Sciences Pour l'Environnement

Parcours Gestion de l'Environnement et Écologie Littorale

Analyse de données / Data analysis

B. Simon-Bouhet

Université de La Rochelle

First semester

6. Analyzing proportions

Outline

4. Estimating with uncertainty
   - The sampling distribution of an estimate
   - Measuring the uncertainty of an estimate
   - Confidence intervals
   - Error bars

5. Hypothesis testing
   - Making and using statistical hypotheses
   - Hypothesis testing: an example
   - Errors in hypothesis testing
   - When the null hypothesis is not rejected
   - One-sided tests
   - Hypothesis testing *vs.* confidence intervals

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

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

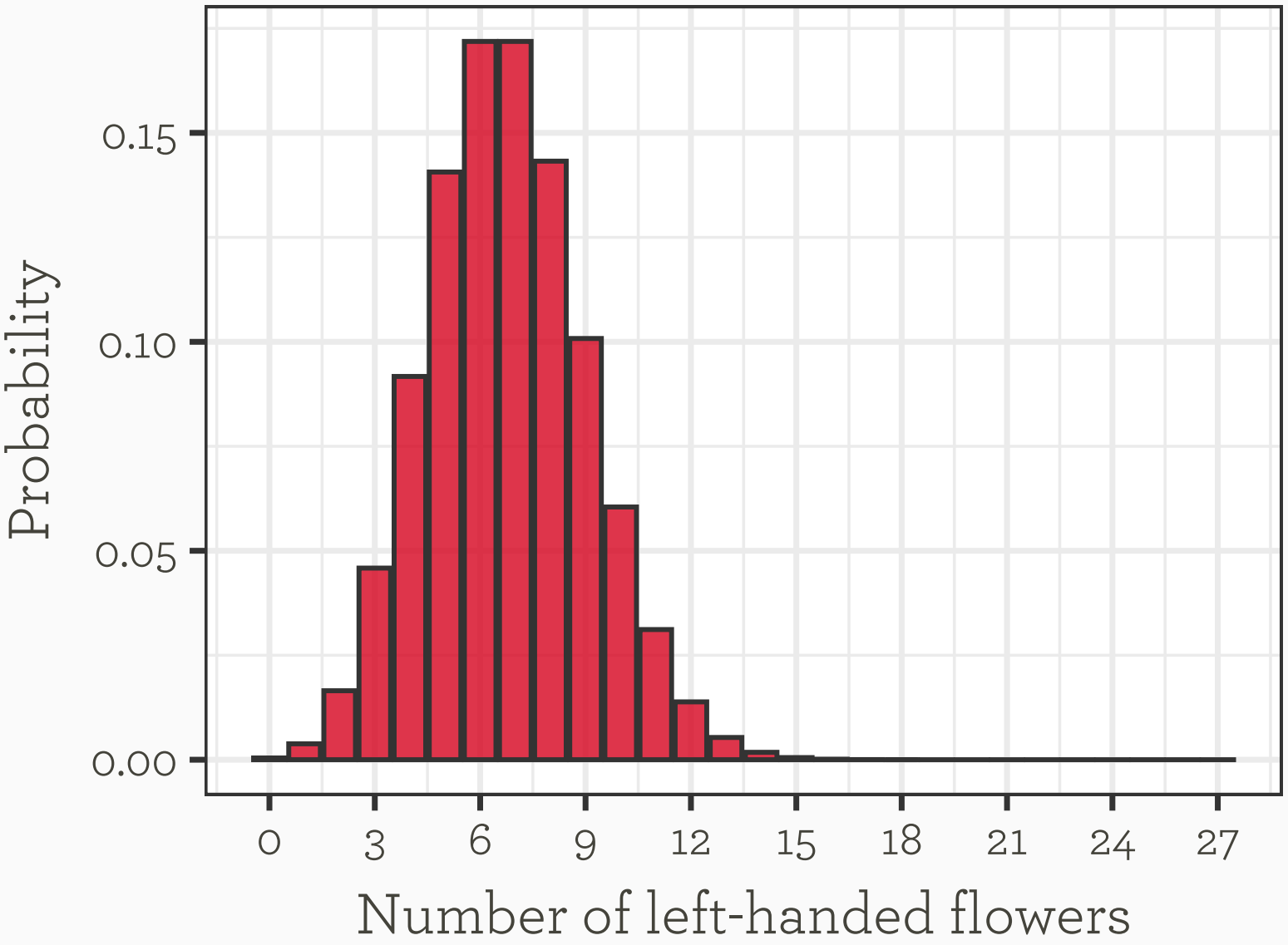
By using the same function, we get the sampling distribution of left-handed flowers in the offspring, with $n = 27$ flowers and $p = \frac{1}{4}$:

```
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
```

```
tibble(x = 0:27, AllProb) %>%
  ggplot(aes(x, AllProb)) +
  geom_col(fill = mycol, color = "grey20") +
  labs(x = "Number of left-handed flowers", y = "Probability") +
  scale_x_continuous(breaks = seq(from = 0, to = 27, by = 3))
```

Number of successes in a random sample

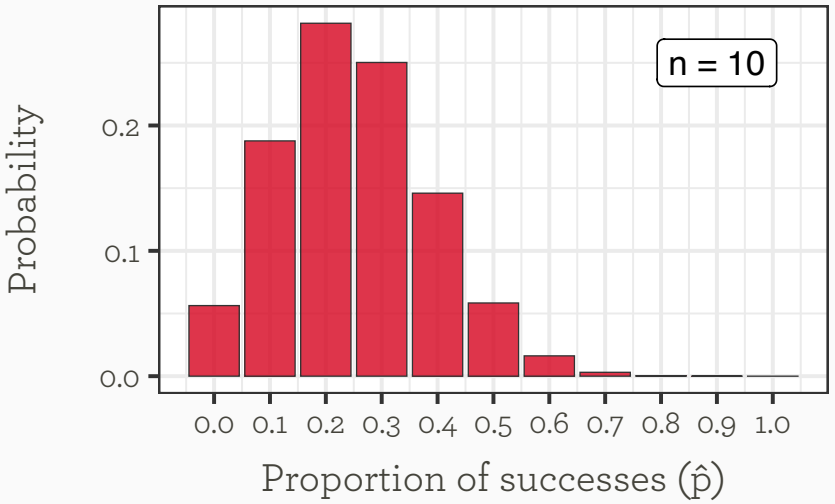


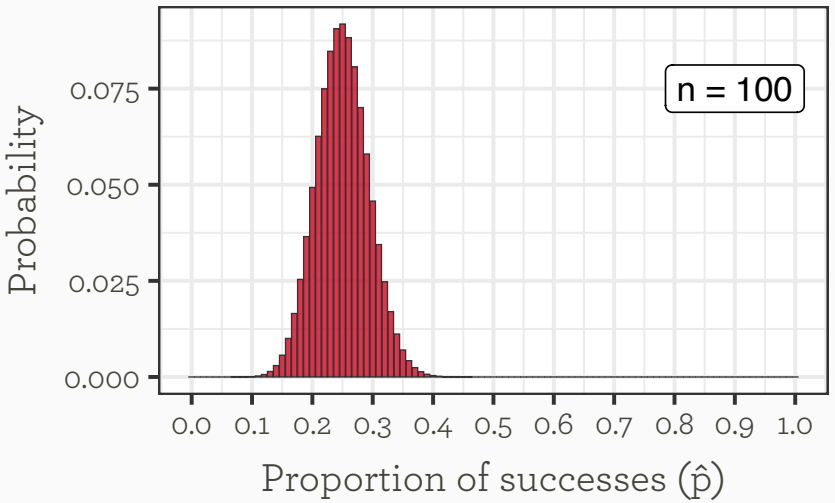
Sampling distribution of the proportion

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

4. Estimating with uncertainty
   - The sampling distribution of an estimate
   - Measuring the uncertainty of an estimate
   - Confidence intervals
   - Error bars

5. Hypothesis testing
   - Making and using statistical hypotheses
   - Hypothesis testing: an example
   - Errors in hypothesis testing
   - When the null hypothesis is not rejected
   - One-sided tests
   - Hypothesis testing *vs.* confidence intervals

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

The binomial test

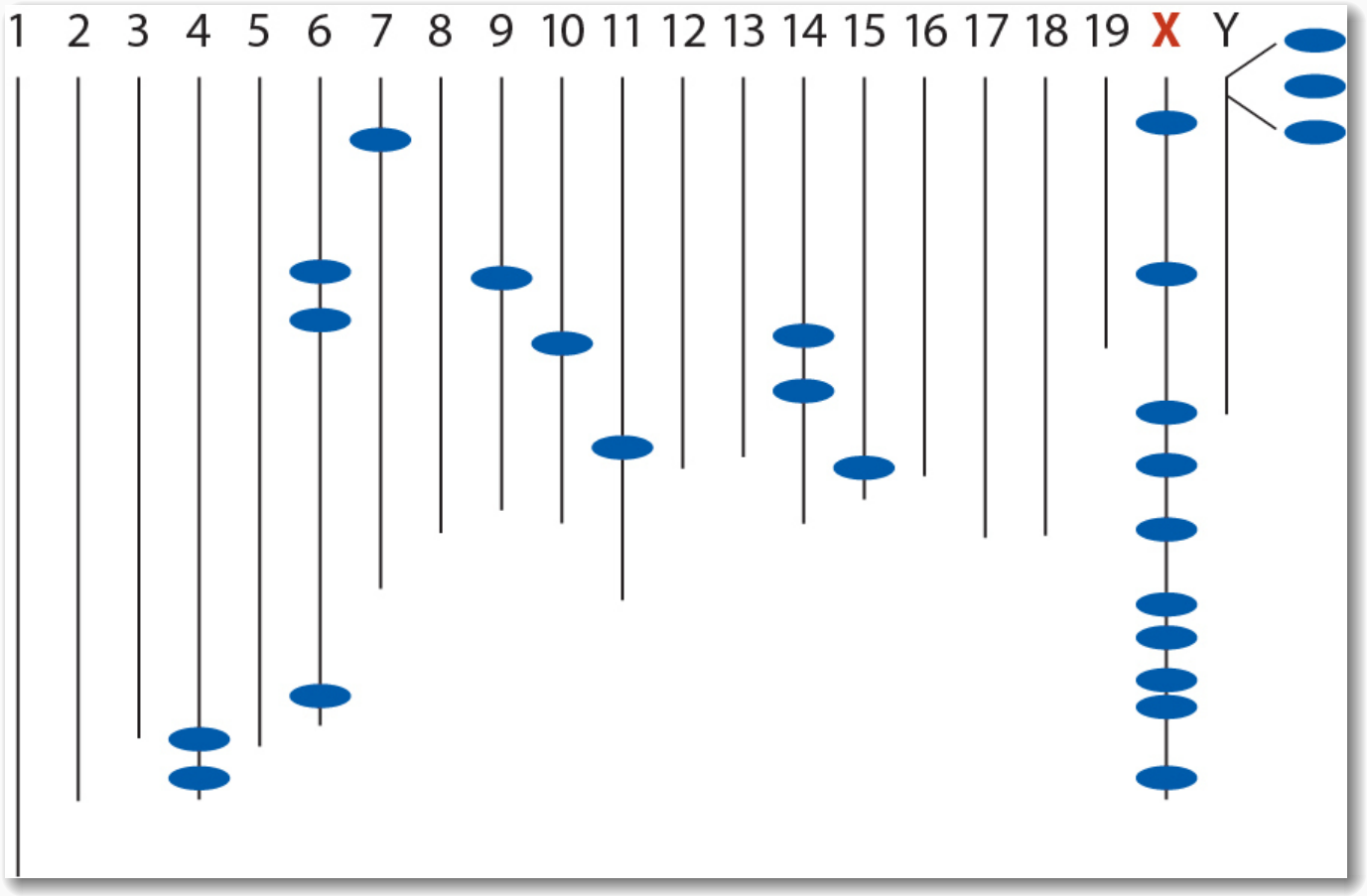
Definition

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

4. Estimating with uncertainty
   - The sampling distribution of an estimate
   - Measuring the uncertainty of an estimate
   - Confidence intervals
   - Error bars

5. Hypothesis testing
   - Making and using statistical hypotheses
   - Hypothesis testing: an example
   - Errors in hypothesis testing
   - When the null hypothesis is not rejected
   - One-sided tests
   - Hypothesis testing *vs.* confidence intervals

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

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.

References

Bisazza A, Cantalupo C, Robins A, Rogers LJ, Vallortigara G (1996) Right-pawedness in toads. *Nature*, **379**, 408.

Frick RW (1996) The appropriate use of null hypothesis testing. *Psychological Methods*, **1**, 379–390.

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

Jesson LK, Barrett SCH (2002) The genetics of mirror-image flowers. *Proceedings of the Royal Society of London, Series B: Biological Sciences*, **269**, 1835–1839.

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