Hypothesis Testing

Paul M. Magwene

Introduction

In hypothesis testing we:

- compare statistical properties of our observed data to the same properties we would expect to see under a *null hypothesis*.

More specifically:

- compare our point estimate of a statistic of interest, based on the observed data, to the *sampling distribution of that statistic* under a given null hypothesis.

Null and alternative hypotheses

In statistical hypothesis testing we must formulate a:

- "null hypothesis"
- "alternative hypothesis"

These jointly describe the possible values for a statistic of interest.

Other assumptions

We must also make some assumptions, either based on theory or inferred from the data, about the distributional properties of the sampling distribution of the statistic of interest.

Statistical hypotheses are not scientific hypotheses

- Statistical hypotheses are statistical statements about a population, not "statements about the existence and possible causes of natural phenomena" (Whitlock and Schluter)
- Can help us to determine which predictions stemming from scientific hypotheses are consistent with the data

Null hypotheses

Whitlock & Schuluter: "A **null hypothesis** is a specific statement about a population parameter made for the purpose of argument. A good null hypothesis is a statement that would be interesting to reject."

- Null hypotheses typically correspond to outcomes that would suggest "no difference" or "no effect" of the treatment, grouping, or other types of comparisons one makes with data
- Sometimes a null expectation is based on prior observation or from theoretical considerations
- A null hypothesis is always *specific*
- The standard mathematical notation to indicate a null hypothesis is to write H_0 ("H-zero" or "H-naught")

Examples of null hypotheses

- H_0: The density of dolphins is the same in areas with and without drift-net fishing
- H_0 :The effect of ACE inhibitors on blood pressure does not differ from administering a placebo
- H_0: There is no correlation between maternal smoking and the probability of premature births

Alternative hypotheses

Whitlock & Schluter: "The **alternative hypothesis** includes all other feasible values for the population parameter besides the value stated in the null hypothesis"

- Alternative hypotheses usually include parameter values that are predicted by a scientific hypothesis, but often include other feasible values as well
- The standard mathematical notation to indicate a null hypothesis is to write H_A

Examples of alternative hypotheses

- H_A: The density of dolphins differs in areas with and without drift-net fishing
- H_A :The effect of ACE inhibitors on blood pressure differs from administration of a placebo
- H_A: There is a non-zero correlation between maternal smoking and the probability of premature births

Rejecting / failing to reject null hypotheses

When carrying out statistical hypothesis testing **the null hypothesis is the only statement being tested with the data**.

- If the data are consistent with the null hypothesis, we have "failed to reject the null hypothesis". This is *not* the same as accepting the null hypothesis.
- If the data are inconsistent with the null hypothesis, we "reject the null hypothesis" and say the data support the alternative hypothesis
- Note that because the alternative hypothesis is usually formulated in terms of all other possible values of a parameter of interest, rejecting the null hypothesis does not allow us to make a probabilistic statement about the value of that parameter.

P-values

The p-value is the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis is true.

- **Small p-values** give us evidence to support the *rejection* of the null hypothesis.

Outcomes of hypothesis tests

	do not reject H_0	reject H_0
H_0 true	okay	Type 1 error (false positive), α
H_A true	Type 2 error (false negative), β	okay

Significance thresholds

It is convention when carrying out hypothesis testing to try to control the rate of false positives (i.e. the rate at which we expect to reject the null hypothesis when it is true).

We specify a "significance threshold" α that specifies the false positive rate we're willing to live with.

p-values smaller than the threshold α are "statistically significant"

Evolving views on significance thresholds

- $\alpha = 0.05$, has been the conventional significance threshold for many studies, but there is growing consensus that this is too liberal a threshold given that real world data often violates sampling and distributional assumptions that underlie conventional hypothesis testing.
- A focus on *p*-values alone ignores the magnitude of differences or effects of interest
- A number of recent studies propose that a more appropriate default convention for statistical significance should be $\alpha = 0.005$

For more in depth discussion of the debate around *p*-values see:

- Benjamin et al. 2018, Nat Hum Behav
- Ioannidis 2018, JAMA

Example: Mean tail length in possums

A prior study established that the distribution of tail length in bushtail possums in Queensland is $N(37.9, 1.71)$. A sample of 5 possums sampled from the state of Victoria gives a mean tail length of 35.9 cm. Is there evidence to suggest that Victorial possum tail lengths differ from those of Queensland possums?

Null and alternative hypotheses

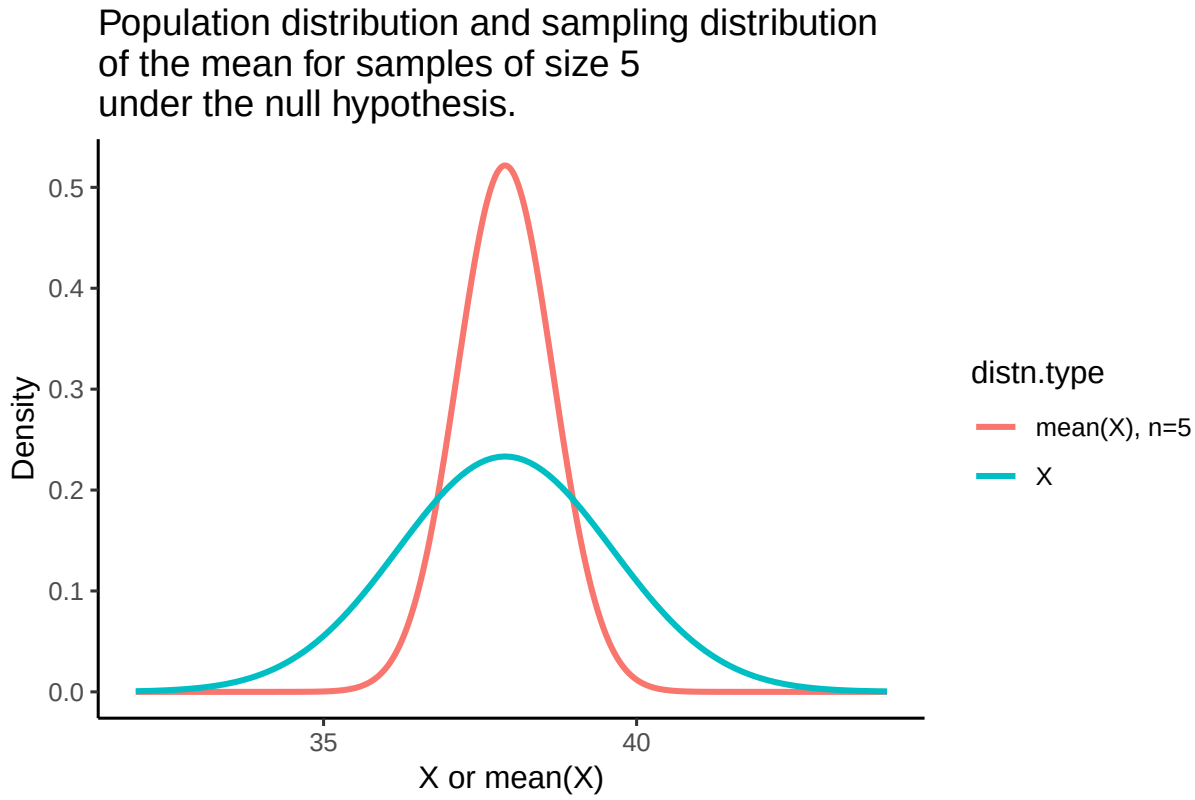
- H_0: There is no difference in mean tail lengths of Queensland and Victoria possums; i.e. $\overline{x} = 37.9$.
- H_A: There is a difference in mean tail lengths of Queensland and Victoria possums, i.e. $\overline{x} \neq 37.9$.

Mean tail length example, continued

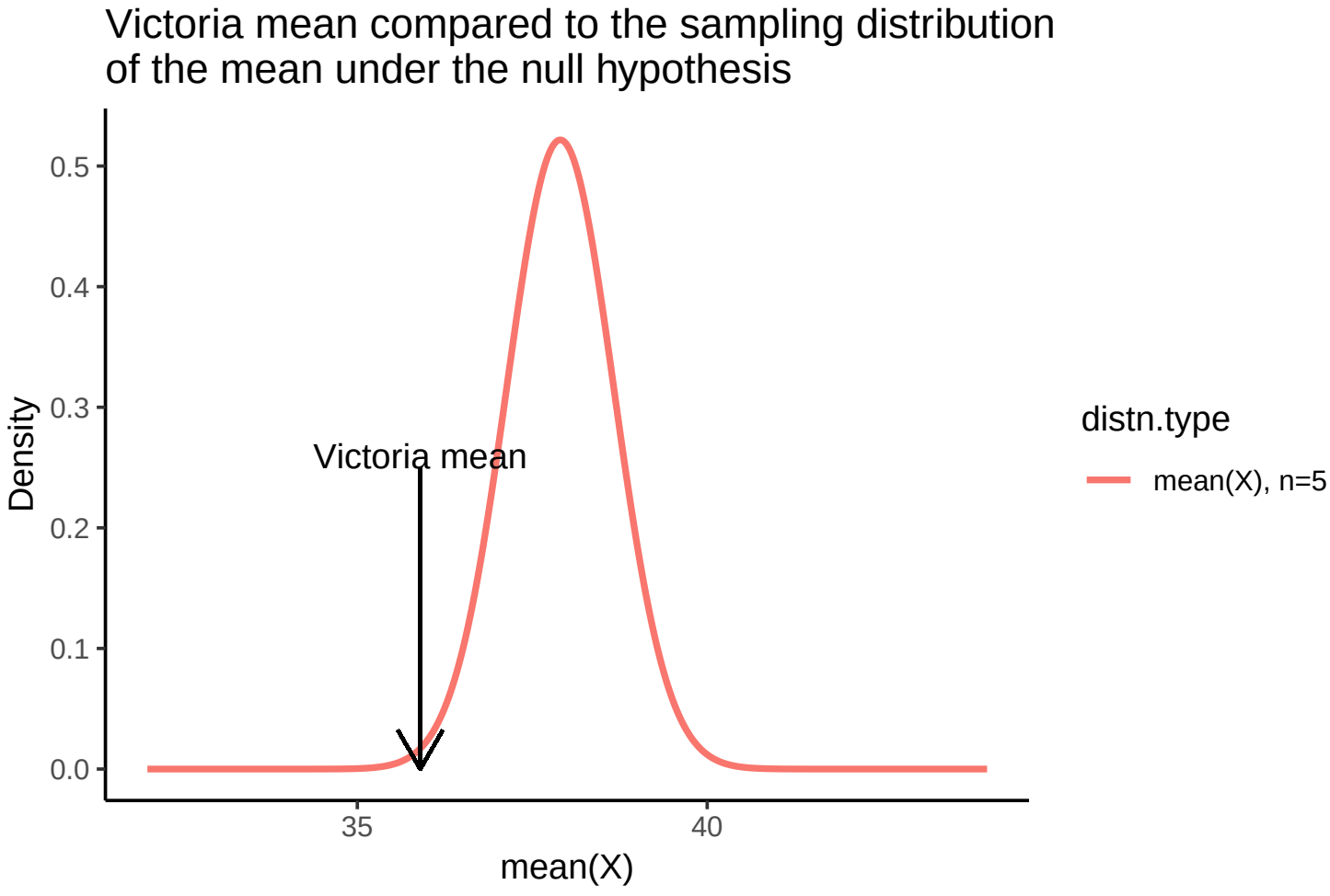
Sampling distribution of the mean

If the underlying population distribution is $N(\mu, \sigma)$, then the sampling distribution of the mean for samples of size n is $N(\mu, \sigma/\sqrt{n})$.

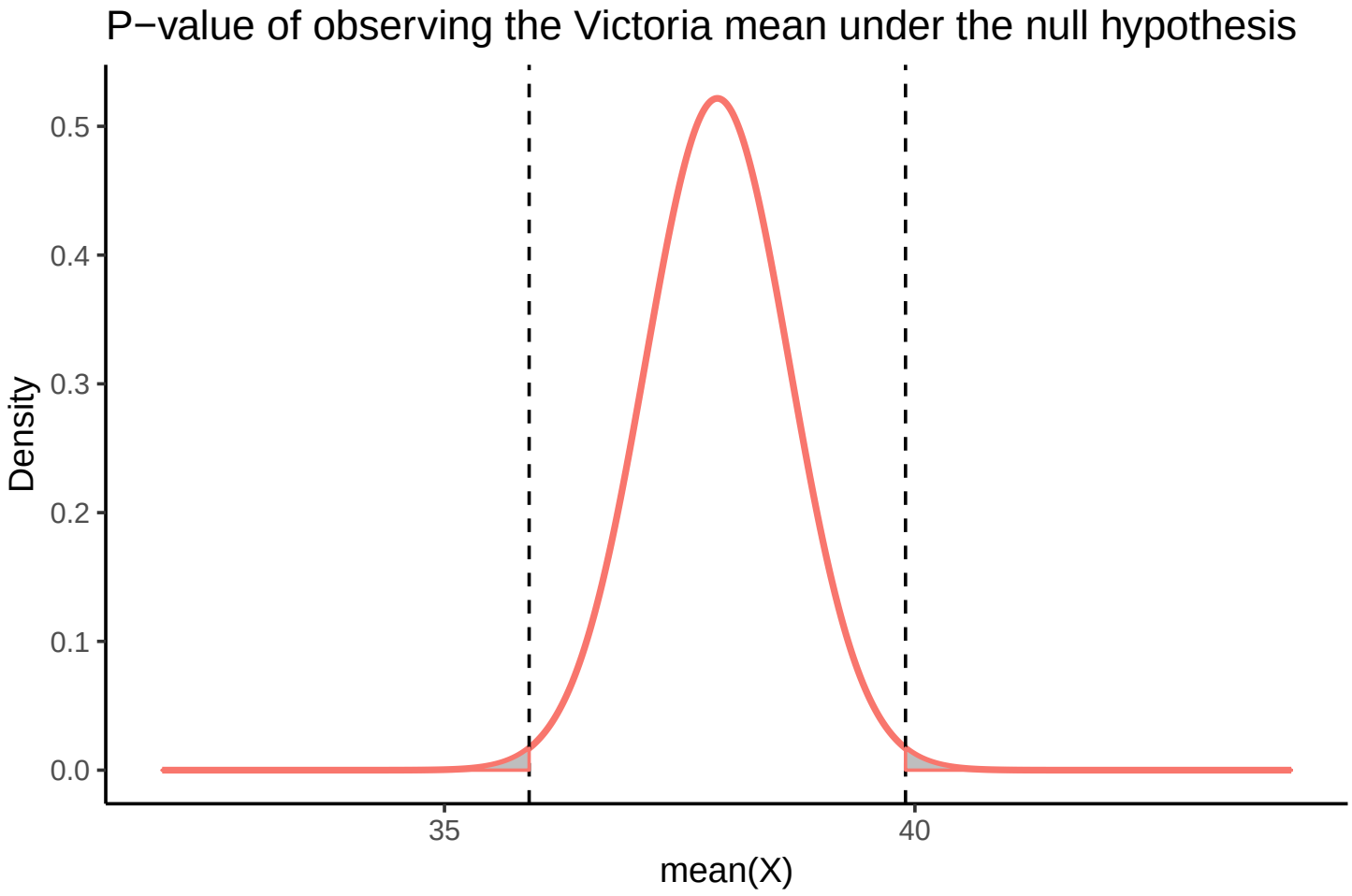
Tail lengths, null distributions of interest



Mean tail length example, continued



Mean tail length, cont.



Example: Toad handedness

Do toads exhibit biased handedness? Bisazza et al. (1996) tested the possiblity of biased handedness in European toads, *Bufo bufo* by using a behavioral assay to determine the preferred hand each individual frog used to perform a task.

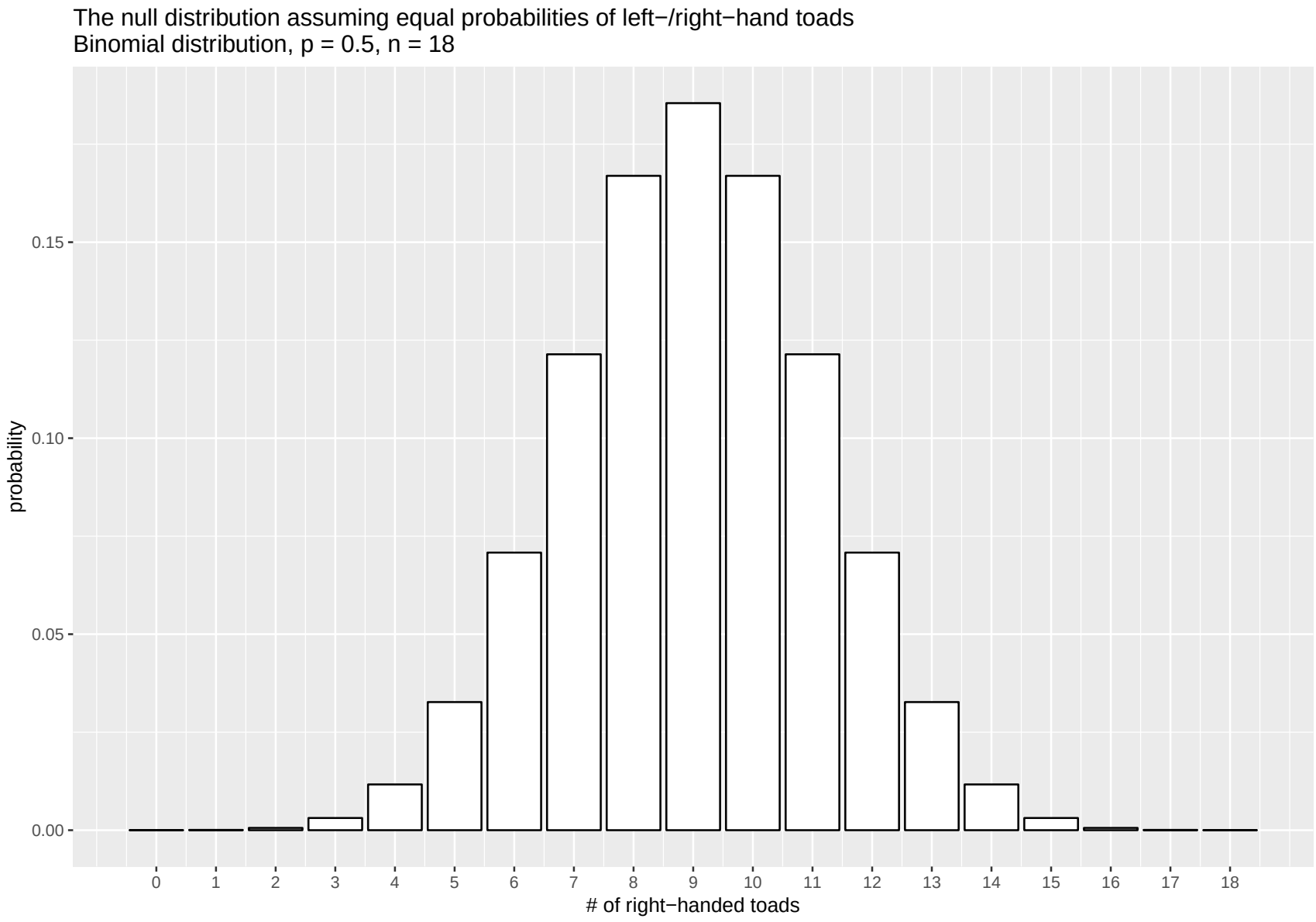
Null and alternative hypotheses

- H_0: There is no difference in the proportion of right-handed and left-handed frogs, i.e. p(right handed) $= 0.5$
- H_A: There is no difference in the proportion of right-handed and left-handed frogs, i.e. p(right haned) $\neq 0.5$

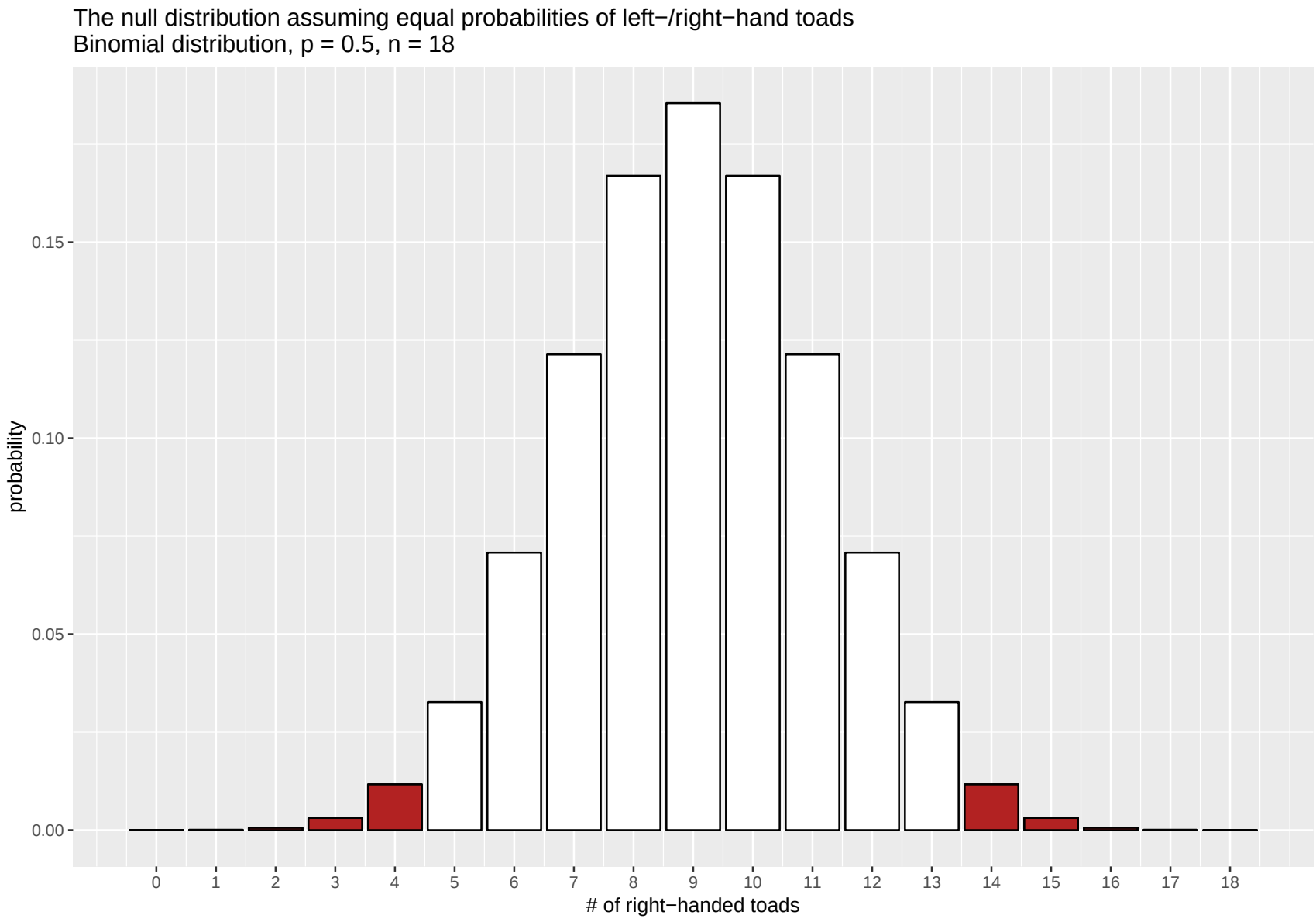
Toad data

- 18 toads analyzed
- 14 showed right-hand preference, 4 showed left-hand preference

Null distribution under the binomial distribution



Calculating the p-value



Carrying out a binomial test in R

```r
binom.test(4, 18, p=0.5, alternative = "two.sided")
```

```
    Exact binomial test

data:  4 and 18
number of successes = 4, number of trials =
18, p-value = 0.03088
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.06409205 0.47637277
sample estimates:
probability of success
             0.2222222
```

One-tailed tests

Where there are strong a priori predictions about the direction of difference from a null hypothesis, the use of a "one-tailed" hypothesis test is sometimes justified.

- The decision to use a one-tailed test should really be made before a study is carried out; not after looking at the data!
- A two tailed test is always more conservative

Example

A factory that packages cereal; cereal boxes are supposed to have an average mass of 300g. Truth-in-advertising laws require the factory to stop the production lines when the average mass of cereal in a sample of 50 boxes can be statistically shown to be less than 300g (overfilling is costly to the company but doesn't hurt the consumer). A one-tailed test is justified here because, from the perspective of compliance, only deviations in one direction from the mean are of interest.