

# Statistical Analysis of Proportions

Bret Hanlon and Bret Larget

Department of Statistics  
University of Wisconsin—Madison

September 9–16, 2010

# Recombination

## Example

In the fruit fly *Drosophila melanogaster*, the gene *white* with alleles  $w^+$  and  $w$  determines eye color (red or white) and the gene *miniature* with alleles  $m^+$  and  $m$  determines wing size (normal or miniature). Both genes are located on the X chromosome, so female flies will have two alleles for each gene while male flies will have only one. During meiosis (in animals, the formation of gametes) in the female fly, if the X chromosome pair do not exchange segments, the resulting eggs will contain two alleles, each from the same X chromosome. However, if the strands of DNA cross-over during meiosis then some progeny may inherit alleles from different X chromosomes. This process is known as *recombination*. There is biological interest in determining the proportion of recombinants. Genes that have a positive probability of recombination are said to be *genetically linked*.

## Recombination (cont.)

### Example

In a pioneering 1922 experiment to examine genetic linkage between the *white* and *miniature* genes, a researcher crossed  $wm^+ / w^+ m$  female flies with male  $wm^+ / Y$  chromosomes and looked at the traits of the male offspring. (Males inherit the Y chromosome from the father and the X from the mother.) In the absence of recombination, we would expect half the male progeny to have the  $wm^+$  haplotype and have white eyes and normal-sized wings while the other half would have the  $w^+ m$  haplotype and have red eyes and miniature wings. This is not what happened.

## Recombination (cont.)

### Example

The phenotypes of the male offspring were as follows:

Eye color	Wing Size	
	normal	miniature
red	114	202
white	226	102

There were  $114 + 102 = 216$  recombinants out of 644 total male offspring, a proportion of  $216/644 \doteq 0.335$  or 33.5%. Completely linked genes have a recombination probability of 0, whereas unlinked genes have a recombination probability of 0.5. The *white* and *miniature* genes in fruit flies are *incompletely linked*. Measuring recombination probabilities is an important tool in constructing *genetic maps*, diagrams of chromosomes that show the positions of genes.

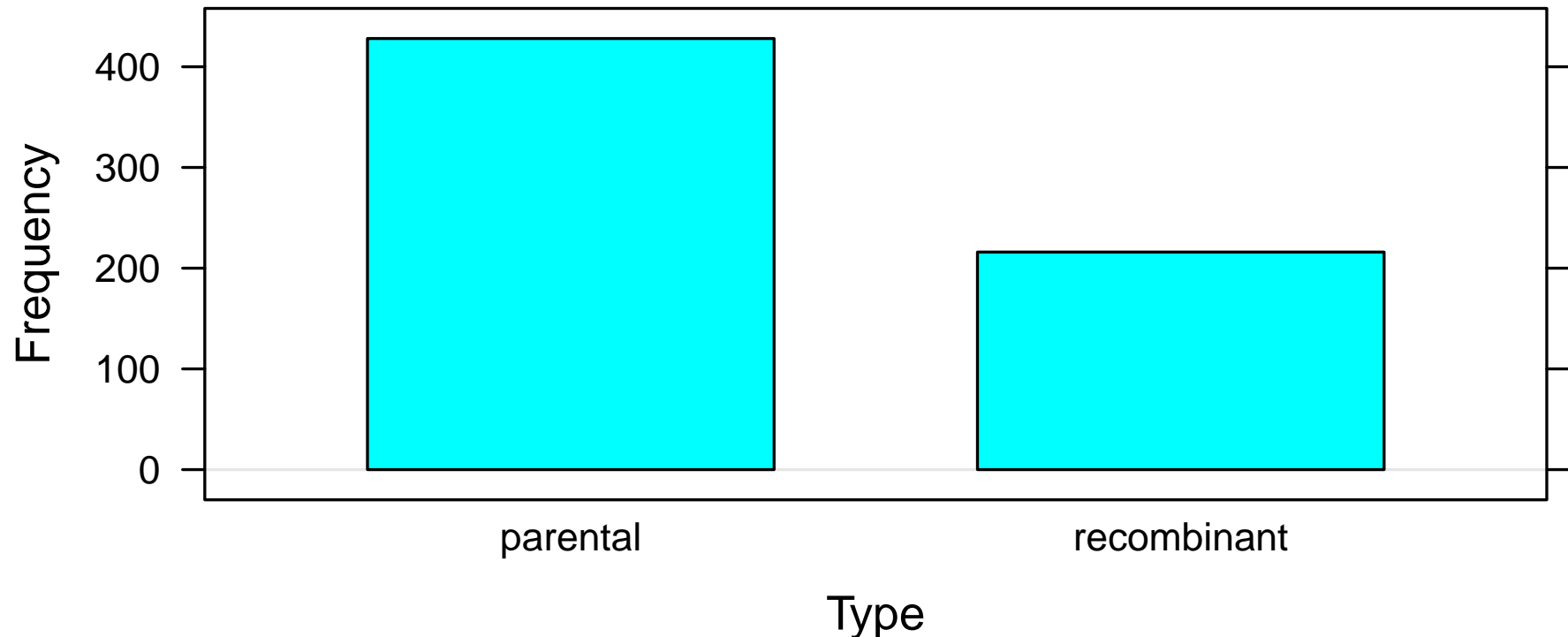
# Proportions in Biology

- Many problems in biology fit into the framework of using sampled data to estimate *population proportions* or *probabilities*.
- In reference to our previous discussion about data, we may be interested in knowing what proportion of a population are in a specific category of a categorical variable.
- For this fly genetics example, we may want to address the following questions:
  - ▶ How close is the population recombination probability to the observed proportion of 0.335?
  - ▶ Are we sure that these genes really linked? If the probability was really 0.5, might we have seen this data?
  - ▶ How many male offspring would we need to sample to be confident that our estimated probability was within 0.01 of the true probability?
- To understand statistical methods for analyzing proportions, we will take our first foray into *probability theory*.

# Bar Graphs

- Proportions are fairly simple statistics, but *bar graphs* can help one to visualize and compare proportions.
- The following graph shows the relative number of individuals in each group and helps us see that there are about twice as many parental types as recombinants.

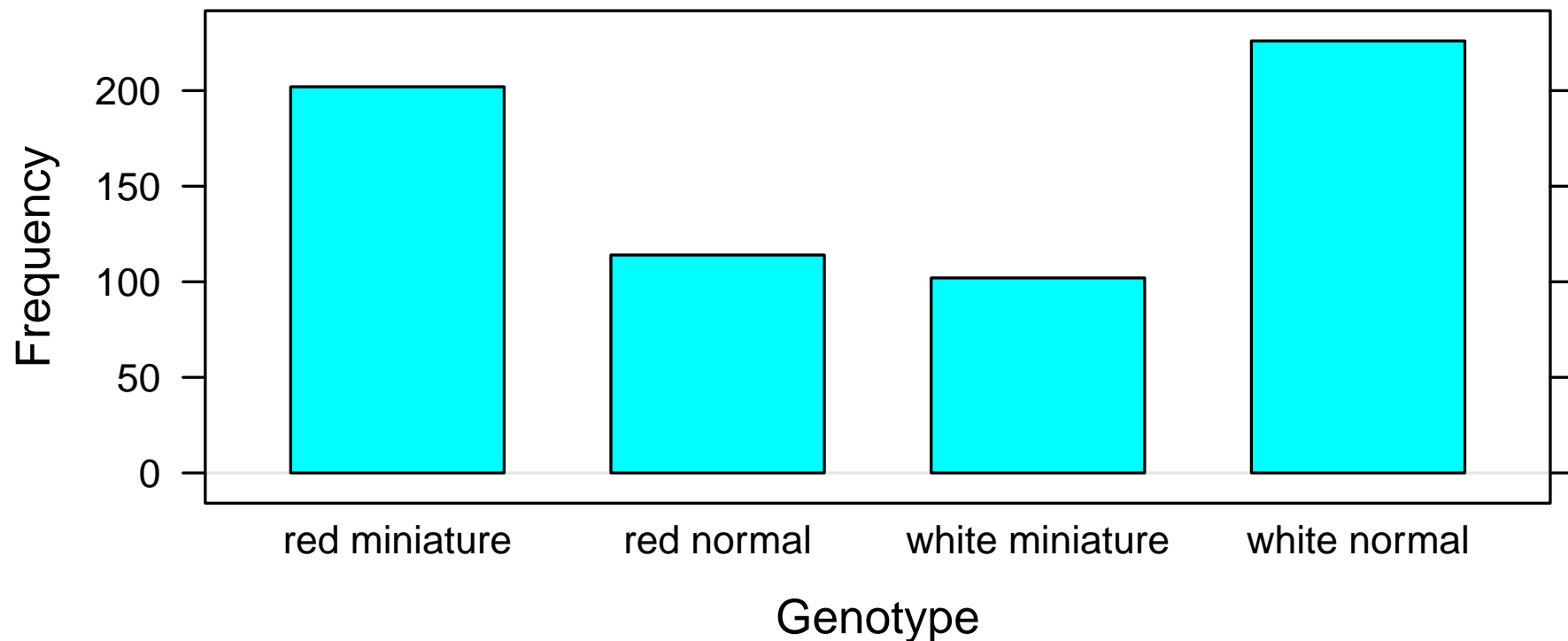
**Male Offspring Types**



## Bar Graphs (cont.)

- The following graph shows the totals in each genotype.
- A later section will describe the R code to make these and other graphs.

### Male Offspring Genotypes



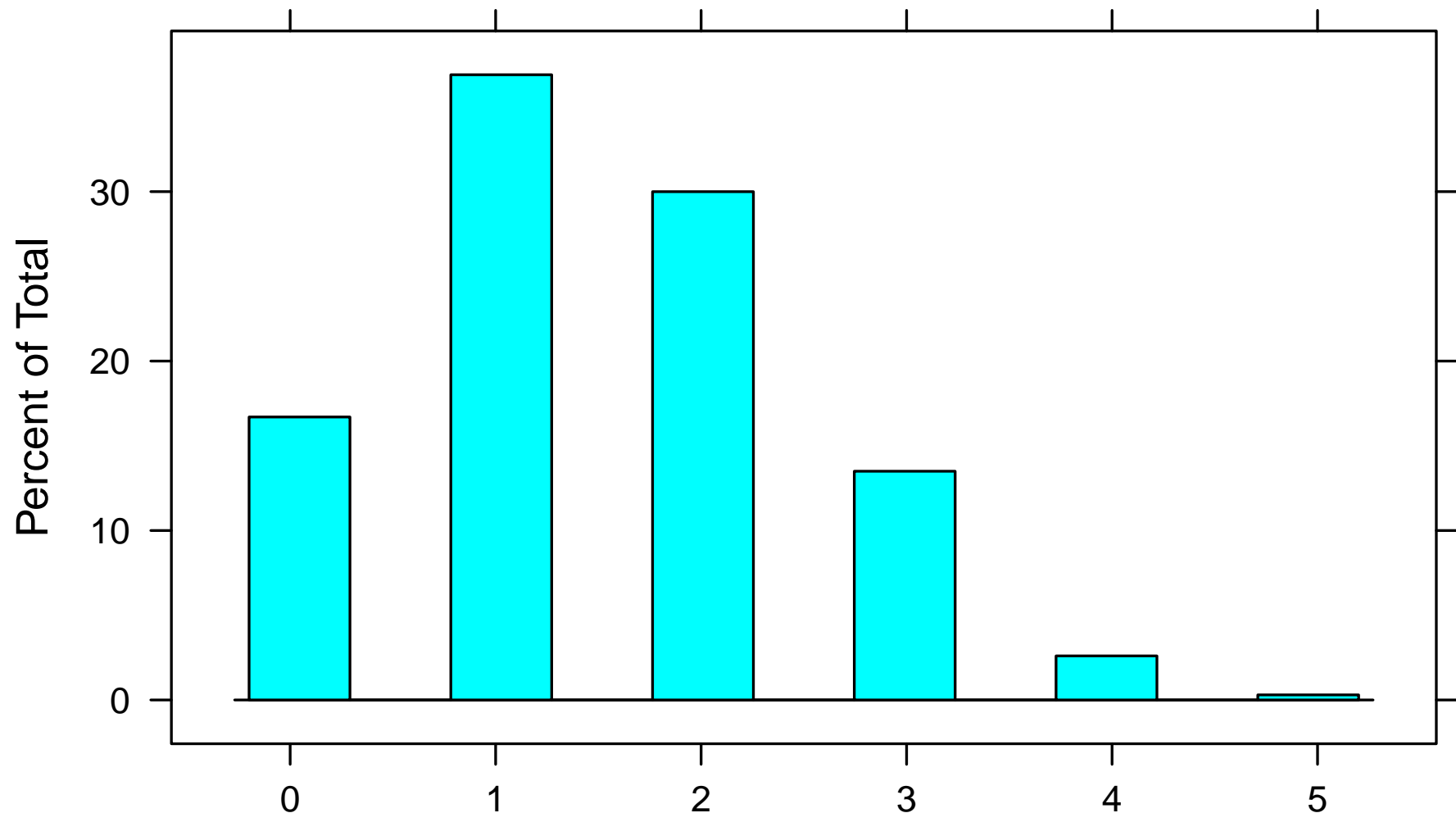
# Motivating Example

- We begin by considering a small and simplified example based on our case study.
- Assume that the true probability of recombination is  $p = 0.3$  and that we take a small sample of  $n = 5$  flies.
- The number of recombinants in this sample could potentially be 0, 1, 2, 3, 4, or 5.
- The chance of each outcome, however, is not the same.



# Simulation

- Using the computer, we can simulate many (say 1000) samples of size 5, for each sample counting the number of recombinants.



# Simulation Results

- If we let  $X$  represent the number of recombinants in the sample, we can describe the *distribution* of  $X$  by specifying;
  - ▶ the set of possible values; and
  - ▶ a probability for each possible value.
- In this example, the possible values and the probabilities (as approximated from the simulation) are:

0	1	2	3	4	5
0.17	0.36	0.31	0.13	0.03	0.00

- Rather than depending on simulation, we will derive a mathematical expression for these probabilities.

# The Binomial Distribution Family

- The *binomial distribution* family is based on the following assumptions:
  - ① There is a *fixed sample size* of  $n$  separate *trials*.
  - ② Each trial has *two possible outcomes* (or classes of outcomes, one of which is counted, and one of which is not).
  - ③ Each trial has the *same probability*  $p$  of being in the class of outcomes being counted.
  - ④ The trials are *independent*, which means that information about the outcomes for some subset of the trials does not affect the probabilities of the other trials.
- The values of  $n$  (some positive integer) and  $p$  (a real number between 0 and 1) determine the full distribution (list of possible values and associated probabilities).

# Binomial Probability Formula

## Binomial Probability Formula

If  $X \sim \text{Binomial}(n, p)$ , then

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad \text{for } k = 0, \dots, n$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{is the number of ways to choose } k \text{ objects from } n.$$

# Example

- In the example, let p represent a parental type and R a recombinant type.
- There are 32 possible samples in order of these types, organized below by the number of recombinants.

$$\overbrace{ppppp}^{\binom{5}{0}=1}$$

$$\overbrace{\begin{array}{l} ppppR \\ pppRp \\ ppRpp \\ pRppp \\ Rpppp \end{array}}^{\binom{5}{1}=5}$$

$$\overbrace{\begin{array}{l} pppRR \\ ppRpR \\ ppRRp \\ pRppR \\ pRpRp \\ pRRpp \\ RpppR \\ RppRp \\ RpRpp \\ RRppp \end{array}}^{\binom{5}{2}=10}$$

$$\overbrace{\begin{array}{l} ppRRR \\ pRpRR \\ pRRpR \\ pRRRp \\ RppRR \\ RpRpR \\ RpRRp \\ RRppR \\ RRpRp \\ RRRpp \end{array}}^{\binom{5}{3}=10}$$

$$\overbrace{\begin{array}{l} pRRRR \\ RpRRR \\ RRpRR \\ RRRpR \\ RRRRp \end{array}}^{\binom{5}{4}=5}$$

$$\overbrace{RRRRR}^{\binom{5}{5}=1}$$

## Example (cont.)

- In the example, p has probability 0.7 and R has probability 0.3;
- The sequence ppppp has probability  $(0.7)^5$
- Since this is the only sequence with 0 Rs,  
 $P(X = 0) = 1 \times (0.3)^0(0.7)^5 \doteq 0.1681$ .
- The sequence ppRpR has probability  $(0.3)^2(0.7)^3$  as do each of the 10 sequences with exactly two Rs, so  
 $P(X = 2) = 10 \times (0.3)^2(0.7)^3 \doteq 0.3087$ .
- The complete distribution is:

0	1	2	3	4	5
0.1681	0.3601	0.3087	0.1323	0.0284	0.0024

- In the general formula  $P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ :
  - ▶  $\binom{n}{k}$  is the number of different patterns with exactly  $k$  of one type; and
  - ▶  $p^k (1 - p)^{n-k}$  is the probability of any single such sequence.

# Random Variables

## Definition

A *random variable* is a rule that attaches a numerical value to a chance outcome.

- In our example, we defined the random variable  $X$  to be the number of recombinants in the sample.
- This random variable is *discrete* because it has a finite set of possible values.
- (Random variables with a countably infinite set of possible values, such as 0, 1, 2, ... are also discrete, but with a continuum of possible values are called *continuous*. We will learn more about continuous random variables later in the semester.)
- Associated with each possible value of the random variable is a *probability*, a number between 0 and 1 that represents the long-run relative frequency of observing the given value.
- The sum of the probabilities for all possible values is one.

# Discrete Probability Distributions

- The *probability distribution* of a random variable is a full description of a unit of probability is distributed on the real number line.
- For a discrete random variable, the probability is broken into discrete chunks and placed at specific locations.
- To describe the distribution, *it is sufficient to provide a list of all possible values and the probability associated with each value.*
- The sum of these probabilities is one.
- Frequently (as with the binomial distribution), there is a formula that specifies the probability for each possible value.



# The Mean (Expected Value)

## Definition

The *mean* or *expected value* of a random variable  $X$  is written as  $E(X)$ . For discrete random variables,

$$E(X) = \sum_k kP(X = k)$$

where the sum is over all possible values of the random variable.

- Note that the expected value of a random variable is a *weighted average* of the possible values of the random variable, weighted by the probabilities.
- A general discrete weighted average takes the form

$$\sum_i (\text{value})_i (\text{weight})_i \quad \text{where} \quad \sum_i (\text{weight})_i = 1$$

- The mean is location where the probabilities *balance*.

# The Variance and Standard Deviation

## Definition

The *variance* of a random variable  $X$  is written as  $\text{Var}(X)$ . For discrete random variables,

$$\text{Var}(X) = E\left((X - E(X))^2\right) = \sum_k (k - \mu)^2 P(X = k) = E(X^2) - \left(E(X)\right)^2$$

where the sum is over all possible values of the random variable and  $\mu = E(X)$ .

- The variance is a weighted average of the squared deviations between the possible values of the random variable and its mean.
- If a random variable has units, the units of the variance are those units squared, which is hard to interpret.
- We also define the *standard deviation* to be the square root of the variance, so it has the same units as the random variable.
- A notation is  $\text{SD}(X) = \sqrt{\text{Var}(X)}$ .

# Chalkboard Example

- Find the mean, variance, and standard deviation for a random variable with this distribution.

$k$	0	1	5	10
$P(X = k)$	0.1	0.5	0.1	0.3

# Formulas for the Binomial Distribution Family

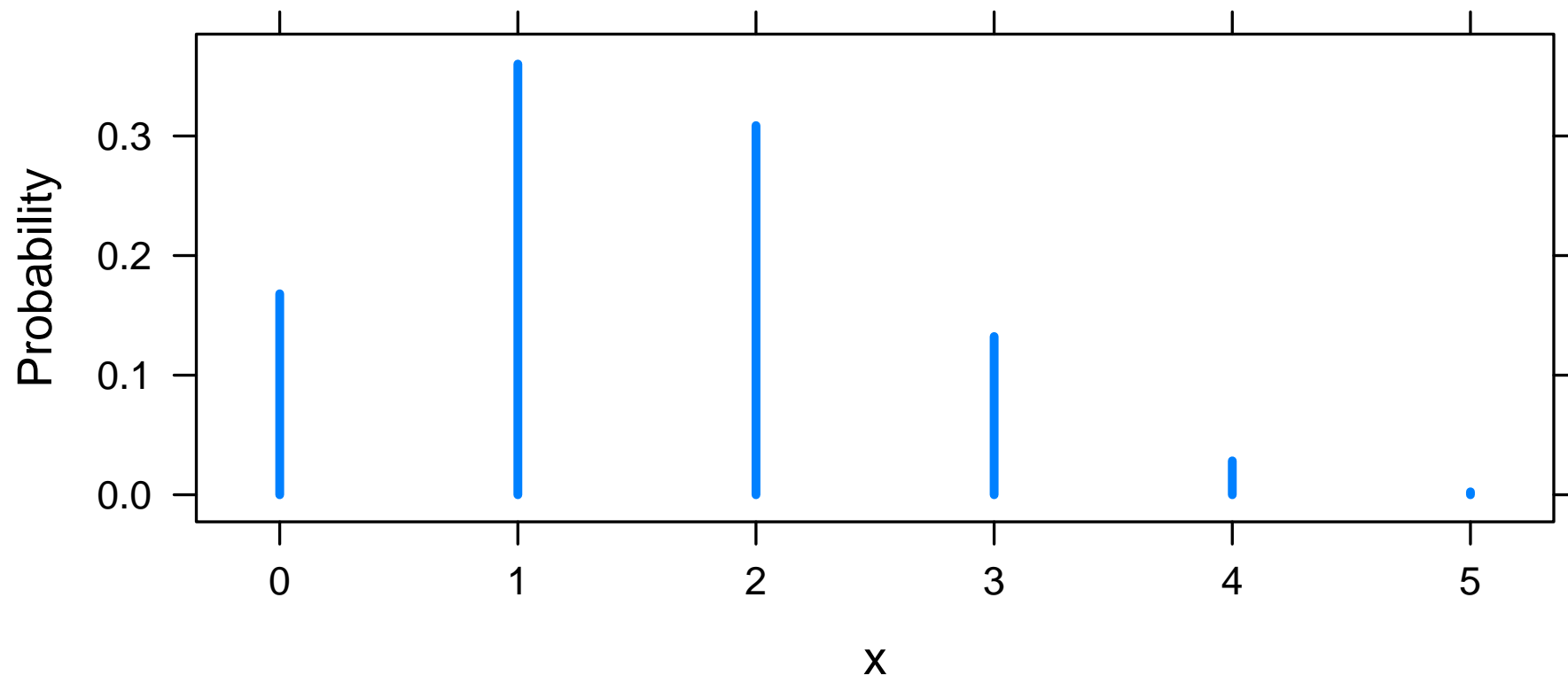
## Moments of the Binomial Distribution

If  $X \sim \text{Binomial}(n, p)$ , then  $E(X) = np$ ,  $\text{Var}(X) = np(1 - p)$ , and  $\text{SD}(X) = \sqrt{np(1 - p)}$ .

- Each of these formulas involves considerable algebraic simplification from the expressions in the definitions.
- The expression for the mean is intuitive: for example, in a sample where  $n = 5$  and we expect the proportion  $p = 0.3$  of the sample to be of one type, then it is not surprising that the distribution is centered at 30% of 5, or 1.5.

## Example

- Here is a plot of the distribution in our small example.
- The exact probabilities are very close to the values from the simulation.



# What you should know (so far)

You should know:

- when a random variable is binomial (and if so, what its parameters are);
- how to compute binomial probabilities;
- how to find the mean, variance, and standard deviation from the definition for a general discrete random variable;
- how to use the simple formulas to find the mean and variance of a binomial random variable;
- that the expected value is the mean (balancing point) of a probability distribution;
- that the expected value is a measure of the center of a distribution;
- that variance and standard deviation are measures of the spread of a distribution.

# Sampling Distribution

## Definition

A *statistic* is a numerical value that can be computed from a sample of data.

## Definition

The *sampling distribution* of a statistic is simply the probability distribution of the statistic when the sample is chosen at random.

## Definition

An *estimator* is a statistic used to estimate the value of a characteristic of a population.

- We will explore these ideas in the context of using *sample proportions* to estimate *population proportions* or *probabilities*.

# The Sample Proportion

- Let  $X$  count the number of observations in a sample of a specified type.
- For a random sample, we often model  $X \sim \text{Binomial}(n, p)$  where:
  - ▶  $n$  is the sample size; and
  - ▶  $p$  is the population proportion.

- The sample proportion is

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

- Adding a *hat* to a population parameter is a common statistical notation to indicate an estimate of the parameter calculated from sampled data.
- What is the sampling distribution of  $\hat{p}$ ?



# Sampling distribution of $\hat{p}$

- The possible values of  $\hat{p}$  are  $0 = 0/n, 1/n, 2/n, \dots, n/n = 1$ .
- The probabilities for each possible value are the binomial probabilities:

$$P\left(\hat{p} = \frac{k}{n}\right) = P(X = k)$$

- The mean of the distribution is  $E(\hat{p}) = p$ .
- The variance of the distribution is  $\text{Var}(\hat{p}) = \frac{p(1-p)}{n}$ .
- The standard deviation of the distribution is  $\text{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ .
- We connect these formulas to the binomial distribution.

# Expected Values and Constants

- While it is intuitively clear that the expected value of all sample proportions ought to be equal to the population proportion, it is helpful to understand why.
- First, for any constant  $c$ ,  $E(cX) = cE(X)$ .
- This follows because constants can be factored out of sums.
- The number  $1/n$  is a constant, so

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{1}{n}(np) = p$$

# Expected Values and Sums

## Expectation of a Sum

If  $X_1, X_2, \dots, X_n$  are random variables, then  
 $E(X_1 + \dots + X_n) = E(X_1) + \dots + E(X_n)$ .

- *The expected value of a sum is the sum of the expected values.*
- This follows because sums can be rearranged into other sums.
- For example,

$$(a_1 + b_1) + (a_2 + b_2) + \dots + (a_n + b_n) = (a_1 + \dots + a_n) + (b_1 + \dots + b_n)$$

- There is also a naturally intuitive explanation of this result: for example, if we expect to see 5 recombinants on average in one sample and 6 recombinants on average in a second, then we expect to see 11 on average when the samples are combined.

# The Binomial Moments Revisited

$k$	0	1
$P(X = k)$	$1 - p$	$p$

- If  $n = 1$ , then the binomial distribution is as above and

$$E(X) = 0(1 - p) + 1(p) = p .$$

- In addition,

$$\text{Var}X = (0 - p)^2(1 - p) + (1 - p)^2p = p^2(1 - p) + p(1 - p)^2 = p(1 - p)$$

# The Binomial Mean Revisited

- For larger  $n$ , a sample of size  $n$  can be thought of as combining  $n$  samples of size 1, so

$$X = X_1 + X_2 + \cdots + X_n$$

where each  $X_i$  has possible values 0 and 1 (the  $i$ th element of the sample is not counted or is).



$$\begin{aligned} E(X) &= E(X_1 + \cdots + X_n) \\ &= E(X_1) + \cdots + E(X_n) \\ &= \underbrace{p + \cdots + p}_{n \text{ times}} = np \end{aligned}$$

# Variance and Sums

## Variance of a Sum

If  $X_1, X_2, \dots, X_n$  are random variables, and if *the random variables are independent*, then  $\text{Var}(X_1 + \dots + X_n) = \text{Var}(X_1) + \dots + \text{Var}(X_n)$ .

- In words, *the variance of a sum of independent random variables is the sum of the variances*.
- Later sections will explore variances of general sums.

# The Binomial Variance Revisited

- For  $X \sim \text{Binomial}(n, p)$ , we can think of  $X$  as a sum of independent random variables

$$X = X_1 + X_2 + \cdots + X_n$$

where each  $X_i$  has possible values 0 and 1, and



$$\begin{aligned}\text{Var}(X) &= \text{Var}(X_1 + \cdots + X_n) \\ &= \text{Var}(X_1) + \cdots + \text{Var}(X_n) \\ &= \underbrace{p(1-p) + \cdots + p(1-p)}_{n \text{ times}} = np(1-p)\end{aligned}$$

# Constants and Variance

- As the variance squares units, when a constant is factored out, its value is also squared.

$$\text{Var}(cX) = c^2 \text{Var}(X)$$

- This can be understood from the definition.

$$\begin{aligned}\text{Var}(cX) &= E\left((cX - E(cX))^2\right) \\ &= E\left((cX - cE(X))^2\right) \\ &= E\left(c^2(X - E(X))^2\right) \\ &= c^2 \text{Var}(X)\end{aligned}$$



# The Variance of $\hat{p}$

$$\begin{aligned}\text{Var}(\hat{p}) &= \text{Var}\left(\frac{X}{n}\right) \\ &= \frac{1}{n^2} \text{Var}(X) \\ &= \frac{1}{n^2} np(1 - p) \\ &= \frac{p(1 - p)}{n}\end{aligned}$$

- Then,  $\text{SD}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$ .

# Standard Error

## Definition

The standard deviation of the sampling distribution of an estimate is called the *standard error* of the estimate.

- A standard error can be thought of as the size of a the typical distance between an estimate and the value of the parameter it estimates.
- Standard errors are often estimated by replacing parameter values with estimates.
- For example,

$$SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}, \quad \widehat{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

# Problem

## Problem

In a large population, the frequency of an allele is 0.25. A cross results in a random sample of 8 alleles from the population.

- ① Find the mean, variance, and standard error of  $\hat{p}$ .
- ② Find  $P(\hat{p} = 0.4)$ .
- ③ Find  $P(\hat{p} = 0.5)$ .
- ④ Find  $P(|\hat{p} - p| > 2SE(\hat{p}))$ .

# Solutions

- ①  $E(\hat{p}) = 0.25$ ,  $\text{Var}(\hat{p}) \doteq 0.0234$ ,  $\text{SE}(\hat{p}) \doteq 0.1531$ .
- ②  $P(X = 3.2) = 0$
- ③  $P(X = 4) \doteq 0.0865$
- ④  $P(X \geq 5) \doteq 0.0273$ .

# What you should know (so far)

You should know:

- that the sampling distribution of the sample proportion (from a random sample) is simply a rescaled binomial distribution;
- the two linearity rules of expectation:
  - ▶  $E(cX) = cE(X)$ ;
  - ▶  $E(X_1 + \cdots + X_n) = E(X_1) + \cdots + E(X_n)$
- how the expectation rules work for variances:
  - ▶  $\text{Var}(cX) = c^2\text{Var}(X)$ ;
  - ▶  $\text{Var}(X_1 + \cdots + X_n) = \text{Var}(X_1) + \cdots + \text{Var}(X_n)$  if  $X_1, \dots, X_n$  are independent.
- how to do probability calculations for small sample proportion problems.

# The Big Picture for Estimation

- In some settings, we may think of a population as a large bucket of colored balls where the population proportion of red balls is  $p$ .
- In a random sample of  $n$  balls from the population, if there are  $X$  red balls in the sample, then  $\hat{p} = \frac{X}{n}$  is the sample proportion.
- $\hat{p}$  is an estimate of  $p$ .
- We wish to quantify the uncertainty in the estimate.
- We will do so by expressing a *confidence interval*; a statement such as *we are 95% confident that  $0.28 < p < 0.39$* .
- Confidence intervals for population proportions are based on the sampling distribution of sample proportions.

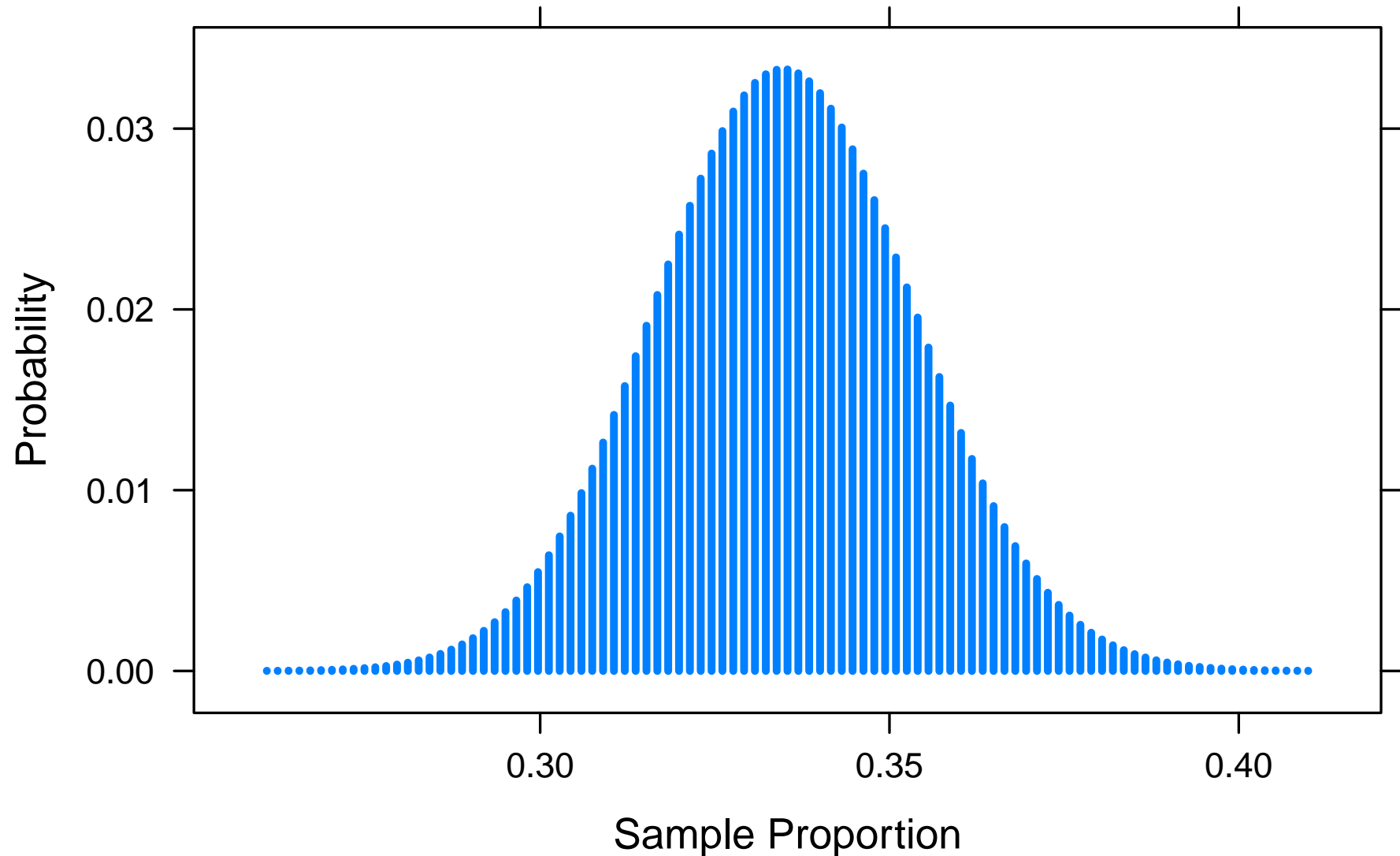
# Recombination Example

## Example

Recall our previous example involving recombination in fruit flies; In a genetics experiment, 216 of 644 male progeny were recombinants. We estimate the recombination probability between the *white* and *miniature* genes to be  $\hat{p} = 216/644 \doteq 0.335$ . How confident are we in this estimate?

# Sampling Distribution

- If  $p = 0.335$ , the sampling distribution of  $\hat{p}$  would look like this.





# Comments on the Sampling Distribution

- The shape of the graph of the discrete probabilities is well described by a *continuous, smooth, bell-shaped curve* called a *normal curve*.
- The mean of the sampling distribution is  $E(\hat{p}) = p = 0.335$ .
- The standard deviation of the sampling distribution is  $SE(\hat{p}) = \sqrt{\frac{0.335(1-0.335)}{644}} \doteq 0.019$ , which is an estimate of the size of the difference between  $p$  and  $\hat{p}$ .
- Even if  $p$  were not exactly equal to 0.335, the numerical value of  $SE(\hat{p})$  would be very close.
- In an ideal normal curve, 95% of the probability is within  $z = 1.96$  standard deviations of the mean.
- As long as  $n$  is large enough, the sampling distribution of  $\hat{p}$  will be approximately normal.
- A rough rule of thumb for *big enough* is that  $X$  and  $n - X$  are each at least five; here  $X = 216$  and  $n - X = 428$ .

# Confidence Interval Procedure

- A 95% confidence interval for  $p$  is constructed by taking an interval centered at an estimate of  $p$  and extending 1.96 standard errors in each direction.
- Statisticians have learned that using an estimate  $p' = \frac{X+2}{n+4}$  results in more accurate confidence intervals than the more natural  $\hat{p}$ .
- The  $p'$  estimate is the sample proportion if the sample size had been four larger and if two of the four had been of each type.

## 95% Confidence Interval for $p$

A 95% confidence interval for  $p$  is

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

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

# Application

- Using our example data,  $p' = (216 + 2)/(644 + 4) \doteq 0.336$ .
- Notice this is shifted a small amount toward 0.5 from  $\hat{p} = 0.335$ .
- The estimated standard error is  $\sqrt{\frac{0.336(1-0.336)}{648}} \doteq 0.019$ .
- This means that the true  $p$  probably differs from our estimate by about 0.019, give or take.
- The *margin of error* is  $1.96 \times 0.019 \doteq 0.036$ .
- We then construct the following 95% confidence interval for  $p$ .

$$0.300 < p < 0.373$$

- This is understood in the context of the problem as:

*We are 95% confident that the recombination probability for the white and miniature genes in fruit flies is between 0.300 and 0.373.*

# Interpretation

- *Confidence* means something different than *probability*, but the distinction is subtle.
- From a frequentist point of view, the interval  $0.300 < p < 0.373$  has nothing random in it since  $p$  is a fixed, unknown constant.
- Thus, it would be wrong to say there is a 95% chance that  $p$  is between 0.300 and 0.373: it is either 100% true or 100% false.
- The 95% confidence arises from using a procedure that has a 95% chance of capturing the true  $p$ .
- There is a 95% chance that *some confidence interval* will capture  $p$ ; we are 95% confident that the fixed interval (0.300, 0.373) based on our sample is one of these.
- From a Bayesian statistical point of view, all uncertainty is described with probability and it would be perfectly legitimate to say simply that there is a 95% probability that  $p$  is between 0.300 and 0.373.
- *Most biologists and many statisticians do not get overly concerned with this distinction in interpretations.*

## A Second Example

### Example

Male radiologists may be exposed to much more radiation than typical people, and this exposure might affect the probability that children born to them are male. In a study of 30 “highly irradiated” radiologists, 30 of 87 offspring were male (Hama et al. 2001). Treating this data as a random sample, find a confidence interval for the probability that the child of a highly irradiated male radiologist is male.

# Calculation

- We find  $\hat{p} = 30/87 \doteq 0.345$  and  $p' = 32/91 \doteq 0.352$ .
- The estimated standard error is  $\sqrt{\frac{0.352(1-0.352)}{91}} \doteq 0.050$ .
- The estimated margin of error is  $1.96 \times SE \doteq 0.098$ .
- The confidence interval is  $0.254 < p < 0.450$ .

*We are 95% confident that the proportion of children of highly irradiated male radiologists that are boys is between 0.254 and 0.450.*

This confidence interval does not contain 0.512, the proportion of male births in the general population. The inference is that exposure to high levels of radiation in men may decrease the probability of having a male child.

# Probability Models

## Definition

A probability model  $P(x | \theta)$  relates possible values of data  $x$  with parameter values  $\theta$ .

- If  $\theta$  is fixed and  $x$  is allowed to vary, the probability model describes the probability distribution of a random variable.
- The total amount of probability is one.
- For a discrete random variable with possible values  $x_1, x_2, \dots$ , and a fixed parameter  $\theta$ , this means that

$$\sum_i P(x_i | \theta) = 1$$

In words, the sum of the probabilities of all possible values are one.

- Each different fixed value of  $\theta$  corresponds to a possibly different probability distribution.

# Likelihood

## Definition

The *likelihood* is a function of the parameter  $\theta$  that takes a probability model  $P(x | \theta)$ , but treats the data  $x$  as fixed while  $\theta$  varies.

$$L(\theta) = P(x | \theta), \quad \text{for fixed } x$$

- Unlike probability distributions, there is no constraint that the total likelihood must be one.
- Likelihood can be the basis of the estimation of parameters: parameter values for which the likelihood is relatively high are potentially good explanations of the data.



# Log-Likelihood

## Definition

The *log-likelihood* is the natural logarithm of the likelihood.

- As probabilities for large sets of data often become very small and as probability models often consist of products of probabilities, it is common to represent likelihood on the natural log scale.

$$\ell(\theta) = \log L(\theta)$$

- (Note that texts in statistics and mathematics often assume that  $\log$  represents the natural log and not the base 10 log. We will follow this convention.)

# Likelihood and Proportions

- The estimate  $\hat{p}$  can also be justified on the basis of *likelihood*.
- The binomial probability model for data  $x$  and parameter  $p$  is

$$P(x | p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

where  $x$  takes on possible values  $0, 1, \dots, n$  and  $p$  is a real number between 0 and 1.

- For fixed  $x$ , the likelihood model is

$$L(p) = \binom{n}{x} p^x (1 - p)^{n-x}$$

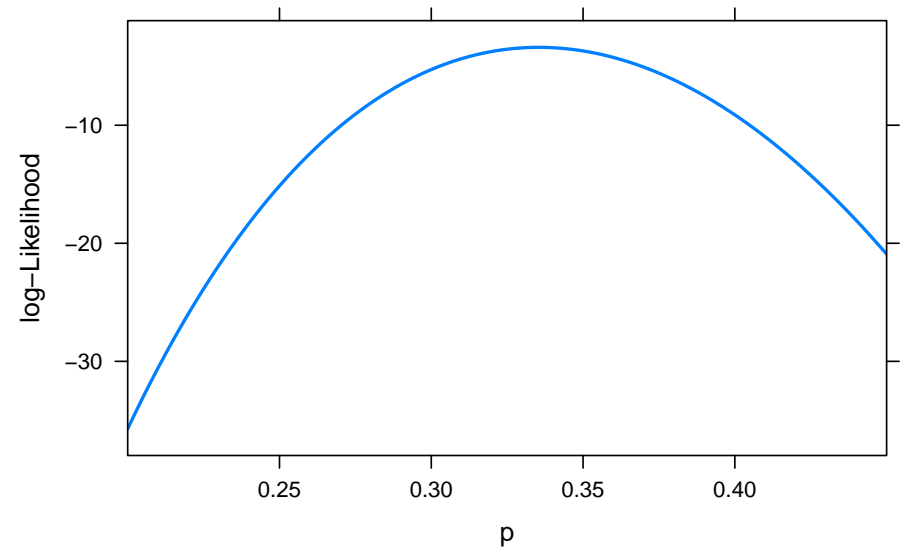
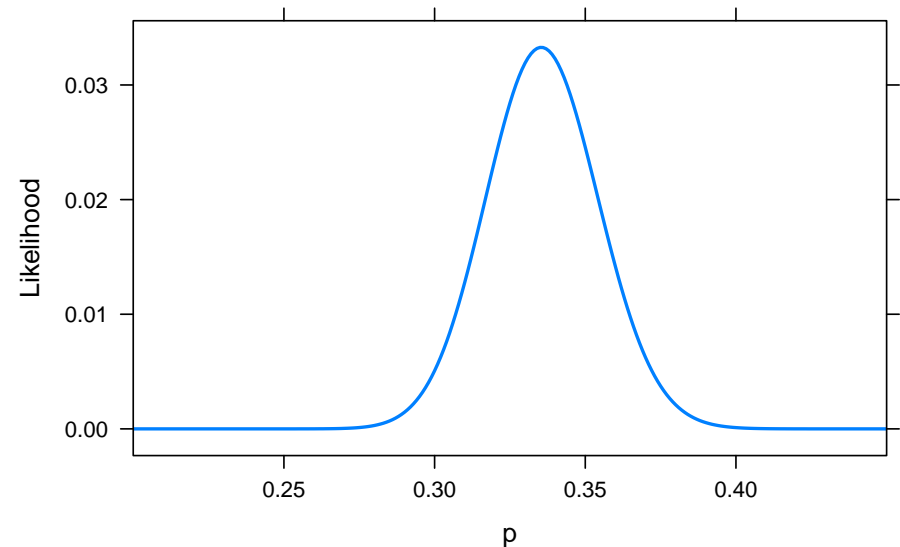
- The log-likelihood is

$$\ell(p) = \log \binom{n}{x} + x \log(p) + (n - x) \log(1 - p)$$

- Recall these facts about logarithms:
  - ▶  $\log(ab) = \log(a) + \log(b)$ ;
  - ▶  $\log(a^b) = b \log(a)$ .

# Graphs

- In our example,  $x = 216$  recombinants out of  $n = 644$  fruit flies.
- The top graph shows the likelihood.
- The bottom graph shows the log-likelihood.
- Note that even though the shapes of the curves are different and the scales are quite different, the curves are each maximized at the same point.



# Maximum Likelihood Estimation

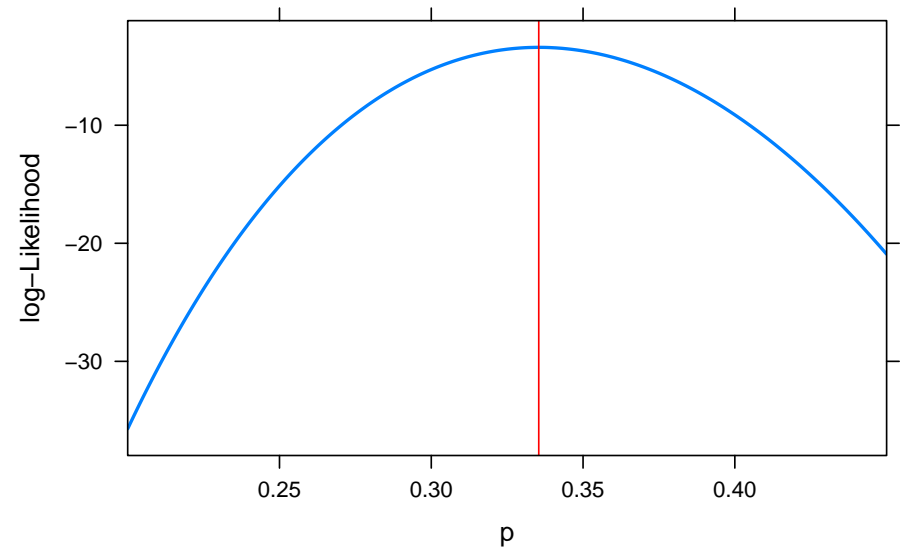
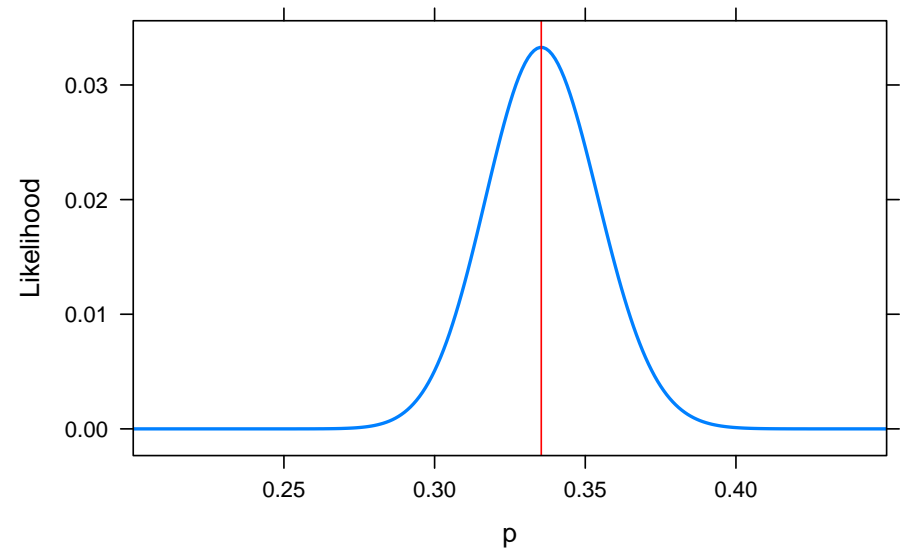
## Definition

The *maximum likelihood estimate* of a parameter is the value of the parameter that maximizes the likelihood function.

- The *likelihood principle* states that all information in data about parameters is contained in the likelihood function.
- The principle of maximum likelihood says that the best estimate of a parameter is the value that maximizes the likelihood.
- This is the value that makes the probability of the observed data as large as possible.

# Example

- In our example, the sample proportion is  $\hat{p} = 216/644 \doteq 0.335$ .
- The vertical lines are drawn at this value.
- We see that  $p = \hat{p}$  is the maximum likelihood estimate.



# Derivation

If you recall your calculus...

$$\ell(p) = \log \binom{n}{x} + x \log(p) + (n - x) \log(1 - p)$$

$$\ell'(p) = \frac{x}{p} - \frac{n - x}{1 - p} = 0$$

$$\frac{x}{p} = \frac{n - x}{1 - p}$$

$$x - xp = np - xp$$

$$p = \frac{x}{n}$$

So,  $\hat{p} = \frac{x}{n}$ .

# Case Study

## Example

Mouse genomes have have 19 non-sex chromosome pairs and X and Y sex chromosomes (females have two copies of X, males one each of X and Y). The total percentage of mouse genes on the X chromosome is 6.1%. There are 25 mouse genes involved in sperm formation. An evolutionary theory states that these genes are more likely to occur on the X chromosome than elsewhere in the genome (in an independence chance model) because recessive alleles that benefit males are acted on by natural selection more readily on the X than on autosomal (non-sex) chromosomes. In the mouse genome, 10 of 25 genes (40%) are on the X chromosome. This is larger than expected by an independence chance model, but how unusual is it?

# The Big Picture

- For proportions, the typical scenario is that there is a population which can be modeled as a large bucket with some proportion  $p$  of red balls.
- A *null hypothesis* is that the proportion is exactly equal to  $p_0$ .
- In a *random sample* of size  $n$ , we observe  $\hat{p} = X/n$  red balls.
- The sample proportion  $\hat{p}$  is typically not exactly equal to the null proportion  $p_0$ .
- A *hypothesis test* is one way to explore if the discrepancy can be explained by *chance variation consistent with the null hypothesis* or if there is statistical evidence that the null hypothesis is incorrect and that the data is better explained by *an alternative hypothesis*.
- Some legitimate uses of hypothesis tests for proportions does not fit into this framework, such as the next example.
- It is very important to interpret results carefully.



# Hypothesis Tests

- Conducting a hypothesis test consists of these steps:
  - ① State null and alternative hypotheses;
  - ② Compute a test statistic;
  - ③ Determine the null distribution of the test statistic;
  - ④ Compute a p-value;
  - ⑤ Interpret and report the results.
- We will examine these steps for this case study.

# Null and Alternative Hypotheses

## Definition

A *hypothesis* is a statement about a probability model.

## Definition

A *null hypothesis* is a specific statement about a probability model that would be interesting to reject. A null hypothesis is usually consistent with a model indicating no relationship between variables of interest. For proportions, a null hypothesis almost always takes the form  $H_0: p = p_0$ .

## Definition

An *alternative hypothesis* is a set of hypotheses that contradict the null hypothesis. For proportions, one-sided alternative hypotheses almost always takes the form  $H_A: p < p_0$  or  $H_A: p > p_0$  whereas two-sided alternative hypotheses take the form  $H_A: p \neq p_0$ .

# Stating Hypotheses

- In the example, the null hypothesis and alternative hypotheses are as follows.

$$H_0: p = 0.061$$

$$H_A: p > 0.061$$

- We choose the one-sided hypothesis  $p > 0.061$  because this is the interesting biological conclusion in this setting.

# Hypothesis Test Framework

- Note that here  $p_0 = 0.061$  refers to the probability in a hypothetical probability model, not an unknown proportion in a large population: we have observed all 25 genes in the population of mouse spermatogenesis genes and the observed proportion  $10/25 = 0.40$  of them on the X chromosome. These 25 genes are not a random sample from some larger population of mouse spermatogenesis genes.
- Here is the question of interest is:  
*If the location of genes in the mouse genome were independent of the function of the genes, would we expect to see as many spermatogenesis genes on the X chromosome as we actually observe?*

We are comparing the observed proportion to its expected value under a hypothetical probability model.

# Compute a Test Statistic

- The observed number of genes, on the X chromosome, here  $X = 10$ , is the test statistic.
- Other approaches for proportions will use other test statistics, such as one based on the normal distribution.

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

# Null Distribution

## Definition

The *null distribution of a test statistic* is the sampling distribution of the test statistic, assuming that the null hypothesis is true.

- Here, we assume  $X \sim \text{Binomial}(25, 0.061)$ .
- The expected value of this distribution is  $E(X) = 25(0.061) \doteq 1.52$ .
- The standard deviation is  $SD(X) = \sqrt{25(0.061)(0.939)} \doteq 1.2$ .
- We note that the observed value is quite a few standard deviations above the mean.

# Compute the P-value

## Definition

The *p-value* is the probability of observing a test statistic at least as extreme as that actually observed, assuming that the null hypothesis is true. The outcomes *at least as extreme* as that actually observed are determined by the alternative hypothesis.

- In the example, observing ten or more genes on the X chromosome,  $X \geq 10$ , would be at least as extreme as the observed  $X = 10$ .
- The null distribution is  $X \sim \text{Binomial}(25, 0.061)$ , so

$$\begin{aligned} P(X \geq 10) &= P(X = 10) + P(X = 11) + \cdots + P(X = 25) \\ &\doteq 9.9 \times 10^{-7} \end{aligned}$$

- In other words, only about 1 in a million random  $X$ s from  $\text{Binomial}(25, 0.061)$  distributions take on the value 10 or more.
- This is a very small probability.

# Interpretation

- If the p-value is very small, this is used as evidence that the null hypothesis is incorrect and that the alternative hypothesis is true.
- The logic is that if the null hypothesis were true, we would need to accept that a rare, improbable event just occurred; since this is very unlikely, a better explanation is that the alternative hypothesis is true and what actually occurred was not uncommon.
- There is no universal cut-off for a *small p-value*, but  $P < 0.05$  is a commonly used range to call the results of a hypothesis test *statistically significant*.
- More formally, we can say that a result is *statistically significant at the  $\alpha = 0.05$  level* if the p-value is less than 0.05. (Other choices of  $\alpha$ , such as 0.1 or 0.01 are also common.)
- Note, however, that results  $P = 0.051$  and  $P = 0.049$ , while on opposite sides of 0.05, quantify strength of evidence against the null hypothesis almost identically.



# Reporting Results

- The report of a hypothesis test should include:
  - ▶ the value of the test statistic;
  - ▶ the sample size;
  - ▶ the p-value; and
  - ▶ the name of the test.

In the example,

*The proportion of spermatogenesis genes on the X chromosome,  $10/25 = 0.40$ , is significantly larger than the proportion of all genes on the X chromosome, 0.061, (binomial test,  $P = 9.9 \times 10^{-7}$ ).*

# Applicability

- The *binomial test* for proportions assumes a binomial probability model.
- The binomial distribution is based on assumptions of *a fixed number of independent, equal-probability, binary outcomes*.
- The assumption of independence is questionable; genes that work together are often located near each other as operons, clusters of related genes that are coregulated.
- A hypothesis that many of the genes would cluster together, whether on the X chromosome or not, is an alternative biological explanation of the observed results.
- The conclusion the observed  $X$  is inconsistent with a  $\text{Binomial}(25, 0.061)$  model could be because the true  $p$  is larger, but the binomial model fits, but also because of lack of appropriateness of the binomial model itself.
- It would help to know more about the specific genes and the underlying biology to better assess the strength of support for the evolutionary hypothesis.

# Using R to Compute the P-Value

- In this example, computing the p-value by hand would be quite tedious as it requires summing many separate binomial probabilities.
- R automates this calculation.
- The functions, `sum()`, `dbinom()`, and the colon operator combine to compute the p-value.
- Here `10:25` creates a sequence from 10 to 25.
- `dbinom()` (d for density, binom for binomial) takes three arguments: first one or more possible values of the random variable, second the sample size  $n$ , and third the success probability  $p$ . The expression `dbinom(10:25, 25, 0.061)` creates a vector of the individual probabilities.
- Finally, use `sum()` to sum the probabilities.  

```
> sum(dbinom(10:25, 25, 0.061))  
  
[1] 9.93988e-07
```

## Another Example

### Example

Example 6.4 on page 138 describes the mud plantain *Heteranthera multiflora* in which the female sexual organ (the style) and male sexual organ (the anther) deflect to different sides. The effect is that if a bee picks up pollen from an anther on the right, it will only deposit the pollen on a plant with a style on the right, and thus avoid self-pollination. The *handedness* (left or right) of the plants describes the location of the style. Crosses of pure-strain left- and right-handed plants result in only right-handed offspring. Under a simple one-gene complete dominance/recessive genetic model,  $p = 0.25$  of the offspring from a second cross between offspring of the first cross should be left-handed. In the experiment there are 6 left-handed offspring and 21 right-handed offspring. Test the hypothesis that  $p = 0.25$ .

# The Hypotheses

- The hypotheses are:

$$H_0: p = 0.25$$

$$H_A: p \neq 0.25$$

where  $p$  is the probability that an offspring is left-handed from the given cross of right-handed  $F_1$  generation plants.

- We select a two-sided test as it is biologically interesting if the true probability is either smaller or larger than 0.25, and we have no a priori reason to expect a deviation in either direction.

# The Test

- Let  $X = \#$  of left-handed offspring.
- Under  $H_0$ ,  $X \sim \text{Binomial}(27, 0.25)$ .
- The expected value of this distribution is  $\mu = E(X) = 27(0.25) = 6.75$ .
- The observed value  $X = 6$  is 0.75 below the mean.
- The value  $6.75 + 0.75 = 7.5$  is the same distance above the mean.
- The probability of being at least as far from the expected value as the actual data is

$$\begin{aligned} P &= P(X \leq 6) + P(X \geq 7.5) \\ &= P(X \leq 6) + P(X \geq 8) \\ &= 1 - P(X = 7) \\ &\doteq 0.828 \end{aligned}$$

- $P = 0.828$  is not a small p-value.
- The data *is consistent with the null hypothesis*.

# Interpretation

*The proportion of left-handed offspring,  $\hat{p} = 6/27 \doteq 0.222$  is consistent with the probability  $p = 0.25$  predicted by the one-gene complete dominance model ( $P = 0.828$ , binomial test).*

# Comparison with the Text Method

- The text describes finding p-values for the binomial test by doubling the p-value from a one-sided test.
- As the binomial distribution is only perfectly symmetric when  $p = 0.5$ , this method employs a needless approximation, but the numerical values will be close to those computed by the method in the notes.



## Comparison with R

- The function `binom.test()` in R determines extreme values using a likelihood-based criterion: the p-value is the sum of probabilities of all outcomes with probabilities equal to or less than that of the outcome.
- In this example,  $P(X = 6) = P(X = 7)$ , so the p-value is computed as  $P(X \leq 6) + P(X \geq 7) = 1$ .

```
> binom.test(6, 27, p = 0.25, alternative = "two.sided")
```

Exact binomial test

data: 6 and 27

number of successes = 6, number of

trials = 27, p-value = 1

alternative hypothesis: true probability of success is not equal to 0.25

95 percent confidence interval:

0.08621694 0.42258306

sample estimates:

probability of success

0.2222222

# What you should know

You should know:

- how to construct a confidence interval for  $p$ ;
- how to conduct a hypothesis test about  $p$  with a binomial test;
- how to interpret confidence intervals and hypothesis tests in a biological context;
- what assumptions are inherent to these inference methods.

# Cautions

- Statistical inference about proportions assumes a definition of a population proportion or probability  $p$ ; make sure it is understood what this represents.
- The methods assume random sampling from the population of interest; when the data is not collected from a random sample, other background information is necessary to justify inference to populations of interest;
- Inference based on the binomial distribution assumes independent, equal-probability, fixed-sample-size, binary trials; if the assumptions are not met, inference can mislead.

# Extensions

- All focus so far has been on single populations; however, many interesting biological questions involve *comparisons between two or among three or more populations*. This topic comes soon.
- When individuals are classified by two categorical variables with two or more levels for each variable, the resulting data can be summarized in a *contingency table*. This topic comes soon as well.
- If there are also available other variables measured on individuals (quantitative or categorical or both), more advanced statistical methods will model individual probabilities as functions of these covariates.
- A class of statistical methods with binary response variables and some covariates are known as *logistic regression models*.

# R Appendix

- See the R handout to learn to:
  - ▶ Create bar graphs with the function `barchart()`;
  - ▶ Calculate binomial probabilities with `dbinom()` and `pbinom()`;
  - ▶ Generate random binomial samples with `rbinom()`;
  - ▶ Write a function to graph the binomial distribution;
  - ▶ Write a function for confidence intervals using the text method;
  - ▶ Use the function `binom.test()` for exact binomial hypothesis tests and confidence intervals.