Laboratory & Professional skills for Bioscientists
Term 2: Data Analysis in R

Week 4: Chi-squared tests

Overview of topics

Week	Topic
2	Introduction to module, statistics and RStudio including first figure
3	Hypothesis testing, variable types; functions (inbuilt), different ways of getting data into RStudio, getting help in RStudio
4	**Chi-squared tests**
5	The normal distribution, summary statistics and confidence intervals; user-defined functions, RStudio
6 and 7	One- and two-sample t-tests and their non-parametric equivalents (2 lectures)
8	One-way ANOVA and Kruskal-Wallis
9	Two-way ANOVA incl understanding the interaction
10	Correlation and regression

Follow up from last week's practical

- Independent study: seal myoglobin exercise at the end of this lecture…
- But first………

Summary of this week

- We start significance testing
- We will introduce the analysis of counts of things falling into mutually exclusive categories using two types of chi-squared test

Learning objectives for the week

By actively following the lecture and practical and carrying out the independent study the successful student will be able to:

- recognise when to use chi-squared Goodness of Fit and Contingency tests (MLO 2)
- be able to carry out, interpret and report scientifically both types of test in R (MLO 3 and 4)

Why chi-squared?

- When we count the number of things in categories and compare the numbers we observe to numbers we expect under a null hypothesis.
- H_0 might expect numbers to
  - be the same, or
  - follow a particular pattern, or
  - match the pattern in another group
- Chi-squared allows us to make the comparison statistically

Our two example scenarios

- The Candy-striped spider can be plain or striped
  - 2 alleles at one locus, striped dominant to plain
  - We perform: Ss x ss = Ss, Ss, ss,ss
  - We expect the ratio of striped : plain to be 1:1

Example scenarios

- Food choice by pig breeds
  - We don’t know what proportions are expected but do expect it to be same for each breed

Welsh

Tamworth

Essex

cabbage

sugarbeet

swede

Two types of scenario thus two types of χ^2 test

- We know what the proportions should be (known as *a priori* expectations)

Goodness of fit (e.g., candy striped spiders)

- We don't know what the proportions should be (without *a priori* expectations) but we know they should be the same in each group

Contingency (e.g., pigs and food)

The Chi-squared formula

$$\chi^2_{[d.f]} = \sum \frac{(O-E)^2}{E}$$

O – observed number

E – expected numbers

Σ – take the sum of

The Chi-squared formula

$$\chi^2_{[d.f]} = \sum \frac{(O-E)^2}{E}$$

The difference between what we see and what we expect to see if H_0 is true

…squared so positive

……..relative to expected value

Gets bigger as the difference increases.

Also as number of categories increase therefore d.f. matter

χ^2 Goodness of fit test

- The expected values (null hypothesis) are derived from some theory
- We test the fit of our data to the theory
- The ‘theory’ can be a uniform distribution
- In our first example the theory is Mendel’s Law (and happens to be uniform too)

χ^2 Goodness of fit test: example

- The Candy-striped spider: Striped : plain is 1:1
  - 63 offspring

Observed	28	35
Expected	31.5	31.5

χ^2 Goodness of fit test: example

At least two ways to conduct in R.

1. By coding the formula
2. By using the inbuilt function

We’ll do both; you can use either.

χ^2 Goodness of fit test: example

1. By coding the formula
   a) Observed values


Observed	28	35
expected	31.5	31.5

```
##################################################
# CHI-SQUARED BY CODING THE FORMULA              #
##################################################

# the observed data
obs <- c(28, 35)

# total number of observations
total <- sum(obs)
```

χ^2 Goodness of fit test: example

1. By coding the formula
   b) Expected values

Observed	28	35
expected	31.5	31.5

```
# calculated the expected values
# the H0 is for a 1:1 ratio
# i.e., half the total in each
exp <- c(total / length(obs), total / length(obs))
# I've used length(obs) rather than 2
# because it makes the code more reusable
```

χ^2 Goodness of fit test: example

1. By coding the formula
   c) Code the formula

$$\chi^2_{[d.f]} = \sum \frac{(O-E)^2}{E}$$

Observed	28	35
expected	31.5	31.5

```
# code the formula
chi <- sum(((obs - exp)^2) / exp)
# [1] 0.7777778
```

χ^2 Goodness of fit test: example

1. By coding the formula
   d) Find the probability of getting a χ^2 of 0.778 or more extreme (bigger)


Observed	28	35
expected	31.5	31.5

```
# look up the probability of getting a chi squared
# of 0.778 or more extreme (bigger)
#
# the degrees of freedom are the number of
# categories minus 1
df <- length(obs) - 1
pchisq(chi, df = df, lower.tail = FALSE)
# [1] 0.3778216
```

χ^2 Goodness of fit test: example

Conclusion

- $\chi^2 = 0.78$; *d.f.* = 1; $p = 0.38$
  - $p > 0.05$, therefore the test is not significant
  - Results are consistent with a 1:1 ratio

"There was no significant difference between the observed and the expected ratio."

χ^2 Goodness of fit test: example

Conclusion

- IF you had $\chi^2 = 4.6$; *d.f.* = 1; $p = 0.032$
  - $p < 0.05$ therefore the test is significant
  - Results are NOT consistent with a 1:1 ratio

"There was a significant difference between the observed and expected ratio ($\chi^2 = 4.6$; *d.f.* = 1; $p = 0.032$)." ✓

"There were significantly more xxxx and fewer xxxx than expected ($\chi^2 = 4.6$; *d.f.* = 1; $p = 0.032$)." ✓ ✓ includes direction

χ^2 Goodness of fit test: example

1. By using the inbuilt function

Observed	28	35
expected	31.5	31.5

```
###################################################
# CHI-SQUARED BY CODING THE FORMULA               #
###################################################
# we can use the same obs vector
chisq.test(obs)

#       Chi-squared test for given probabilities
#
# data:  obs
# X-squared = 0.77778, df = 1, p-value = 0.3778
```

χ^2 Goodness of fit test: example

But what to use?? What you prefer but….

1. By coding the formula

   Useful when your expected are derived from a more complex theory/idea (e.g., poisson distribution, binomial distribution) or you need to alter the d.f.

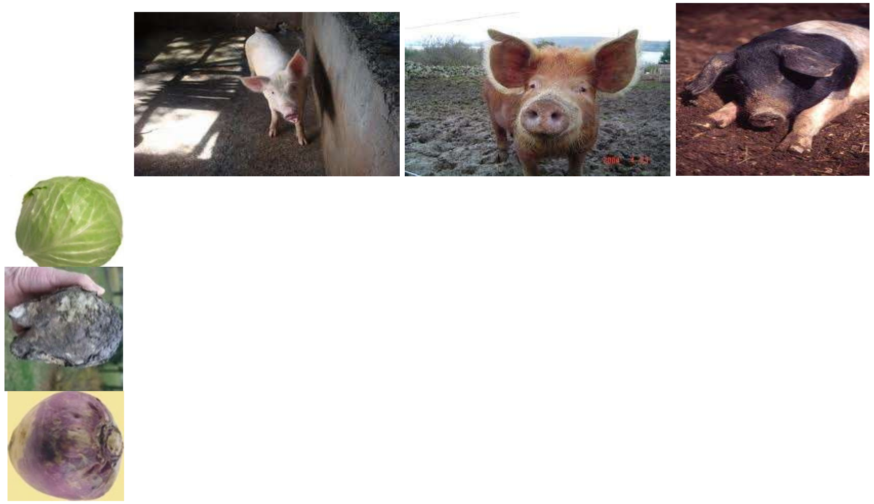
2. By using the inbuilt function

   Easy when the ratio is 1:1, 1:1:1, 1:1:1 etc

   But take care – other H_0 must be specified

χ^2 Contingency test

- Food choice by pig breeds
  - We don't know what proportions are expected but do expect it to be same for each breed



- Null hypothesis: proportion of foods taken by each breed is the same, *i.e.*, no association between breed and food type

χ^2 Contingency test: example

The Data

	Welsh 	Tamworth 	Essex 	
cabbage 	11	19	22	**52**
sugarbeet 	21	16	8	**44**
swede 	7	12	11	**30**
	39	**47**	**41**	**127**

Expected values are derived from the data

Overall pref for cabbage = 52/127

We expect (the H_0)same for each breed

χ^2 Contingency test example

Where do the expected values come from?

	Welsh	Tamworth	Essex	
cabbage	11	19	22	**52**
sugarbeet	21	16	8	**44**
swede	7	12	11	**30**
	38	**47**	**41**	**127**

Overall preference for cabbage = 45/127

Thus: Exp no. of welsh preferring cabbage = 52/127 * 38 = 15.97

Exp no. of tamworth preferring cabbage 52/127 * 47 =19.24

Exp no. of essex preferring cabbage 52/127 * 41 = 16.79

RULE: Expected number for each cell:

Row total * Column total / Overall total

χ^2 Contingency test example

Where do the expected values come from?

Wow, that's a pain!

R to the rescue!

@allison_horst

χ^2 Contingency test example

R's inbuilt function will do that!

First, add the data

```
# create the data
food_pref <- matrix(c(11, 19, 22,
                      21, 16, 8,
                      7, 12, 11),
                    nrow = 3)

#       [,1] [,2] [,3]
# [1,]    11   21    7
# [2,]    19   16   12
# [3,]    22    8   11
```

Note: this is the only time we'll use a matrix datatype – we normally use dataframes.

χ^2 Contingency test example

It’s helpful to name the rows and columns

```
# make a list object to hold two vectors
# a list is useful because the vectors can be
# of different lengths
vars <- list(breed = c("welsh",
                       "tamworth",
                       “essex"),
             food = c("cabbage",
                      "sugarbeet",
                      "swede"))
food_pref <- matrix(c(11, 19, 22,
                      21, 16, 8,
                      7, 12, 11),
                    nrow = 3,
                    dimnames = vars)
```

And this is partly why! Dataframes always have named columns.

χ^2 Contingency test example

Now we have…

```
#            food
# breed       cabbage sugarbeet swede
#   welsh          11        21     7
#   tamworth       19        16    12
#   essex          22         8    11
```

Run the inbuilt test

```
chisq.test(food_pref)

#          Pearson's Chi-squared test
#
# data:  food_pref
# X-squared = 10.64, df = 4, p-value = 0.03092
```

χ^2 Contingency test: example

degrees of freedom

- Degrees of freedom are not number of categories – 1 but

  (rows - 1)(cols - 1) = 2 * 2 = 4

- $\chi^2{}_{[4]} = 10.64$

χ^2 Contingency test

Conclusion

- Thus the test is significant (we reject the null hypothesis)
- Conclude: evidence of a preference for particular foods by different breeds

- But in what way? ("direction of effect")

  Who likes what?

χ^2 Contingency test

Conclusion

In what way – examine the observed and expected values.

Observed:

```
#             food
# breed        cabbage sugarbeet swede
#   welsh           11        21     7
#   tamworth        19        16    12
#   essex           22         8    11
```

Expected:

```
chisq.test(food_pref)$expected
#             food
# breed        cabbage sugarbeet swede
#   welsh      14.47619  13.87302  9.650794
#   tamworth 17.90476  17.15873 11.936508
#   essex      15.61905  14.96825 10.412698
```

χ^2 Contingency test

Conclusion

Higher than expected
Less than 1 different
Lower than expected

Direction of deviations; size of deviation

Observed:

```
#              food
# breed         cabbage  sugarbeet  swede
#    welsh           11         21      7
#    tamworth        19         16     12
#    essex           22          8     11
```

Expected:

```
chisq.test(food_pref)$expected
#              food
# breed         cabbage  sugarbeet  swede
#    welsh     14.47619   13.87302   9.650794
#    tamworth  17.90476   17.15873  11.936508
#    essex     15.61905   14.96825  10.412698
```

χ^2 Contingency test

Conclusion

Different pig breeds showed a significant preference for the different food types (χ^2 = 10.64; *d.f.* = 4; *p* = 0.031) with Essex much preferring cabbage and disliking sugarbeet, Tamworth showing a small preference for Cabbage and Welsh showing a strong preferencing for sugarbeet.

```
#            food
# breed       cabbage  sugarbeet  swede
#   welsh          11         21      7
#   tamworth       19         16     12
#   essex          22          8     11
```

Summary

Two types of scenario thus two types of χ^2 test

- Goodness of fit
  - We know what the proportions should be (known as *a priori* expectations); fit to a theory or distribution
  - Single row/column of observations. One explanatory
- Contingency
  - We don't know what the proportions should be (without *a priori* expectations) but we know they should be the same in each
  - At least 2 x 2. Two explanatory variables

Learning objectives for the week

By actively following the lecture and practical and carrying out the independent study the successful student will be able to:

- recognise when to use chi-squared Goodness of Fit and Contingency tests (MLO 2)
- be able to carry out, interpret and report scientifically both types of test in R (MLO 3 and 4)

Follow up from last week's practical

- Independent study: seal myoglobin exercise Live demo.
- And why ggplot rocks!