Formation R niveau 2

Comparing proportions

B. Simon-Bouhet

La Rochelle Université

Avril 2023

What will we talk about?…

1. Analyzing proportions
   - The binomial distribution
   - Testing a proportion: the binomial test
   - Estimating proportions

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

1. Analyzing proportions

Outline

Number of successes in a random sample

Definition

The binomial distribution provides the probability distribution for the number of "successes" in a fixed number of independent trials, when the probability of success is the same in each trial.

$$Pr(X \text{ successes}) = \binom{n}{X} \cdot p^X \cdot (1-p)^{(n-X)}$$

avec

$$\binom{n}{X} = \frac{n!}{X!(n-X)!}$$

Example: probability of observing 4 left-handed flowers in the offspring, with $n = 27$ flowers and $p = \frac{1}{4}$:

$$Pr(X = 4) = \frac{27!}{4!(27-4)!} \cdot 0.25^4 \cdot (1-0.25)^{(27-4)} = 0.09171$$

Number of successes in a random sample

Example: probability of observing 4 left-handed flowers in the offspring, with $n = 27$ flowers and $p = \frac{1}{4}$:

$$Pr(X = 4) = \frac{27!}{4!(27-4)!} \cdot 0.25^4 \cdot (1 - 0.25)^{(27-4)} = 0.09171$$

In R:

```
dbinom(x = 4, size = 27, prob = 0.25)
```

```
[1] 0.09171623
```

Number of successes in a random sample

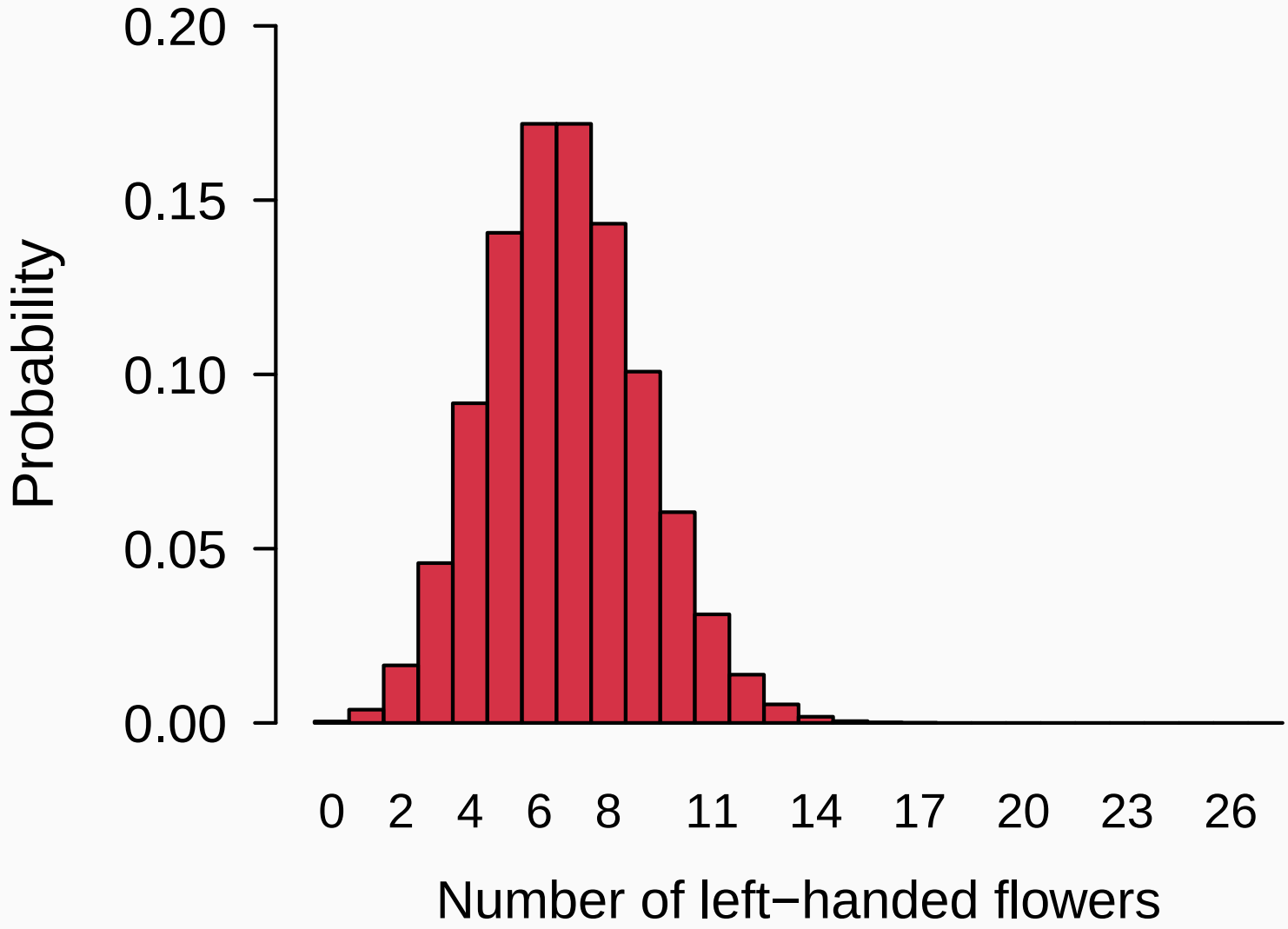
By using the same function, we get the sampling distribution of left-handed flowers in the offspring, with $n = 27$ flowers and $p = \frac{1}{4}$:

```
AllProb <- dbinom(x = 0:27, size = 27, prob = 0.25)
AllProb
```

```
 [1] 4.233057e-04 3.809751e-03 1.650892e-02 4.585812e-02
 [5] 9.171623e-02 1.406316e-01 1.718830e-01 1.718830e-01
 [9] 1.432358e-01 1.007956e-01 6.047736e-02 3.115500e-02
[13] 1.384667e-02 5.325641e-03 1.775214e-03 5.128395e-04
[17] 1.282099e-04 2.765311e-05 5.120947e-06 8.085705e-07
[21] 1.078094e-07 1.197882e-08 1.088984e-09 7.891188e-11
[25] 4.383993e-12 1.753597e-13 4.496403e-15 5.551115e-17
```

```
barplot(AllProb, names = 0:27, cex.names = 0.8, las = 1,
    col = rgb(0.8, 0, 0.1, 0.8), space = 0, ylim = c(0, 0.2),
    xlab = "Number of left-handed flowers", ylab = "Probability")
```

Number of successes in a random sample

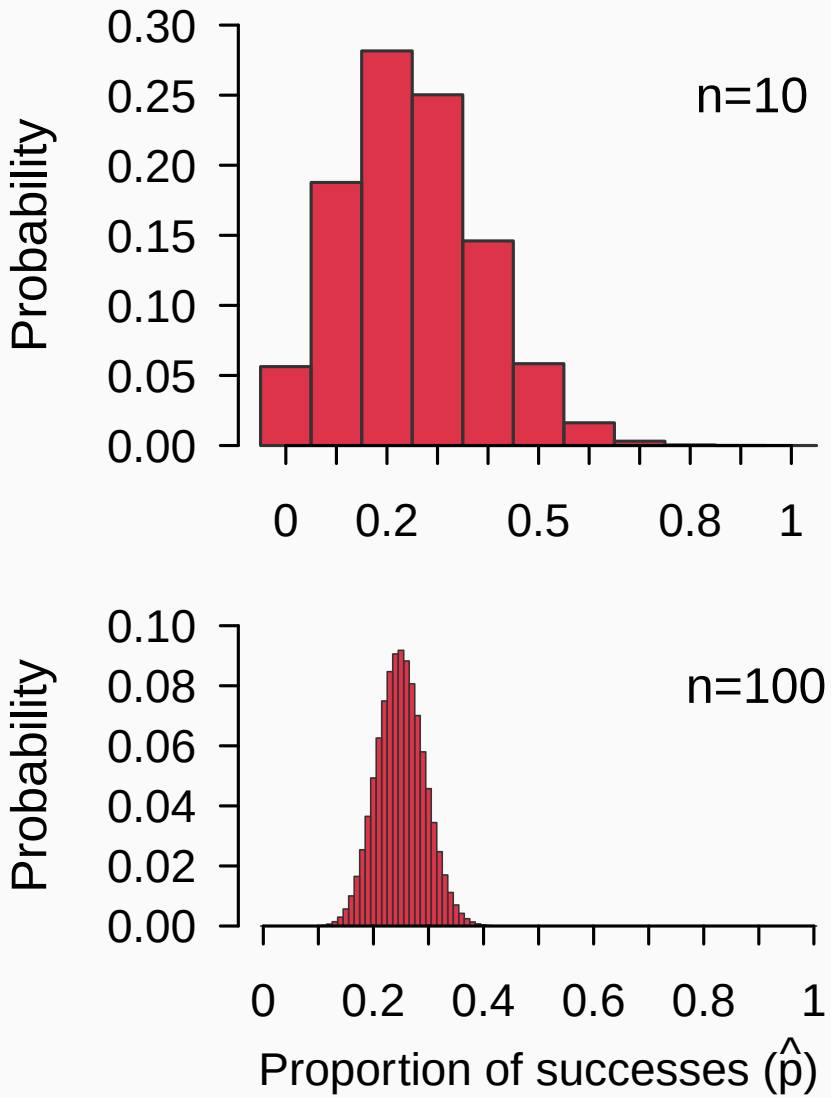


Sampling distribution of the proportion

If there are X successes out of n trials, the estimated proportion of successes is:

$$\hat{p} = \frac{X}{n}$$

We can use the same theoretical population of flowers having a true proportion of success $p = 0.25$ to illustrate the sampling distribution of the sample proportion $\hat{p}$.



Sampling distribution of the proportion

The standard error of the mean quantifies the uncertainty associated with a sample mean:

$$\mathsf{SE}_{\overline{Y}} = \frac{s}{\sqrt{n}}$$

Likewise, the standard error $\sigma_{\hat{p}}$ quantifies the uncertainty associated with an estimated proportion $\hat{p}$:

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

As for the mean, since the true value p is often unknown, the estimated standard error is calculated using the estimate of the proportion $\hat{p}$:

$$\mathsf{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Outline

The binomial test

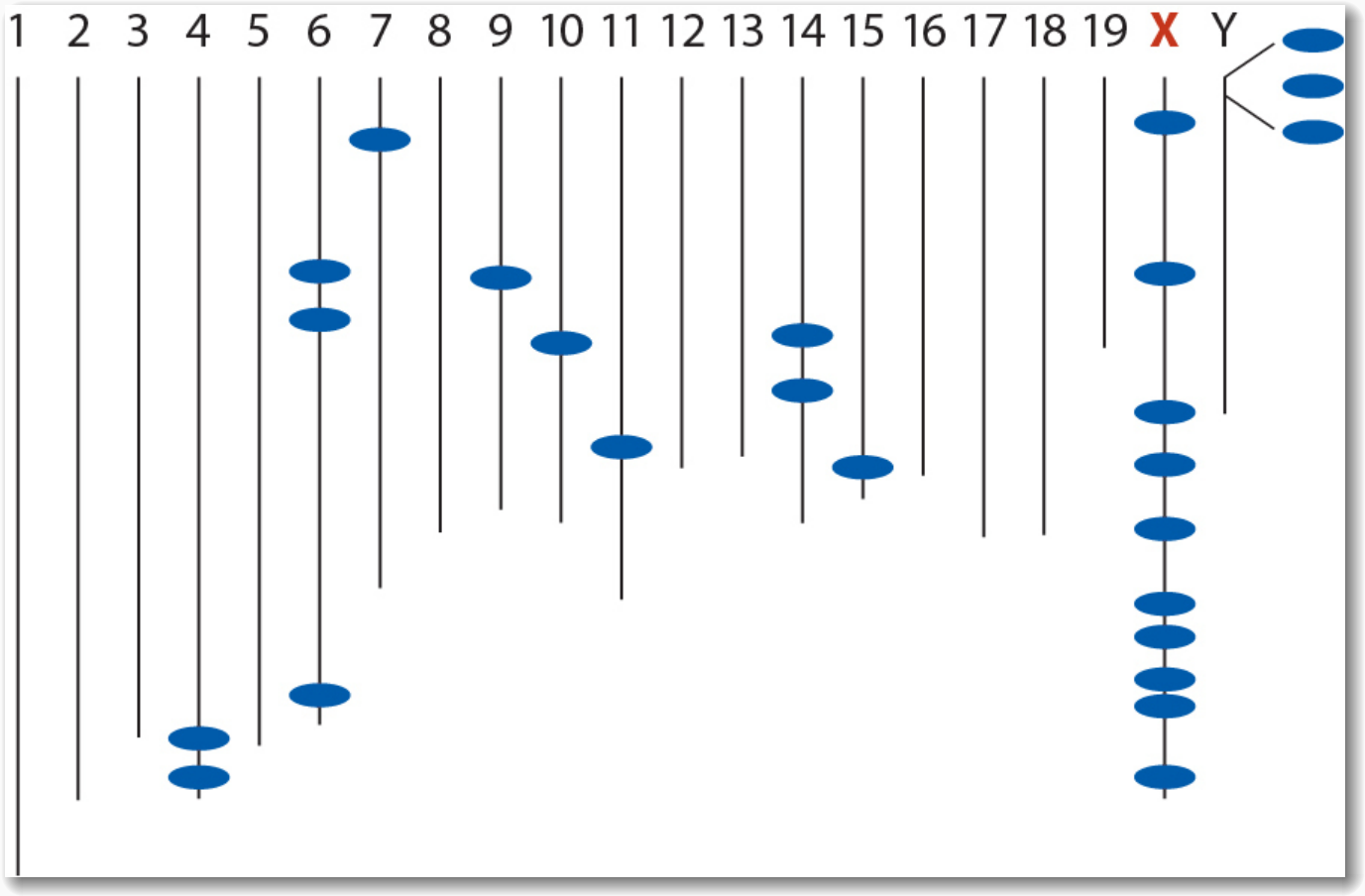
Definition

The binomial test uses data to test whether a population proportion (p) matches a null expectation (p_0) for the proportion.

- H_0: the relative frequency of successes in the population is p_0.
- H_A: the relative frequency of successes in the population is not p_0.

Some data to work with

Sex and the X chromosome



Data from Wang *et al.* (2001).
10 out of 25 genes ($\hat{p} = 0.4$) on the X chromosome.

Binomial test

Sex and the X chromosome

1. State the hypotheses

- H_0: the probability that a spermatogenesis gene falls on the X chromosome is $p = 0.061$.
- H_A: the probability that a spermatogenesis gene falls on the X chromosome is something other than 0.061 ($p \neq 0.061$).

2. Compute the test statistic

For a binomial test, the test statistic is the number of success, here, 10 spermatogenesis genes on the X chrmosome.

Binomial test

Sex and the X chromosome

3. Compute the P-value

Under the null hypothesis, the expected number of spermatogenesis genes falling on the X chromosome is:

$$p_0 \times n = 0.061 \times 25 = 1.525$$

Since we observed more than 1.525 spermatogenesis genes on the X chromosome (i.e. 10), we have to calculate the probability that 10 or more spermatogenesis genes fall on the X chromosome simply by chance, using the binomial distribution:

$$Pr(X \geq 10) = Pr(10) + Pr(11) + \cdots + Pr(25) = 9.9 \times 10^{-7}$$

The P-value is twice that number since the test is two-sided:

$$P = 2 \times Pr(X \geq 10) = 2\left(9.9 \times 10^{-7}\right) = 1.98 \times 10^{-6}$$

Binomial test

Sex and the X chromosome

3. Compute the P-value

In R, we can compute the P-values manually:

```
2*sum(dbinom(x = 10:25, size = 25, p = 0.061))
```

```
[1] 1.987976e-06
```

Binomial test

Sex and the X chromosome

3. Compute the P-value

Or we could use the built-in `binom.test()` function:

```
binom.test(x = 10, n = 25, p = 0.061)


^^IExact binomial test

data:  10 and 25
number of successes = 10, number of trials = 25,
p-value = 9.94e-07
alternative hypothesis: true probability of success is not equal to 0.061
95 percent confidence interval:
 0.2112548 0.6133465
sample estimates:
probability of success
                   0.4
```

Binomial test

Sex and the X chromosome

4. Draw the appropriate conclusion

Here, P is much smaller thant the significance level $\alpha = 0.05$.

We reject H_0. Here is what we would conclude in a report:

> ***Conclusion***
>
> "There is a disproportionate number of spermatogenesis genes on the X chromosome (0.40, SE $= 0.10$; binomial test, $n = 25$, $P < 0.001$)."

Outline

Estimating proportions

Sex and the X chromosome

Proportion:

$$\hat{p} = \frac{X}{n} = \frac{10}{25} = 0.40$$

Standard Error of proportions:

$$\mathsf{SE}_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.4(1-0.4)}{25}} = 0.10$$

Confidence interval of a proportion (Agresti-Coull method):
First, define p' as:

$$p' = \frac{X+2}{n+4}$$

Then, use the following formula:

$$p' - 1.96\sqrt{\frac{p'(1-p')}{n+4}} < p < p' + 1.96\sqrt{\frac{p'(1-p')}{n+4}}$$

Confidence intervals: a lot of methods

Sex and the X chromosome

```
library(binom)
binom.confint(x = 10, n = 25)
```

```
          method  x  n      mean     lower     upper
1  agresti-coull 10 25 0.4000000 0.2336047 0.5930338
2     asymptotic 10 25 0.4000000 0.2079635 0.5920365
3          bayes 10 25 0.4038462 0.2227432 0.5889367
4        cloglog 10 25 0.4000000 0.2128160 0.5812317
5          exact 10 25 0.4000000 0.2112548 0.6133465
6          logit 10 25 0.4000000 0.2304775 0.5974104
7         probit 10 25 0.4000000 0.2265040 0.5962745
8        profile 10 25 0.4000000 0.2247930 0.5947951
9            lrt 10 25 0.4000000 0.2247577 0.5948050
10     prop.test 10 25 0.4000000 0.2181192 0.6110970
11        wilson 10 25 0.4000000 0.2340330 0.5926054
```

Here, all methods confirm that $p_0 = 0.061$ is not among the most likely values for the true population paramter p.

2. Contengency analysis

Outline

The χ^2 contingency test

Trematodes, snails, fishes and birds

Definition

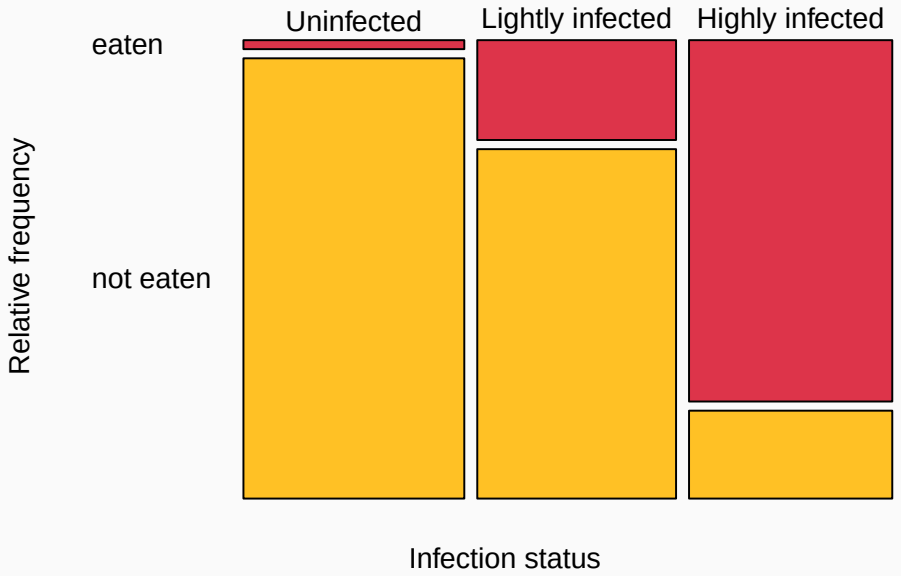
The χ^2 contingency test is the most commonly used test of association between two categorical variables. It tests the goodness of fit to the data of the null model of independance of variables.

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)
```

```
          infection  fate
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).

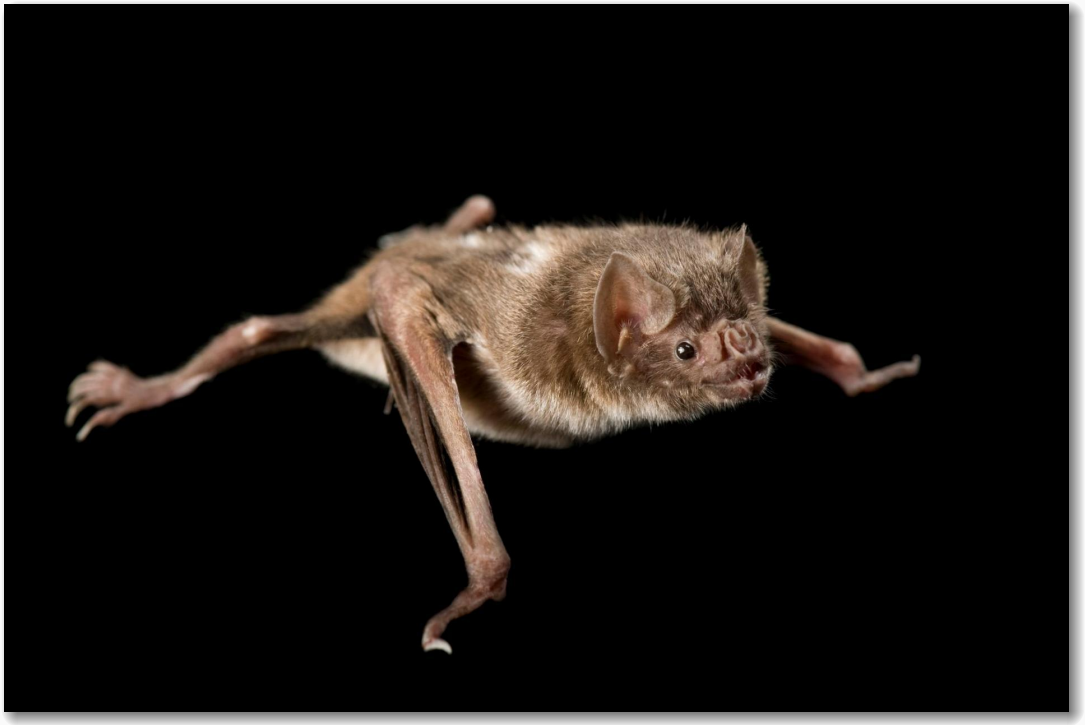


Photo: National Geographic

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

1. Analyzing proportions
   - The binomial distribution
   - Testing a proportion: the binomial test
   - Estimating proportions

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

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

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

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

Wang PJ, McCarrey JR, Yang F, Page DC (2001) An abundance of X-linked genes expressed in spermatogonia. *Nature Genetics*, **27**, 422–426.