Master Sciences Pour l'Environnement

Parcours Gestion de l'Environnement et Écologie Littorale

Analyse de données / Data analysis

Partie 3 / Part 3

B. Simon-Bouhet

La Rochelle Université

First semester

What will we talk about?…

7. Fitting probability models to frequency data
   - The googdness-of-fit test
   - Goodness-of-fit test with only two categories
   - Fitting the binomial distribution
   - Random in space and time: the Poisson distribution

8. Contingency analysis
   - The χ^2 contingency test
   - Fisher's exact test
   - G tests

9. Inference for a normal Population
   - Central limit theorem
   - The t-distribution for sample mean
   - The one-sample t-test
   - Assumptions of the one-sample t-test

7. Fitting probability models to frequency data

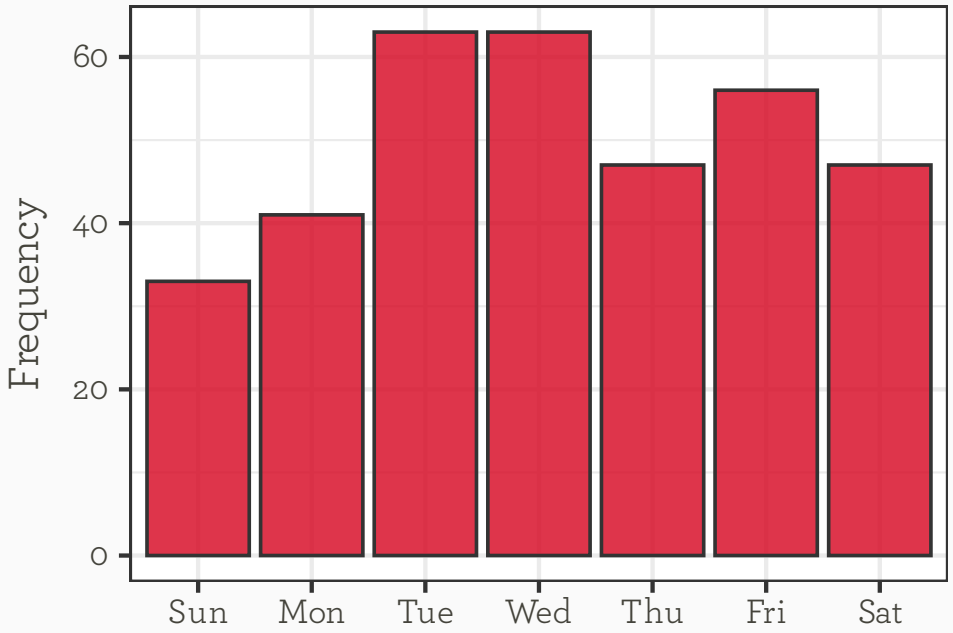
Outline

Some data to work with

Frequency of births along the week

Definition

The proportional model is a simple probability model in which the frequency of occurence of events is proportional to the number of opportunities



Day	Number of births
Sunday	33
Monday	41
Tuesday	63
Wednesday	63
Thursday	47
Friday	56
Saturday	47
Total	350

Data from Ventura *et al.* (2001).

χ^2 goodness-of-fit test

Null and alternative hypotheses

Definition

The χ^2 goodness-of-fit test compares frequency data to a probability model stated by the null hypothesis.

1. State the hypotheses

- H_0: the probability of birth is the same on every day of the week.
- H_A: the probability of birth is not the same on every day of the week.

χ^2 goodness-of-fit test

Observed and expected frequencies

For each day of the week, we count the number of occurences during the year 1999, and we calculate the expected frequencies of births under the null hypothesis of a proportional model:

Day	Number of days in 1999	Proportion of days in 1999	Expected frequency of births
Sunday	52	52/365	49.863
Monday	52	52/365	49.863
Tuesday	52	52/365	49.863
Wednesday	52	52/365	49.863
Thursday	52	52/365	49.863
Friday	53	53/365	50.822
Saturday	52	52/365	49.863
Sum	365	1	350

χ^2 goodness-of-fit test

The χ^2 test statistic

Definition

The χ^2 statistic measures the discrepancy between observed frequencies from the data and expected frequencies from the null hypothesis.

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i}$$

χ^2 goodness-of-fit test

The χ^2 test statistic

2. Compute the test statistic

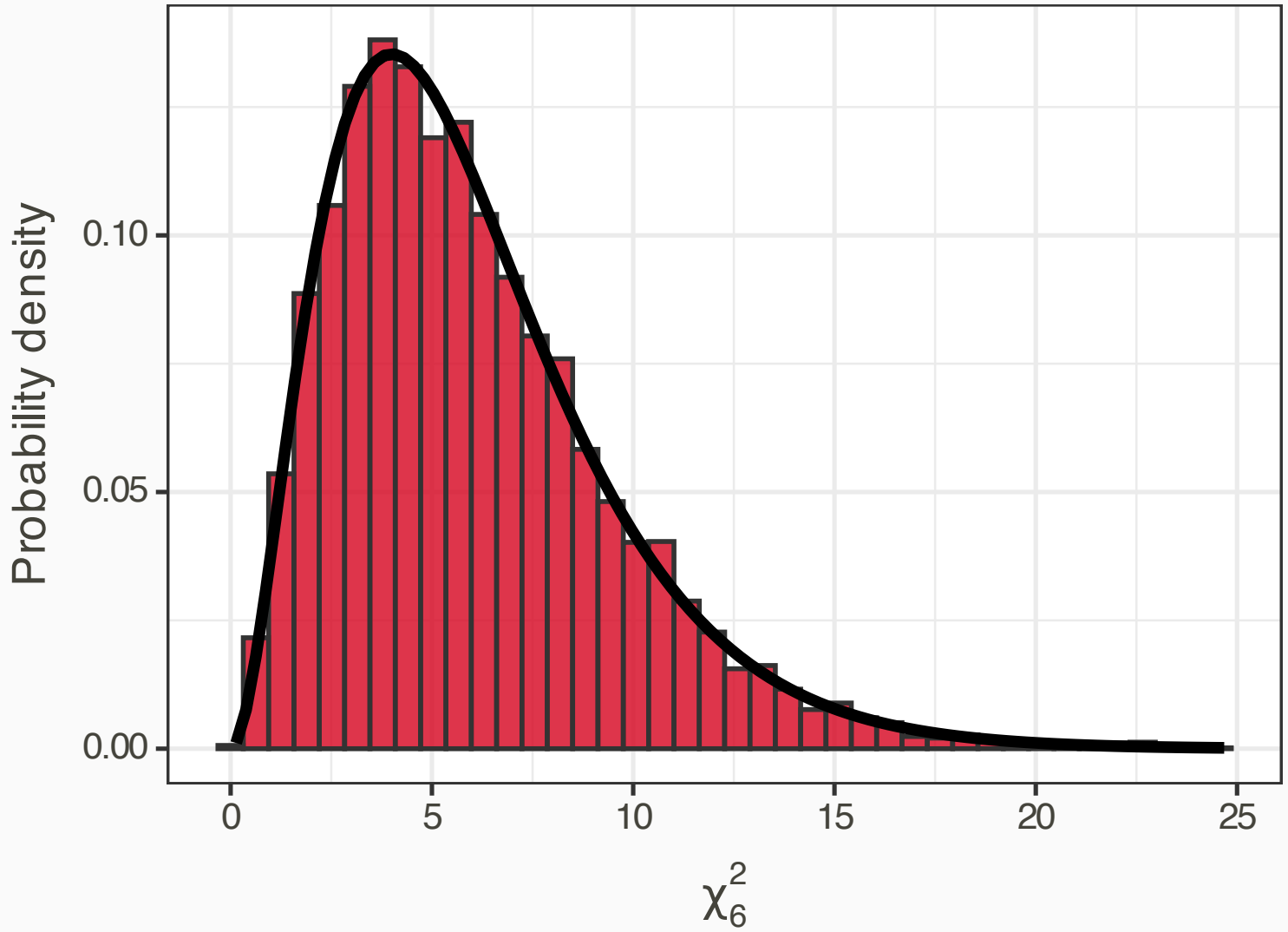
So here we have the following frequencies:

Day	Observed frequency of births	Expected frequency of births
Sunday	33	49.863
Monday	41	49.863
Tuesday	63	49.863
Wednesday	63	49.863
Thursday	47	49.863
Friday	56	50.822
Saturday	47	49.863
Sum	350	350

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} = 15.05$$

χ^2 goodness-of-fit test

The sampling distribution of χ^2 under the null hypothesis



χ^2 goodness-of-fit test

df: degrees of freedom

Definition

The number of degrees of freedom (abbreviated df) of a χ^2 statistic specifies which χ^2 distribution to use as the null distribution.

How is it calculated?

$df = (\text{nb of categories}) - 1 - (\text{nb of parameters estimated from the data})$

Here, $df = 7 - 1 = 6$.

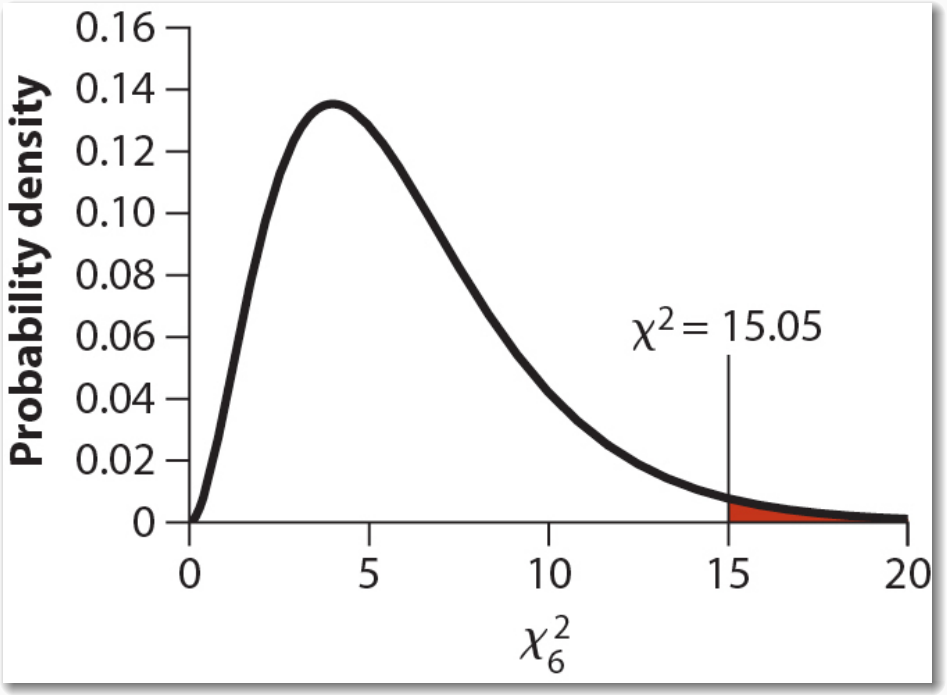
This tells us that we need to compare the value of our test statistic calculated from the data ($\chi^2 = 15.01$) to the χ^2_6 distribution with 6 degrees of freedom.

χ^2 goodness-of-fit test

3. Calculating the P-value

Important

For the χ^2 goodness-of-fit test, the P-value is the probability of getting a χ^2 value greater than the observed χ^2 value calculated from the data, simply by chance (i.e. if H_0 were true).



χ^2 goodness-of-fit test

3. Calculating the P-value

So how do we find the P-value from this?

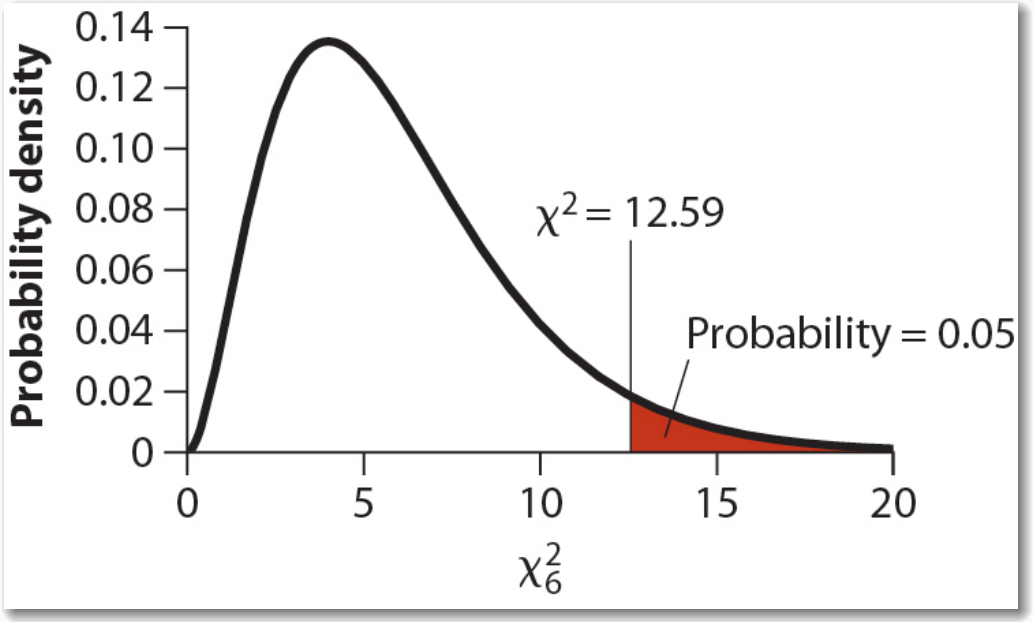
- We read a critical value corresponding to the significance level α in a table of χ^2 critical values.
- Or, we use R!

χ^2 distribution — critical values

	α							
ν	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.02	0.05	0.10	5.99	7.38	9.21	10.60	13.82
3	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.87	1.24	1.64	**12.59**	14.45	16.81	18.55	22.46
7	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12

χ^2 goodness-of-fit test

3. Calculating the P-value



Definition

The critical value read in the table is the value of a test statistic that marks the boundary of a specified area in the tail (or tails) of the sampling distribution under H_0.

$$P = Pr(\chi^2_6 \geq 15.05) < 0.05$$

χ^2 goodness-of-fit test

3. Calculating the P-value

Actually, we can say a bit more than that using the table:

χ^2 distribution — critical values

	α							
ν	0.99	0.975	0.95	0.05	0.025	0.01	0.005	0.001
1	0.00016	0.00098	0.00393	3.84	5.02	6.63	7.88	10.83
2	0.02	0.05	0.10	5.99	7.38	9.21	10.60	13.82
3	0.11	0.22	0.35	7.81	9.35	11.34	12.84	16.27
4	0.30	0.48	0.71	9.49	11.14	13.28	14.86	18.47
5	0.55	0.83	1.15	11.07	12.83	15.09	16.75	20.52
6	0.87	1.24	1.64	12.59	14.45	16.81	18.55	22.46
7	1.24	1.69	2.17	14.07	16.01	18.48	20.28	24.32
8	1.65	2.18	2.73	15.51	17.53	20.09	21.95	26.12

So we can conclude that:

$$0.01 < P < 0.025$$

χ^2 goodness-of-fit test

3. Calculating the P-value

With R, it is much more simple:

```
(birthDayTable <- table(birthDay$day_short))


Sun Mon Tue Wed Thu Fri Sat
 33  41  63  63  47  56  47

chisq.test(birthDayTable, p = c(52, 52, 52, 52, 52, 53, 52)/365)


^^IChi-squared test for given probabilities

data:  birthDayTable
X-squared = 15.057, df = 6, p-value = 0.01982
```

χ^2 goodness-of-fit test

Drawing conclusions and assumptions of the test

4. Draw the appropriate conclusion

Since our P-value ($P = 0.020$) is smaller than α, we reject the null hypothesis H_0 and conclude that births are not equally distributed over the days of the week.

Important

Assumptions of the χ^2 goodness-of-fit test:

- None of the categories should have an expected frequency less than one.
- No more than 20% of the categories should have expected frequencies less than five.

Outline

Some data to work with

Gene content of the human X chromosome

The Human Genome Project (Hubbard *et al.*, 2005):

- 781 genes on the X chromosome out of 20290 genes found so far in the entire genome
- The X chromosome represent 5.2% of the DNA content of the whole genome

Under the proportional model, we would expect 5.2% of the genes to be on the X chromosome. Is this what we observe?

1. State the hypotheses

- H_0: the percentage of human genes on the X chromosome is 5.2%
- H_A: the percentage of human genes on the X chromosome is not 5.2%

Frequencies and assumptions

Gene content of the human X chromosome

2. Check the assumptions

Chromosome	Observed	Expected
X	781	1055
Not X	19509	19235
Total	20290	20290

It would be difficult to use the binomial test, because we would have to calculate:

$$P = 2 \times Pr(X \leq 781)$$

with

$$Pr(X \leq 781) = Pr(X = 0) + Pr(X = 1) + \cdots + Pr(X = 781)$$

Perform the test

Gene content of the human X chromosome

By hand, it is impossible to perform the binomial test. But is it easy to perform the χ^2 goodness-of-fit test to compute the critical χ^2 value, and thus, get the answer we seek.

In R, both tests are equally easy to perform, and both tests lead to the same conclusion:

Perform the test

Gene content of the human X chromosome

By hand, it is impossible to perform the binomial test. But is it easy to perform the χ^2 goodness-of-fit test to compute the critical χ^2 value, and thus, get the answer we seek.

In R, both tests are equally easy to perform, and both tests lead to the same conclusion:

```
chisq.test(c(781, 19509), p = c(1055, 19235)/20290)


^^IChi-squared test for given probabilities

data:  c(781, 19509)
X-squared = 75.065, df = 1, p-value < 2.2e-16
```

$P \ll \alpha$: we reject H_0.

Perform the test

Gene content of the human X chromosome

By hand, it is impossible to perform the binomial test. But is it easy to perform the χ^2 goodness-of-fit test to compute the critical χ^2 value, and thus, get the answer we seek.

In R, both tests are equally easy to perform, and both tests lead to the same conclusion:

```
binom.test(x = 781, n = 20290, p = 0.052)


^^IExact binomial test

data:  781 and 20290
number of successes = 781, number of trials = 20290,
p-value < 2.2e-16
alternative hypothesis: true probability of success is not equal to 0.052
95 percent confidence interval:
 0.03588645 0.04123056
sample estimates:
probability of success
            0.03849187
```

Outline

Some data to work with

The composition of two-child families

Data from Rodgers & Doughty (2001).

Number of boys	Observed number of families
0	530
1	1332
2	582
Total	2444

- H_0: the number of boys in two-child families has a binomial distribution.
- H_A: the number of boys in two-child families does not have a binomial distribution.

We can test the fit of a binomial distribution to the observed data with a χ^2 goodness-of-fit test.

Computing the expected frequencies

The composition of two-child families

We have $2444 \times 2 = 4888$ children in this study.

The number of boys is $530 \times 0 + 1332 \times 1 + 582 \times 2 = 2496$

Hence, the estimated probability of a child being a boy is:

$$\hat{p} = \frac{2496}{4888} = 0.5106$$

Next, we use this value of $\hat{p}$ and the binomial distribution with $n = 2$ to calculate the expected probabilities under the null hypothesis:

```
dbinom(x = 0:2, size = 2, prob = 0.5106)
```

```
[1] 0.2395124 0.4997753 0.2607124
```

Computing the expected frequencies

The composition of two-child families

Finally, we are able to compute the expected frequencies of families by multiplying the probabilities by the total number of families:

Number of boys	Observed number of families	Probability under H_0	Expected number of families
0	530	0.23951	585.3
1	1332	0.49978	1221.4
2	582	0.26071	637.3
Total	2444	1.00000	2444.0

From here, we can perform the χ^2 goodness-of-fit test:

```
chisq.test(c(530, 1332, 582), p = c(585.3, 1221.4, 637.3)/2444)


^^IChi-squared test for given probabilities

data:  c(530, 1332, 582)
X-squared = 20.038, df = 2, p-value = 4.454e-05
```

P-value and conclusion

The composition of two-child families

Here, the correct number of df is:

$$df = 3 - 1 - 1 = 1$$

To get the correct P-value, we need to use both the test statistic obtained earlier (the χ^2 value from the test) and the correct df:

```
1 - pchisq(20.038, df = 1)
```

```
[1] 7.591843e-06
```

Outline

The Poisson distribution

Definition

The Poisson distribution describes the number of successes in blocks of time or space, when successes happen independently of each other and occur with equal probability at every instant in time or point in space.

3 types of spatial distribution of points:

The Poisson distribution

Mass extinctions

The way we calculate probabilities with the Poisson distribution is:

$$Pr(X \text{ successes}) = \frac{e^{-\mu}\mu^{X}}{X!}$$

where μ is the mean number of independant successes in time or space (expressed as count per unit time or space).

In this example, we want to know if:

- extinctions of species occur randomly through the long fossil record of Earth's history.
- there are periods in which extinction rates are unusualy high (mass extinction) compared with background rates.

The Poisson distribution

Mass extinctions

Raup & Sepkoski (1982) studied the remains of hard shells of marine invertebrates.

76 regular blocks of time have been defined and the number of extinctions of marine families has been counted in each one of them.

Number of Extinctions (X)	Frequency
0	0
1	13
2	15
3	16
4	7
5	10
6	4
7	2
8	1
9	2
10	1
11	1

Number of Extinctions (X)	Frequency
12	0
13	0
14	1
15	0
16	2
17	0
18	0
19	0
20	1
>20	0
Total	76

The Poisson distribution

Mass extinctions

1. State the hypotheses

- H_0: The number of extinctions per time interval has a Poisson distribution.
- H_A: The number of extinctions per time interval does not have a Poisson distribution.

In order to compute the expected frequencies under H_0 using the formula for the Poisson distribution:

$$Pr(X \text{ successes}) = \frac{e^{-\mu}\mu^X}{X!}$$

we need to know the value of μ. Since we don't, we calculate it's unbiased estimate, the mean number of extinctions per block of time ($\overline{X}$):

$$\overline{X} = \frac{(0 \times 0) + (13 \times 1) + (15 \times 2) + \cdots}{76} = 4.21$$

The Poisson distribution

Mass extinctions

We can now calculate the probability of any number of successes n, from 0 to 20 or more, using:

$$Pr(n \text{ successes}) = \frac{e^{-\overline{X}}\overline{X}^n}{n!}$$

In R:

```
expectedProportion <- dpois(0:20, lambda = 4.210526)
expectedProportion
```

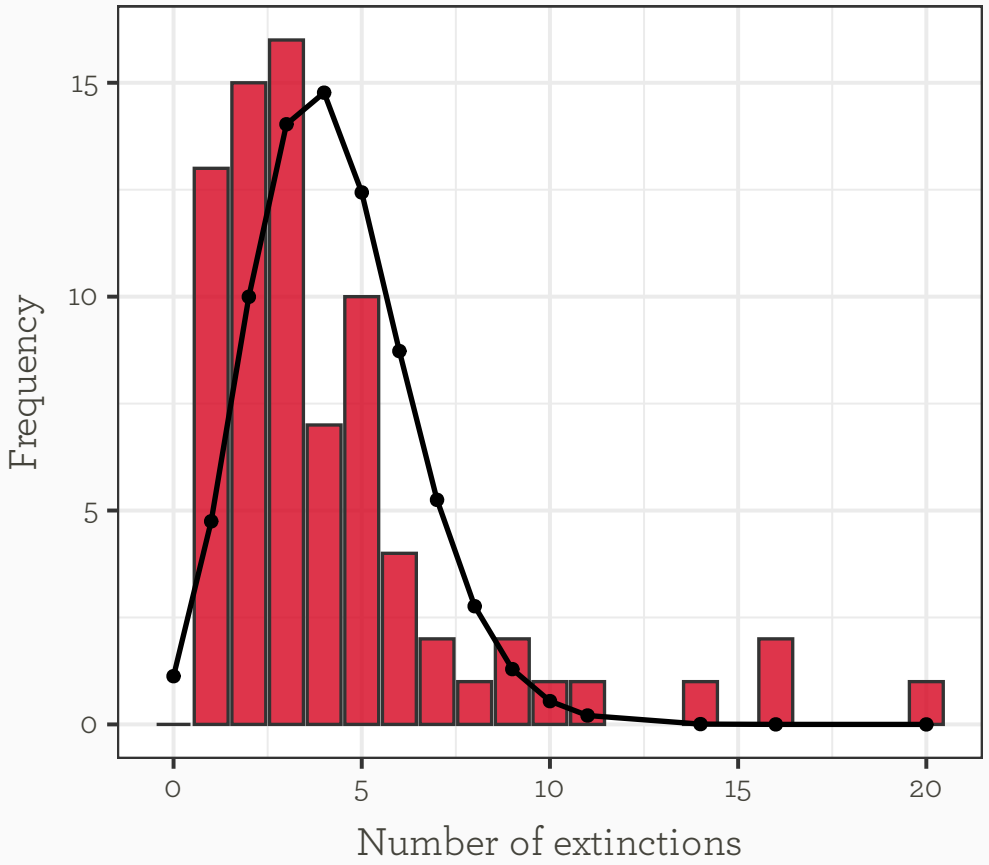
```
 [1] 1.483856e-02 6.247815e-02 1.315329e-01 1.846076e-01
 [5] 1.943238e-01 1.636411e-01 1.148358e-01 6.907418e-02
 [9] 3.635483e-02 1.700811e-02 7.161307e-03 2.741170e-03
[13] 9.618139e-04 3.115187e-04 9.368982e-05 2.629889e-05
[17] 6.920761e-06 1.714120e-06 4.009638e-07 8.885623e-08
[21] 1.870657e-08
```

The Poisson distribution

Mass extinctions

The expected frequencies are obtained by multiplying the probabilities by 76:

Number of extinctions (X)	Observed frequency of time intervals	Expected frequency of time intervals
0	0	1.13
1	13	4.75
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.73
7	2	5.25
8	1	2.76
9	2	1.29
≥ 10	6	0.86
Total	76	76.00



The Poisson distribution

Mass extinctions

Number of extinctions (X)	Observed frequency of time intervals	Expected frequency of time intervals
0 or 1	13	5.88
2	15	10.00
3	16	14.03
4	7	14.77
5	10	12.44
6	4	8.73
7	2	5.25
$\geq$8	9	4.91
Total	76	76.00

Now that we have the observed and expected frequencies, we can perform the χ^2 goodness-of-fit test.

The Poisson distribution

Mass extinctions

```
obsFreqGroup

   0 or 1         2         3         4         5         6         7 8 or more
       13        15        16         7        10         4         2         9

expFreqGroup

   0 or 1         2         3         4         5         6         7 8 or more
 5.876068  9.996501 14.030177 14.768608 12.436722  8.727524  5.249639  4.914761
```

```
chisq.test(obsFreqGroup, p = expFreqGroup/76)


^^IChi-squared test for given probabilities

data:  obsFreqGroup
X-squared = 23.95, df = 7, p-value = 0.001163
```

How many df do we have here?

The Poisson distribution

Mass extinctions

Since we used $\overline{X}$ to estimate μ from the data in order to compute the expected frequencies under H_0, we lost a df.

Thus,

$$df = 8 - 1 - 1 = 6$$

We then compute the true P-value:

```
1 - pchisq(23.95, df = 6)
```

```
[1] 0.0005334336
```

The Poisson distribution

Mass extinctions

Important

A Poisson distribution has its mean μ equal to its variance σ^2.

For an observed frequency distribution:

- If the variance is greater than the mean, then the distribution is clumped.
- If the variance is smaller than the mean, then the distribution is dispersed.

Hence, the ratio $\frac{s^2}{\overline{X}}$ is a measure of clumping or dispersion.
In our case, the variance is greater than the mean:

$$\overline{X} = 4.21 \qquad s^2 = 13.72 \qquad \frac{s^2}{\overline{X}} \gg 1$$

We conclude that the distribution of extinction events in time is highly clumped: extinctions tend to occur in bursts (mass extinctions) rather than randomly or evenly in time.

8. Contengency analysis

Outline

The χ^2 contingency test

Trematodes, snails, fishes and birds

Definition

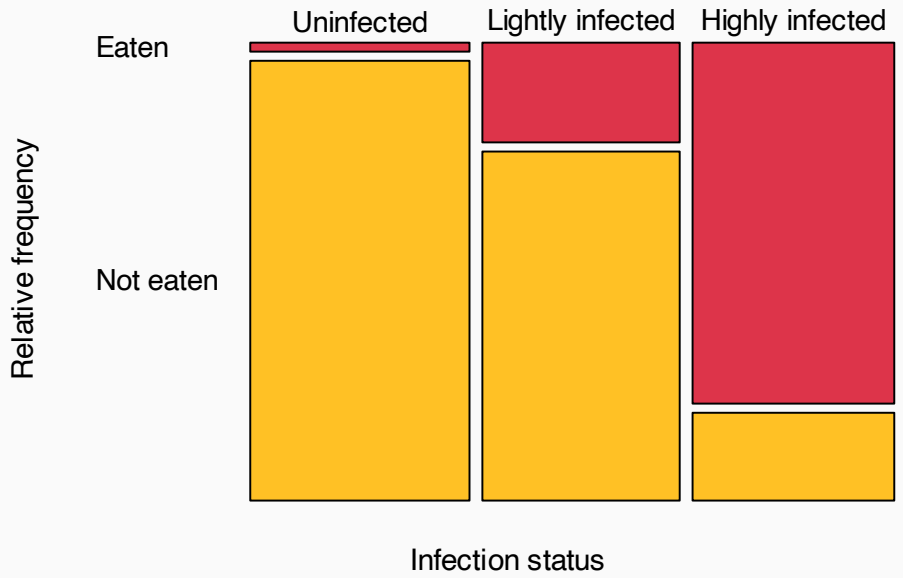
The χ^2 contingency test is the most commonly used test of association between two categorical variables. It tests the goodness of fit to the data of the null model of independance of variables.

Life cycle of the trematodes *Euhaplorchis californiensis* (Lafferty & Morris, 1996):

The χ^2 contingency test

Trematodes, snails, fishes and birds

	Uninfected	Lightly infected	Highly infected	Row *total*
Eaten by birds	1	10	37	*48*
Not eaten by birds	49	35	9	*93*
Column total	*50*	*45*	*46*	*141*



The χ^2 contingency test

Trematodes, snails, fishes and birds

1. State the hypotheses

- H_0: Parasite infection and being eaten are independent.
- H_A: Parasite infection and being eaten are not independent.

2. Compute the test statistic

To compute the χ^2, we need the expected frequencies under H_0:

	Uninfected	Lightly infected	Highly infected	Row *total*
Eaten by birds	17.0	15.3	15.7	*48*
Not eaten by birds	33.0	29.7	30.3	*93*
Column total	*50*	*45*	*46*	*141*

$$\chi^2 = \sum_i \frac{(\text{Observed}_i - \text{Expected}_i)^2}{\text{Expected}_i} \qquad i = \text{cells of the contingency table}$$

The χ^2 contingency test

Trematodes, snails, fishes and birds

Here $\chi^2 = 69.5$

The number of degrees of freedom is

$$df = (\text{nb of rows } - 1)(\text{nb of columns} - 1) = 2$$

In R:

```
head(worm)
```

```
# A tibble: 6 x 2
  infection        fate
  <fct>            <fct>
1 Uninfected       Eaten
2 Lightly infected Eaten
3 Lightly infected Eaten
4 Lightly infected Eaten
5 Lightly infected Eaten
6 Lightly infected Eaten
```

```
dim(worm)
```

```
[1] 141   2
```

The χ^2 contingency test

Trematodes, snails, fishes and birds

In R, we don't need to compute the expected frequencies by hand:

```
chisq.test(worm$fate, worm$infection, correct = FALSE)


^^IPearson's Chi-squared test

data:  worm$fate and worm$infection
X-squared = 69.756, df = 2, p-value = 7.124e-16
```

However, we can print the expected values from the `chisq.test()` function to check the assumptions of the test:

```
chisq.test(worm$fate, worm$infection, correct = FALSE)$expected

           worm$infection
worm$fate   Uninfected Lightly infected Highly infected
  Eaten       17.02128          15.31915         15.65957
  Not eaten   32.97872          29.68085         30.34043
```

Here, $P \ll \alpha$. We reject H_0.

Outline

Fisher's exact test

The feeding habits of vampire bats

Definition

Fisher's exact test examines the independence of two categorical variables even with small expected values.

The common vampire bat *Desmodus rotundus*, data from Turner (1975).

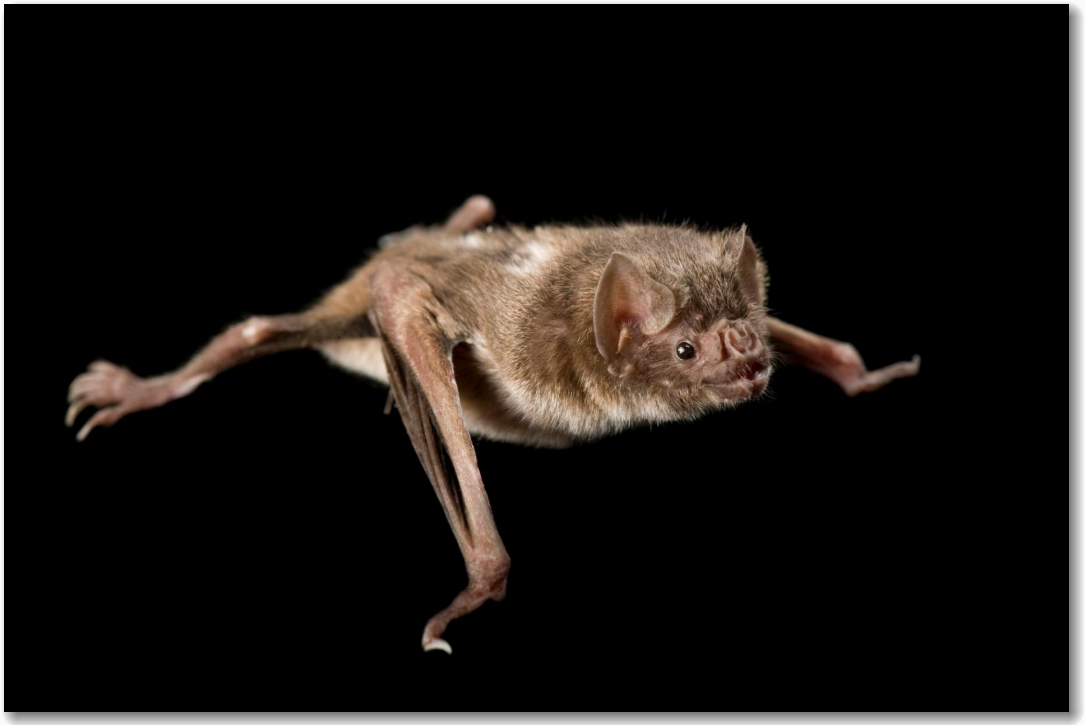


Photo: National Geographic

Fisher's exact test

The feeding habits of vampire bats

	Cows in estrus	Cows not in estrus	Row *total*
Bitten by vampire bats	15	6	*21*
Not bitten by vampire bats	7	322	*329*
Column total	*22*	*328*	*350*

- H_0: State of estrus and vampire bats attacks are independent.
- H_A: State of estrus and vampire bats attacks are not independent.

```
vampire <- read_csv("data/chap09e5VampireBites.csv")
vamp.test <- chisq.test(vampire$bitten, vampire$estrous)$expected
vamp.test
```

```
              vampire$estrous
vampire$bitten estrous no estrous
    bitten        1.32      19.68
    not bitten   20.68     308.32
```

Fisher's exact test

The feeding habits of vampire bats

```
summary(vampire)
```

```
   estrous             bitten
 Length:350         Length:350
 Class :character   Class :character
 Mode  :character   Mode  :character
```

```
dim(vampire)
```

```
[1] 350   2
```

```
vampireTable <- table(vampire$bitten, vampire$estrous)
vampireTable
```

```
             estrous no estrous
  bitten          15          6
  not bitten       7        322
```

Fisher's exact test

The feeding habits of vampire bats

```
fisher.test(vampire$bitten, vampire$estrous)


^^IFisher's Exact Test for Count Data

data:  vampire$bitten and vampire$estrous
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  29.94742 457.26860
sample estimates:
odds ratio
  108.3894
```

Once again, we find that $P \ll \alpha$, hence we reject H_0.

Fisher's exact test

The feeding habits of vampire bats

```
fisher.test(vampireTable)


^^IFisher's Exact Test for Count Data

data:  vampireTable
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  29.94742 457.26860
sample estimates:
odds ratio
  108.3894
```

Once again, we find that $P \ll \alpha$, hence we reject H_0.

Quantifying the strength of an association

A few words about odds and odds ratio

Definition

The odds of success are the probability of success divided by the probability of failure.

$$\widehat{O} = \frac{\hat{p}}{1-\hat{p}}$$

For a cow, the odds of being bitten while in estrus is:

$$\widehat{O_1} = \frac{\frac{15}{22}}{1-\frac{15}{22}} = 2.1429$$

For a cow, the odds of being bitten while not in estrus is:

$$\widehat{O_2} = \frac{\frac{6}{328}}{1-\frac{6}{328}} = 0.0186 \approx \frac{1}{55}$$

Quantifying the strength of an association

A few words about odds and odds ratio

Definition

The odds ratio is the odds of success in one group divided by the odds of success in a second group. It quantifies the strength of the association between two categorical variables.

$$\widehat{OR} = \frac{\widehat{O_1}}{\widehat{O_2}} = \frac{2.1429}{0.0186} = 115$$

This means that it is on average 115 times more likely for a cow to get bitten by a vampire bat when in estrus than when not in estrus

A quicker way to compute $\widehat{OR}$ is:

$$\widehat{OR} = \frac{a \times d}{b \times c}$$

Quantifying the strength of an association

A few words about odds and odds ratio

Like any estimate, standard error and confidence intervals can be calculated for odds ratios.
The formulae involves log transformations... In practice, use R!

```
^^IFisher's Exact Test for Count Data

data:  vampireTable
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  29.94742 457.26860
sample estimates:
odds ratio
  108.3894
```

Conclusion

Vampire bats are 108.3 times more likely to bite cows in estrus than cows that are not in estrus (Fisher's exact test, $P < 0.001$, n=350, odds ratio $CI_{95\%} = [29.9; 457.3]$)

Outline

G test

An alternative to χ^2 tests

The G test is an alternative to the χ^2 goodness-of-fit test based on the principles of likelihood analysis:

$$G = 2 \sum_i \text{Observed}_i \times \ln \frac{\text{Observed}_i}{\text{Expected}_i}$$

It can be used even with small expected frequencies, but has been shown to be less accurate when sample size is small.

In R:

- function `GTest()` from package `DescTools`
- function `G.test()` from package `RVAideMemoire`

References

Hubbard T, Andrews D, Caccamo M, *et al.* (2005) Ensembl 2005. *Nucleic Acids Research*, **33**, D447–D453.

Lafferty KD, Morris AK (1996) Altered behavior of parasitized killifish increases susceptibility to predation by bird final hosts. *Ecology*, **77**, 1390–1397.

Raup DM, Sepkoski JJ (1982) Mass extinctions in the marine fossil record. *Science*, **215**, 1501–1503.

Rodgers JL, Doughty D (2001) Does having boys or girls run in the family? *Chance Magazine*, **Fall**, 8–13.

Shoemaker AL (1996) What's normal? —temperature, gender and heart rate. *Journal of Statistics Education*, **4**.

Turner DC (1975) *The vampire bat: a field study in behavior and ecology*. Johns Hopkins Press, Baltimore, MD.

Ventura SJ, Martin JA, Curtin SC, Menacker F, Hamilton BE (2001) Births: final data for 1999. *National Vital Statistics Reports*, **49**.