi-square contingency table analysis is an analysis of count data
- It is a test of association among two or more categorical variables
- Does not assume the data are normally distributed (non-parametric test)
- Example: ladybirds in rural and industrial sites. Is there an association between ladybird colours and their habitat?

```
lady = read.csv("ladybirds_morph_colour.csv")
lady # look at the data
str(lady) # 20 observations, with a count (number) for each one. 5 sites per habitat
```

Chi-square

- So, do the frequencies of the two colours of ladybirds differ between the two habitat types?
- Using a chi-square, we need to calculate totals and in this case end up with 4 numbers. What would those 4 numbers be?

Chi-square

```
library(dplyr)
totals = lady %>%
    group_by(Habitat,morph_colour) %>%
        summarise(total.number = sum(number))
totals
```

- %>% denotes a pipe. You can read this like "put the answer of the left-hand command into the function on the right"

Chi-square

- Let's plot those totals. In this case, a bar chart will do the job:

```
library(ggplot2)
ggplot(totals, aes(x = Habitat, y = total.number, fill = morph_colour)) +
  geom_bar(stat = 'identity', position = 'dodge') +
  scale_fill_manual(values = c(black = "black", red = "red"))
```

- It looks like we may need to reject the null hypothesis

Chi-square

- Let's look at the data again:

```
totals # now we will make a contingency table:
lady.table = xtabs(number ~ Habitat + morph_colour, data = lady)
# Have a look:
lady.table
# Run the chi-square:
chisq.test(lady.table)
```

Chi-square

- One way to report these results:

"We tested the hypothesis that there is an association between colour morphs of ladybirds and industrial and rural habitats. Ladybird colour morphs are not equally distributed in the two habitats (Chi-square = 19.1, df = 1, $p < 0.001$), with black morphs being more frequent in industrial habitats and red morphs more frequent in rural habitats (Figure 1)."

Chi-square

- There are other ways to run a chi-square. Let's go simpler:

```
df <- read.csv("https://goo.gl/j6lRXD")
head(df, 10)
table(df$treatment, df$improvement)
plot(df$treatment, df$improvement)
chisq.test(df$treatment, df$improvement)
```

Chi-square

- Even simpler. If you already know the frequencies you don't even need to load the whole dataset into R
- Question: Do we have a 1:1 sex ratio in our population?

```
obs = c(50,20) # let's say 50 males and 20 females
expP = c(0.50, 0.50) # we expect 1:1 sex ratio
chisq.test(obs, p = expP)
```

Chi-square

- One last example:

```
M = matrix(c(22, 4, 15, 10), byrow = TRUE, nrow = 2)
```

- What did we do there? Now calculate the chi-square

Chi-square

Exercise

- Corn kernels were counted in an ear of corn.
- There were 295 purple kernels and 86 yellow kernels.
- Are these counts consistent with a 3:1 ratio of purple to yellow?

Correlation

- Measure of association (negative or positive)
- Cannot attribute causality
- Correlation coefficients say nothing about which variable causes the other to change
- 2 types of correlation: bivariate and partial
    - partial correlation controls for variation in other variables
- Parametic correlation => Pearson correlation
- Non-parametric correlation:
    - Spearman's rho
    - Kendall's tau (better than Spearman for small samples)

Correlation

Correlation coefficient

- varies between -1 and +1
    - 0 = no relationship
- it is an effect size
    - +/- 0.1 = small effect
    - +/- 0.3 = medium effect
    - +/- 0.5 = large effect
- R^2 - by squaring the value of R you get the proportion of variance in one variable shared by the other

Correlation

- To compute basic correlation coefficients there are several functions that can be used:

```
cor()
cor.test() # I normally use this one
rcorr()
```

Correlation

Example

- We will use the R dataset iris
- Investigate the dataset, e.g. use str()
- Question: Is sepal length correlated with petal length in any of the 3 species?

```
setosa = subset(iris, Species=="setosa")
plot(setosa$Sepal.Length, setosa$Petal.Length, cex = 1.5, pch = 19,
     xlab = "Sepal Length", ylab = "Petal Length", cex.lab = 1.5)
shapiro.test(setosa$Sepal.Length)
shapiro.test(setosa$Petal.Length)
cor.test(setosa$Sepal.Length, setosa$Petal.Length)
```

"There is not a significant correlation between sepal length and petal length (r = 0.27, p = 0.06)"

- Repeat with any other species

Correlation

Exercise

- Load the "SoaySheepFitness" dataset
- Plot
- Determine normality
- Run correlation

t-test

- Comparison between two groups
- Independent & dependent (or repeated measures) models
- function t.test()

```
# Data for different groups stored in a single column:
model<-t.test(outcome~predictor, paired=FALSE/TRUE)
#For example:
ind.t.test<-t.test(Anxiety~Group)

# Data for different groups stored in two columns:
model<-t.test(scores group 1, scores group 2, paired=FALSE/TRUE)
# For example:
dep.t.test<-t.test(control,treatment, paired=TRUE)
```

t-test

Example

- Heights of 10 plants grown with fertilizer (fert) and 10 plants without (control)
- Independent samples

```
fert = c(110.3, 130.4, 114.0, 135.7, 129.9, 98.2, 109.4, 131.4, 127.9, 125.7)
control = c(64.7, 86.6, 67.1, 62.5, 75.1, 83.8, 71.7, 83.4, 90.3, 82.7)
boxplot(control, fert)
boxplot(control, fert, names = c("Control", "Fertilizer"),
    xlab = "Treatment", ylab = "Plant Height (cm)", cex.lab = 1.4)
```

- Interpret
- Swap the two boxplots
- Check normality for the Fertility and Control samples using shapiro.test()
- Also test that variances are equal using the var.test(group1,group2)
- Conduct the t-test: t.test(x,y, paired= FALSE/TRUE, var.equal=TRUE/FALSE)

One-way ANOVA

- Use ANOVA to compare several means (do not do a series of t-tests on all pairwise combinations of the data! This increases type I error)
- Constancy of variance (homoscedasticity) is the most important assumption underlying regression and analysis of variance
    - var.test() for two samples
    - for more than two samples Bartlett test or Fligner-Killeen test: bartlett.test(); fligner.test()
- ANOVA is an omnibus test
    - it test for an overall difference between groups
    - it tells us that the group means are different
    - it doesn't tell us exactly which means differ
- ANOVA is simply a special case of regression; traditional ANOVA is simply regression with categorical predictors
- R does this through the general linear model (GLM): functions lm() & aov()

Post hoc tests

- Compare each mean against all others
- Tukey and Dunnett's tests can be done using the glht() function in the multcomp package
- Also TukeyHSD()

ANOVA

Example

- We will use the "Daphniagrowth.csv"" dataset. Get it into Rstudio and look at the data
- Question: Do parasites reduce *Daphnia* growth?
- We will use ggplot2 again to make the figures, so make it available
- We will Plot -> Model -> Check Assumptions -> Interpret

ANOVA

Plot first

```
daphnia =read.csv("Daphniagrowth.csv")
library(ggplot2)
ggplot(daphnia, aes(x=parasite, y=growth.rate)) # aes = aesthetics
```

```
ggplot(daphnia, aes(x=parasite, y=growth.rate)) +
  geom_boxplot()
```

```
ggplot(daphnia, aes(x=parasite, y=growth.rate)) +
  geom_boxplot() +
  theme_bw() # gets rid of the grey background
```

```
ggplot(daphnia, aes(x=parasite, y=growth.rate)) +
  geom_boxplot() +
  theme_bw() +
  coord_flip() # Maybe you prefer it this way
```

ANOVA

Next write the model

```
model.growth = lm(growth.rate ~ parasite, data = daphnia)
```

Then check the assumptions

```
install.packages("ggfortify")
library(ggfortify)
autoplot(model.growth)
```

- Residuals vs Fitted, if model is appropriate you should not see any pattern
- Q-Q plot, points should be closed to the line. If so, normality is met
- Scale-location, evaluates the assumption of equal variance. No pattern is good
- Leverage, to see influential points and outliers

ANOVA

Finally, we can interpret our model

- We can use anova() and summary()

```
anova(model.growth)
# F value = between-group variance / within-group variance
summary(model.growth)
```

- Post hoc tests (e.g. Tukey's test)

```
model.aov = aov(model.growth) # TukeyHSD() does not work with lm models
TukeyHSD(model.aov)
```

ANOVA

Exercise

- Use the iris dataset again
- Question: are there any differences in petal length between the 3 species?
- Remember: Plot -> Model -> Check Assumptions -> Interpret
- For plotting, you can use ggplot or base R. For example:

```
boxplot(iris$Petal.Length~iris$Species, ylab= "Petal Length", cex.lab= 1.4)
```

2-way ANOVA

- Two categorical factors and their interaction
- Let's go through an example: "growth.csv". Load and check
- Each combination of diet and supplement was replicated 4 times (fully factorial experimental design)

Let's plot the data

```
cow = read.csv("growth.csv")
library(dplyr)
cow.means = cow %>% group_by(diet, supplement) %>% summarise(meanGrow = mean(gain))
cow.means # see what we have done
library(ggplot2)
ggplot(cow.means, aes(x=supplement, y=meanGrow, color=diet, group=diet)) +
  geom_point() +
  geom_line() +
  theme_bw()
```

2-way ANOVA

Build the model

- H0 => The effect of supplement type on cow weight gain does not depend on the diet
- So we need to include an interaction in the model
- Supplement*diet => 2 main effects + interaction

```
cow.model = lm(gain ~ diet * supplement, data = cow)
```

Check the assumptions and interpret the model

```
library(ggfortify)
autoplot(cow.model, smooth.colour=NA) # assumptions. Looks OK
anova(cow.model)
summary(cow.model)
```

ANCOVA

- Another form of glm which includes covariates (continuous). That is, we have one categorical and one continuous explanatory variables
- Including covariate into model allows us to "partial out" its effect
- Including covariate takes away some of the unexplained variance - allows the effects of the other variables to be seen more clearly

ANCOVA

Example

- Let's use the "limpet.csv" dataset. Examine the data
- Question: "Does the density dependence of egg production differ between spring and summer?"
- First, let's make a plot

```
ggplot(limpets, aes(x=DENSITY, y=EGGS, color=SEASON)) +
  geom_point() +
  scale_color_manual(values = c(spring="dark green", summer="red")) +
  theme_bw()
```

- Now the rest: make the model, check the assumptions and interpret results

```
limpets.mod = lm(EGGS ~ DENSITY*SEASON, data=limpets)
autoplot(limpets.mod)
anova(limpets.mod) # You could remove the interaction (EGGS ~ DENSITY+SEASON, data=limpets)
```

Non-parametric equivalents

When data is normally distributed:

- **t-test**: Two groups (independent or dependent samples)
- **ANOVA**: More than two groups

When data is non-normally distributed:

- Mann-Whitney test - Two independent groups - wilcox.test(x,y)
- Wilcoxon test - Two dependent groups - wilcox.test(x, y,paired=TRUE)
- Kruskal-Wallis - More than two groups - kruskal.test(y~A)
    - For post hoc tests you can use kruskalmc() in the pgirmess package
- https://www.statmethods.net/stats/nonparametric.html (https://www.statmethods.net/stats/nonparametric.html)

Non-parametric equivalents

Example

- Use the mtcars R database
- am => 0 is automatic, 1 is manual
- Question: Do automatic and manual cars differ in their number of cylinders?

```
boxplot(mtcars$cyl~mtcars$am)
#we don't check for normality, as we have count data
wilcox.test(cyl~am,data=mtcars)
```

In-class Continuous Assessment #1

- Complete and submit exercise before you leave
- Work on your own
- You can use your notes, any books you have, and the internet

In-class Continuous Assessment #1

- Complete and submit exercise before you leave
- Work on your own
- You can use your notes, any books you have, and the internet
- Use the R dataset"chickwts"
- Question: Does chick weight depend on feed type?
- Create a new script. Write your name in the first line of the script. Write all the code necessary to address the question above. Make plots as polished as possible, check assumptions and interpret results. Write any notes and interpretation using # symbols