17C

Laboratory & Professional Skills:
Data Analysis

Emma Rand
Data Analysis in R

Chi-squared tests

Summary of this week

Analysing data that are counts falling into mutually exclusive categories’

Two types of chi-squared test and the difference between these two tests.

In the workshop we will discover how to apply chi-squares to data in both frequency tables and in a more raw state (i.e., tabulate the data yourself)

Learning objectives for the week

By the end of this week the successful student should be able to :

- Explain the principles of chi-squared Goodness of Fit and Contingency tests and know when each can be applied (MLO 2)
- Apply and interpret Goodness of fit and Contingency chi-squared tests R on data in frequency tables and in raw format (MLO 3 and 4)
- Summarise and illustrate with appropriate R figures test results scientifically (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.

The null hypothesis might be:

Same number in each category

1:1

smooth	wrinkled
13	17

1:1:1:1

red	blue	green	yellow
8	11	12	9

1:1:1:1:1:1:1

Mon	Tues	Wed	Thurs	Fri	Sat	Sun
23	29	11	14	20	15	21

Goodness of fit

The null hypothesis might be:

Follow a particular pattern

1:3

smooth	wrinkled
13	58

1:3:3:9

Yellow plain	Yellow spot	Red plain	Red spot
4	14	12	30

Goodness of fit

The null hypothesis might be:

match the pattern in another group ?

But same as

	blue	green	yellow
Group A	11	12	9
Group B	3	12	9

Contingency

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

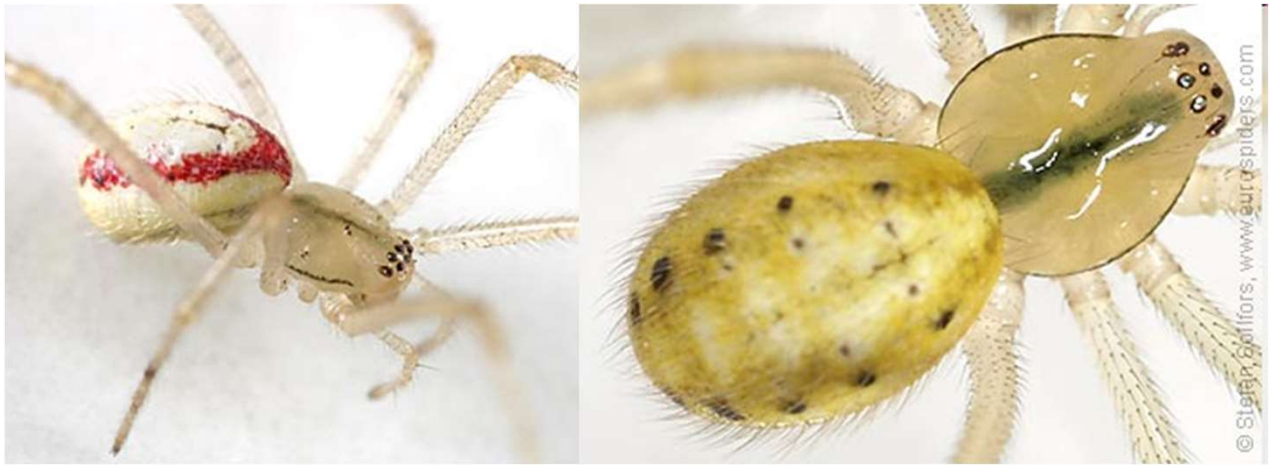
  Goodness of fit

  - match the pattern in another group

  Contingency
- 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

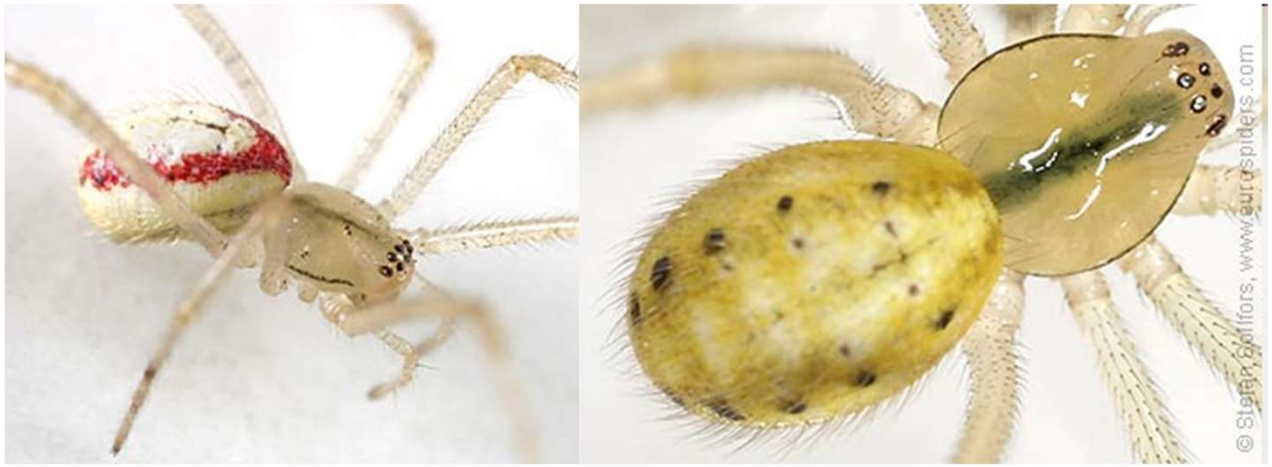
χ^2 Goodness of fit test

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)

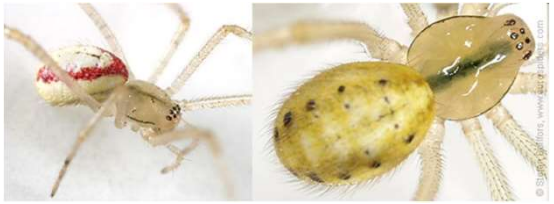
Our two example scenarios

- 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



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

```
# calculate 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 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

Summary

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
- Inbuilt `chisq.test()` function is good for 1:1:1 expectations, otherwise specify `prob`
- For complex cases, calculate expected values and code the test.

χ^2 Contingency test

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

Welsh Tamworth Essex

cabbage

sugarbeet

swede

- 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:

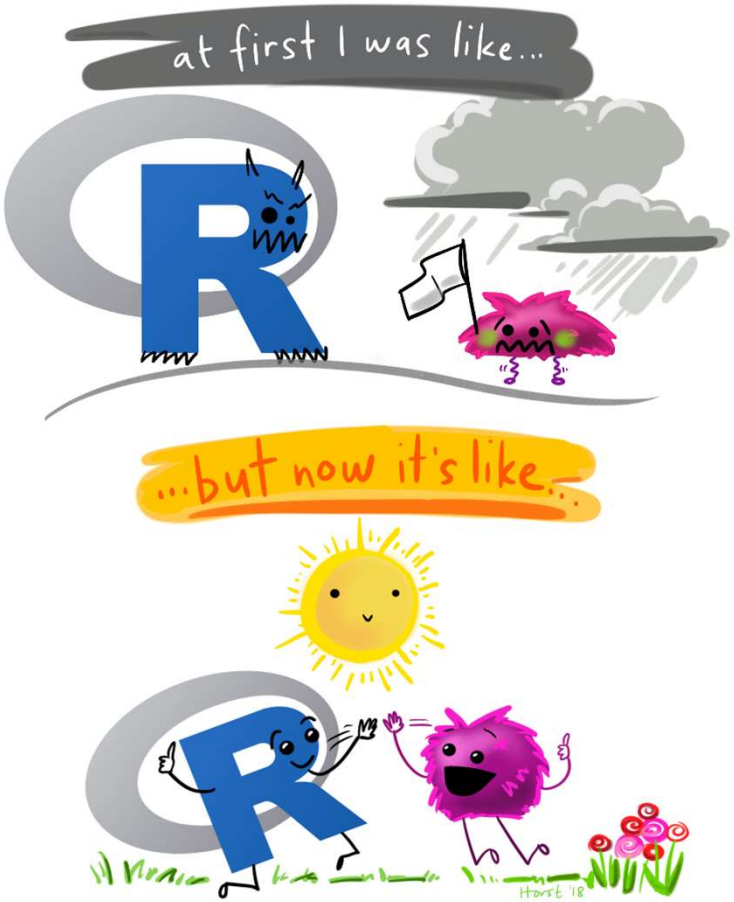
Row total * Column total / Overall total

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, 21, 7,
                      19, 16, 12,
                      22, 8, 11),
                    nrow = 3,
                    byrow = TRUE)

     [,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
# in a list the vectors can be of different lengths
vars <- list(breed = c("welsh",
                       "tamworth",
                       “essex"),
             food = c("cabbage",
                      "sugarbeet",
                      "swede"))
food_pref <- matrix(c(11, 21, 7,
                      19, 16, 12,
                      22, 8, 11),
                    nrow = 3,
                    byrow = TRUE,
                    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    15.96850  13.81890  9.212598
#   tamworth 19.24409  16.65354 11.102362
#   essex    16.78740  14.52756  9.685039
```

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    15.96850  13.81890  9.212598
#   tamworth 19.24409  16.65354 11.102362
#   essex    16.78740  14.52756  9.685039
```

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, Welsh showing a strong preferencing for sugarbeet and a dislike of cabbage and Tamworth showing no clear preference.

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

Summary

Contingency

- We don't know what the proportions should be (without *a priori* expectations) but we know they should be the same in each case
- At least 2 x 2 table
  - Two explanatory
- Inbuilt `chisq.test()` function is good