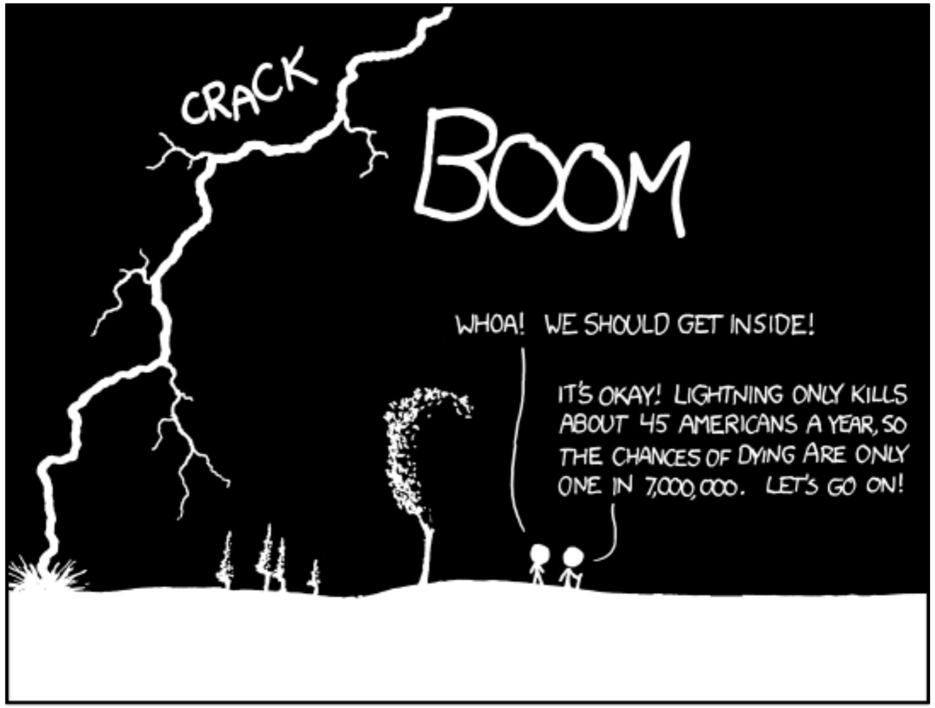
Probability and Discrete Variables

Biology 683

Coding Session: Saturday 14
9-12 BSBE 115


CRACK
BOOM
WHOA! WE SHOULD GET INSIDE!
IT'S OKAY! LIGHTNING ONLY KILLS ABOUT 45 AMERICANS A YEAR, SO THE CHANCES OF DYING ARE ONLY ONE IN 7,000,000. LET'S GO ON!

THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Size is often exponentially distributed in populations. Imagine that you sample a population and get the mean size in your sample. You then repeat this process many times; what will the distribution of your means look like?

What is the definition of a p-value?

Analyzing Proportions

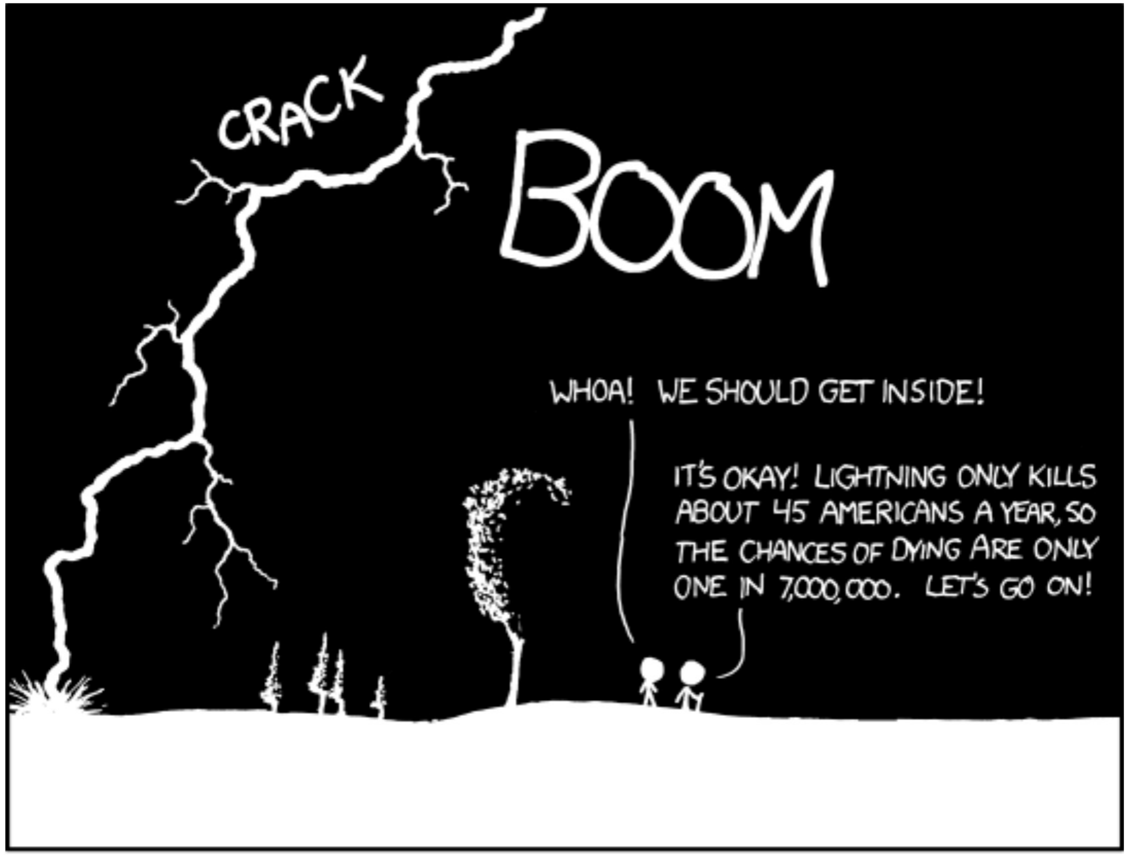
Several chapters in intro books deal with this topic

The experiment boils down to this:

- Your subjects have some alternative outcomes
- Each individual has some probability of each outcome
- You are trying to find the conditions that impact that probability

When would this type of problem come up in the biological sciences?

Terms to know for probability


CRACK
BOOM
WHOA! WE SHOULD GET INSIDE!
IT'S OKAY! LIGHTNING ONLY KILLS ABOUT 45 AMERICANS A YEAR, SO THE CHANCES OF DYING ARE ONLY ONE IN 7,000,000. LET'S GO ON!

THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Conditional probability: The probability of being struck by lightning changes if you know the statistic about 45 American a year being killed by lightning

A simple example

What is the probability of drawing an ace then a king from a deck cards?

What is the sample space?

Are the events independent or dependent?

What is the probability of this event?

What is the conditional probability of drawing a king if we have already drawn an ace?

assume standard deck with no jokers

Binomial Test

A test to determine whether or not the observed proportion adheres to the expected proportion under the null hypothesis

Some possible uses:

- Are frogs equally likely to be right or left handed?
- Is the sex ratio half male and half female?
- Are the offspring phenotypes a 3:1 ratio?
- Do some beetles win more fights?

Binomial Test

As in most statistical tests, a test statistic is compared to a distribution

In this case, the test statistic is just the observed number (number of right-handed toads, number of females in the population, number of fights won)

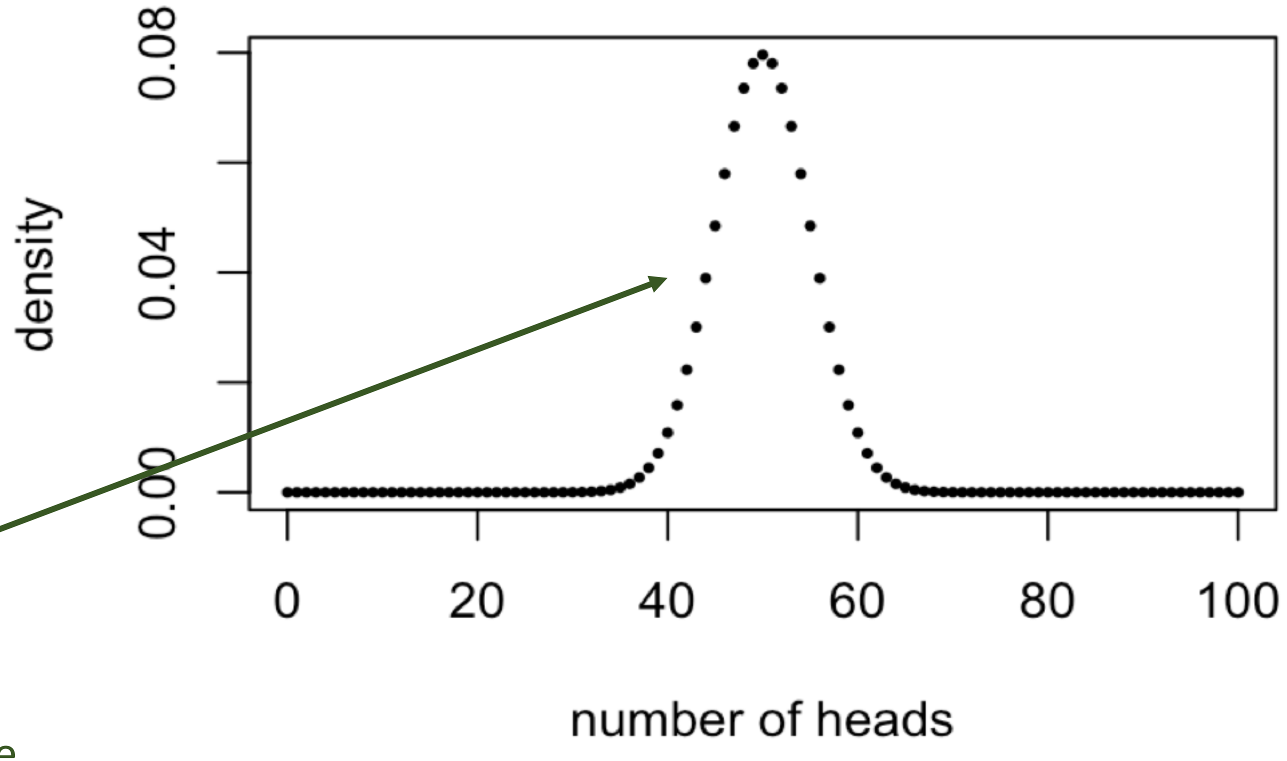
Note that this test is only appropriate when there are two categories of individuals and your hypothesis allows you to provide a probability of the outcomes.

Binomial Test

With the binomial test our null hypothesis is the probability of one of the two outcomes. This probability and the number of observations defines the distribution we will compare our observation to.

Distribution when the null is 50% and we have 100 observations

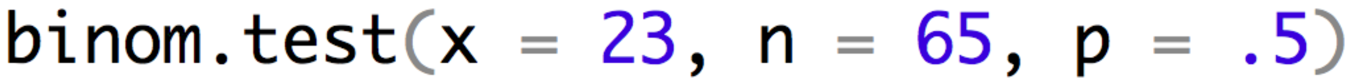
```
x <- 0:100
y <- dbinom(x, size = 100, prob = .5)
plot(y~x, pch=16, cex=.5,
     xlab="number of heads")
```



use a simulation to see if you can replicate this curve

Binomial Test

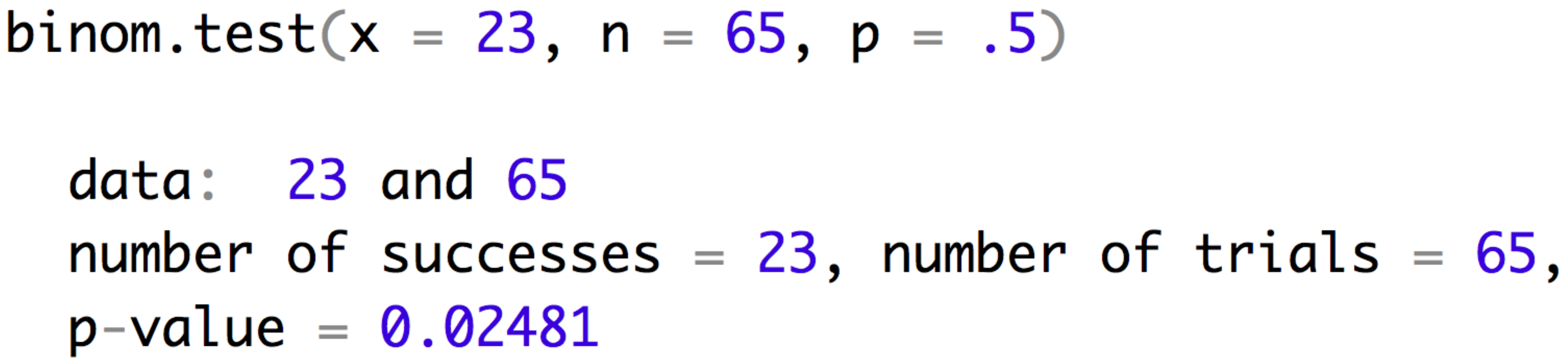
Lets look at an example with sex ratio. You are hybridizing closely related species (with XY sex chromosomes) so you know Haldane’s rule states that the males might be more rare. When you survey the offspring you find 23 males out of 65 offspring. Does this result support Haldane’s rule occuring in your system?



```
binom.test(x = 23, n = 65, p = .5)
```

Binomial Test

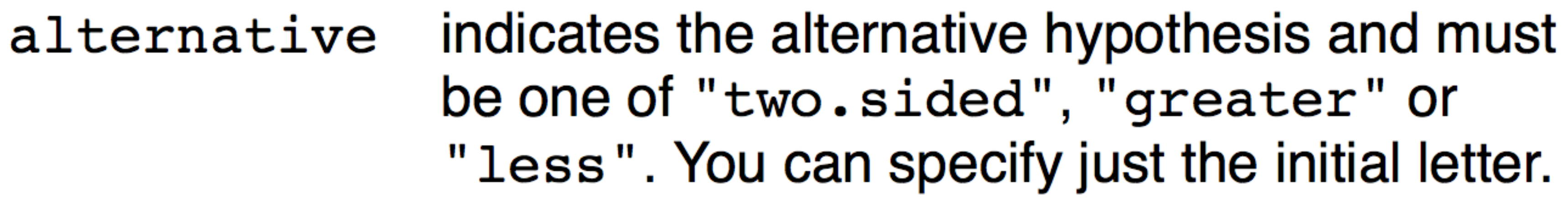
Lets look at an example with sex ratio. You are hybridizing closely related species (with XY sex chromosomes) so you know Haldane’s rule states that the males might be more rare. When you survey the offspring you find 23 males out of 65 offspring. Does this result support Haldane’s rule occuring in your system?



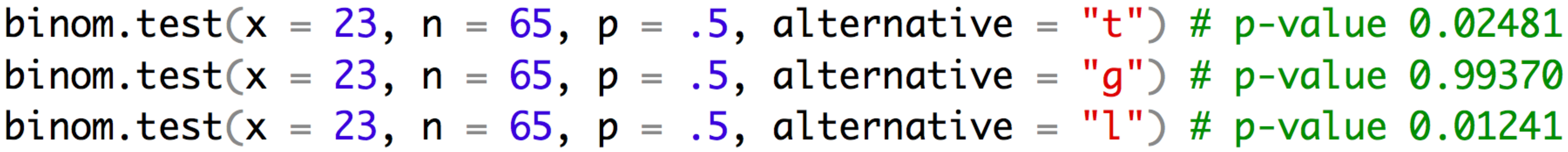
```
binom.test(x = 23, n = 65, p = .5)

  data:  23 and 65
  number of successes = 23, number of trials = 65,
  p-value = 0.02481
```

Binomial Test

binom.test has an argument `alternative`



`alternative` indicates the alternative hypothesis and must be one of `"two.sided"`, `"greater"` or `"less"`. You can specify just the initial letter.



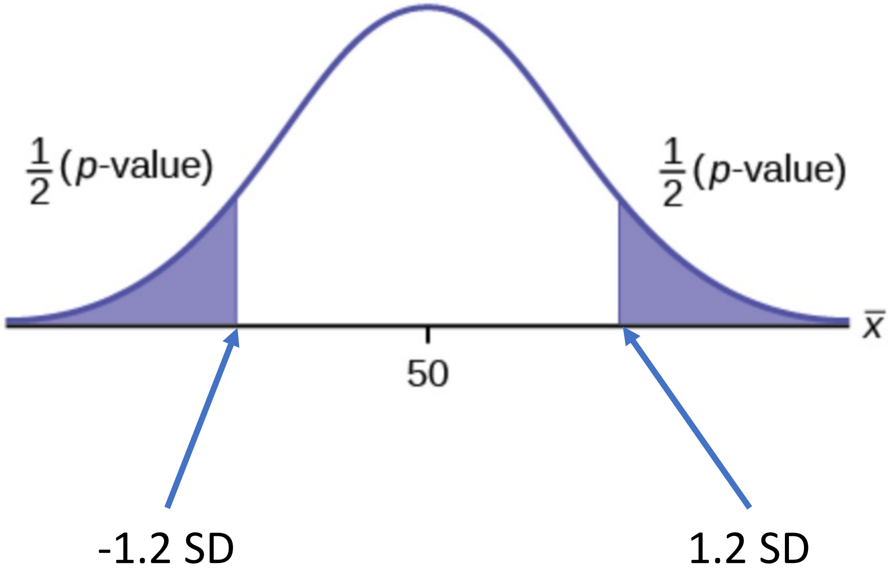
```
binom.test(x = 23, n = 65, p = .5, alternative = "t") # p-value 0.02481
binom.test(x = 23, n = 65, p = .5, alternative = "g") # p-value 0.99370
binom.test(x = 23, n = 65, p = .5, alternative = "l") # p-value 0.01241
```

Binomial Test

Alternative = two.sided

What is the probability that I would see a skew in the sex ratio this great or greater.

In this case our observed number of males was -1.2 standard deviations from the mean. So our p-value is the area under the curves above 1.2SD and below -1.2SD.



-1.2 SD

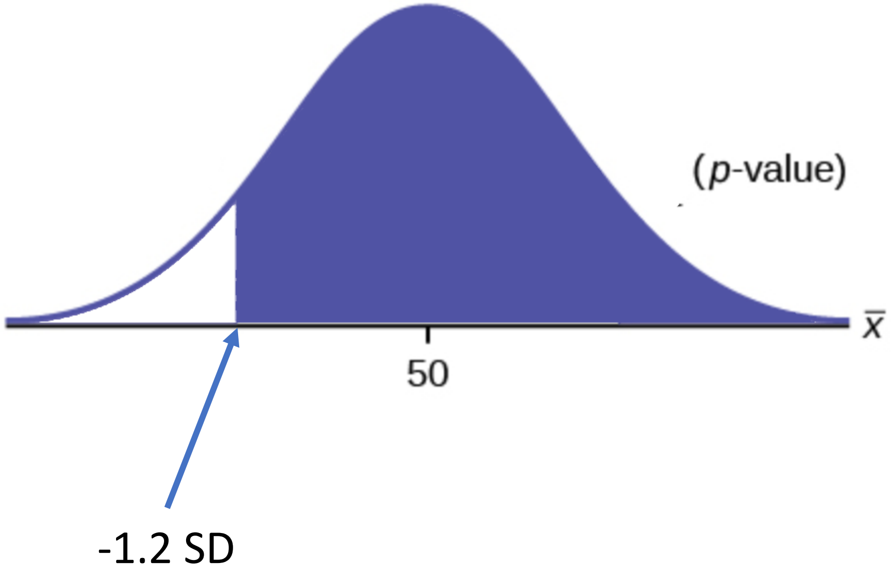
1.2 SD

Binomial Test

Alternative = greater

What is the probability that I would see a larger number of males.

In this case our observed number of males was -1.2 standard deviations from the mean. So our p-value is the area under the curves above -1.2SD.

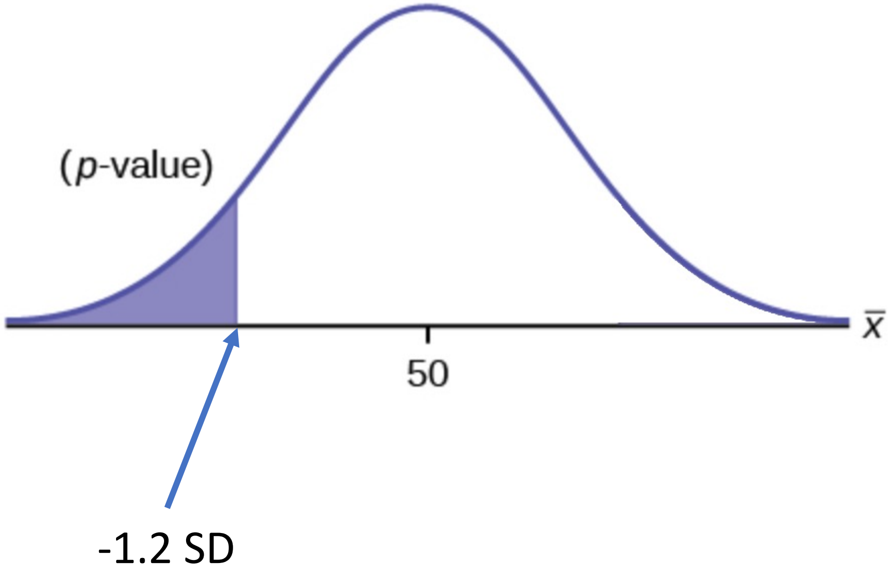


Binomial Test

Alternative = less

What is the probability that I would see this many or fewer males.

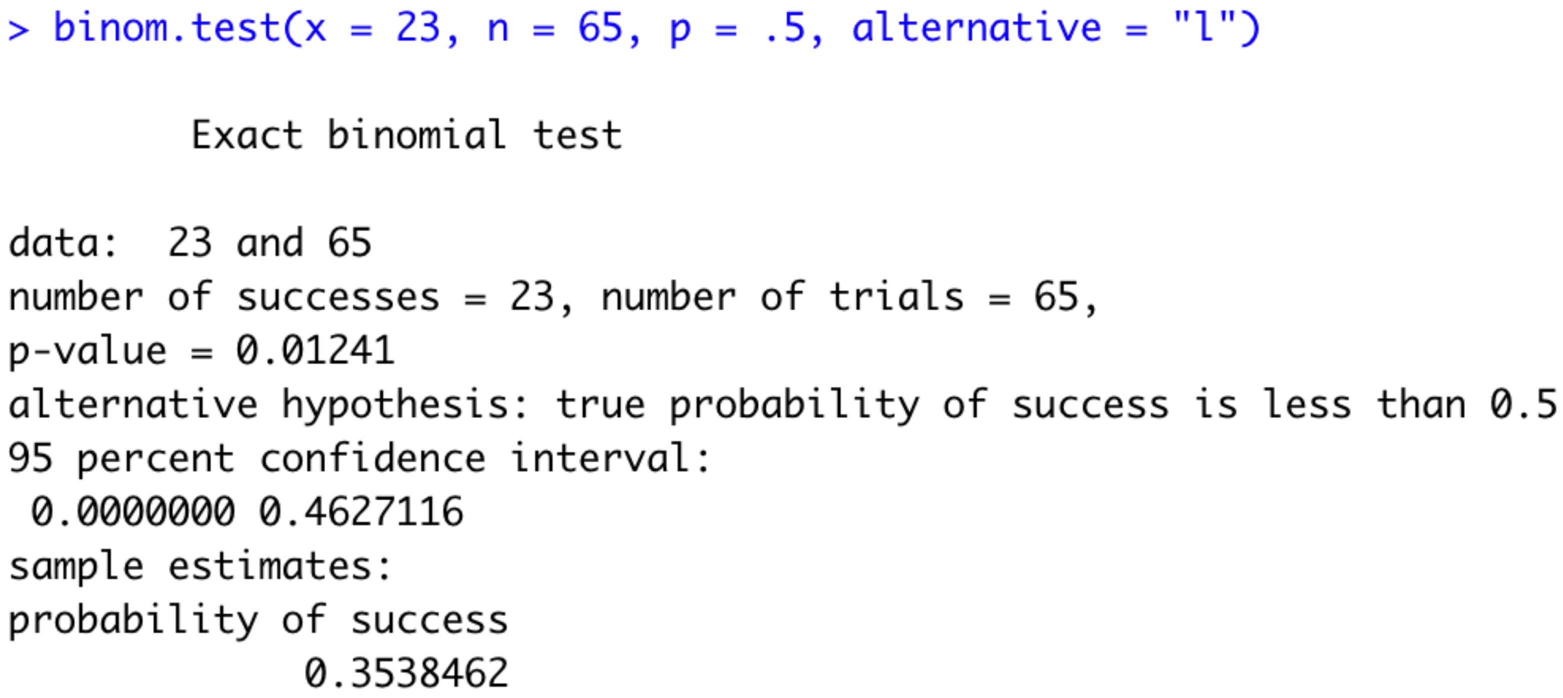
In this case our observed number of males was -1.2 standard deviations from the mean. So our p-value is the area under the curves below -1.2SD.



-1.2 SD

If before we collected this data we wanted to answer the question: “Are males more rare than females?” Then we have an apriori hypothesis that we should see fewer males than females so we could justify using this more powerful test.

Reporting the Results



```
> binom.test(x = 23, n = 65, p = .5, alternative = "l")

	Exact binomial test

data:  23 and 65
number of successes = 23, number of trials = 65,
p-value = 0.01241
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.4627116
sample estimates:
probability of success 
             0.3538462 
```

Our offspring ratio shows a significantly fewer males than would be expected under a 1:1 sex ratio (0.35, 95% CI: 0.24-0.48, binomial test, $n = 65$, $p < 0.025$).

Reporting the Results

This populations shows a significantly fewer males than would be expected under a 1:1 sex ration (0.35, 95% CI: 0.24-0.48, binomial test, $n = 65$, $p < 0.025$).

For very small *p*-values, we just say that *p* is very small (< 0.001 or < 0.0001).

Most journals/subdisciplines will have conventions about how certain tests are reported.

Most journals italicize mathematical variables, so *n* and *p* would be italicized. They also normally would be lower case.

χ^2 Test

This test compares the observed number in each category to expectations based on the null hypothesis (if there are only two categories, it approximates the binomial test with probability of 50%)

It can also be used to test for independence of two variables, and then it is called a contingency χ2-test.

We will use data from the Titanic and see if some females were more likely to survive than others.

Female adults on the Titanic

	Survived	Died
1st	140	4
2nd	80	13
3rd	76	89
Crew	20	3

χ^2 Test

To calculate the statistic we just sum up the standardized deviations from the expected values in each category.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

χ^2 Test

To calculate the statistic we just sum up the standardized deviations from the expected values in each category.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Female adults on the Titanic			
	Survived	Died	total
1st	140	4	144
2nd	80	13	93
3rd	76	89	165
Crew	20	3	23
total	74.4%	25.6%	

χ^2 Test

To calculate the statistic we just sum up the standardized deviations from the expected values.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Female adults on the Titanic			
	Survived	Died	
1st	140	4	144
2nd	80	13	93
3rd	76	89	165
Crew	20	3	23
total	74.4%	25.6%	

Expected		
	Survived	Died
1st	.744 x 144	.256 x 144
2nd	.744 x 93	.256 x 93
3rd	.744 x 165	.256 x 165
Crew	.744 x 23	.256 x 23

χ^2 Test

To calculate the statistic we just sum up the standardized deviations from the expected values.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

Female adults on the Titanic			
	Survived	Died	
1st	140	4	144
2nd	80	13	93
3rd	76	89	165
Crew	20	3	23
total	74.4%	25.6%	

Expected		
	Survived	Died
1st	107	37
2nd	69	24
3rd	123	42
Crew	17	6

χ^2 Test

To calculate the statistic we just sum up the standardized deviations from the expected values.

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

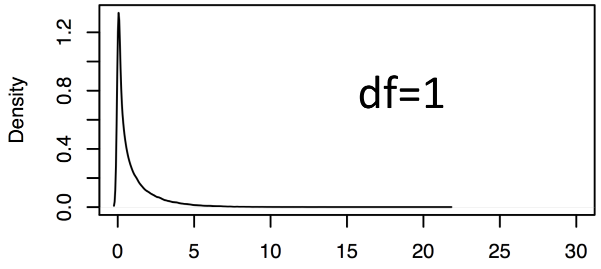
Observed		
	Survived	Died
1st	140	4
2nd	80	13
3rd	76	89
Crew	20	3

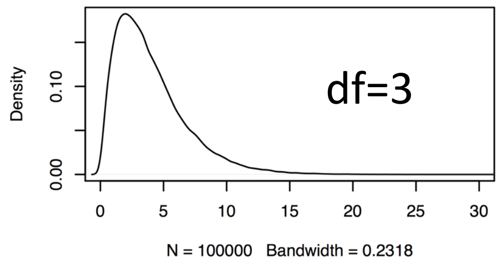
Expected		
	Survived	Died
1st	107	37
2nd	69	24
3rd	123	42
Crew	17	6

$$\chi^2 = 117$$

χ^2 Test

The shape of the chi square distribution depends on the degrees of freedom (df).





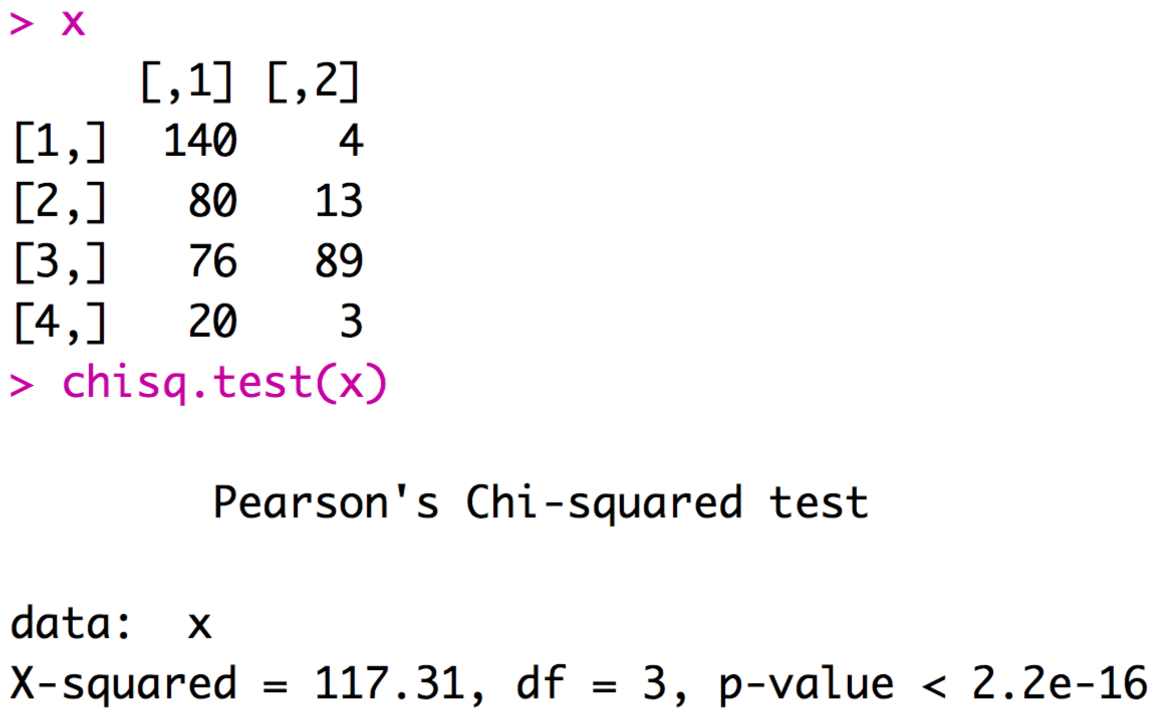
df = (no. rows – 1)(no. cols -1)

Female adults on the Titanic

	Survived	Died
1st	140	4
2nd	80	13
3rd	76	89
Crew	20	3

df = (4– 1)(2 -1)

df = 3



```
> x
     [,1] [,2]
[1,]  140    4
[2,]   80   13
[3,]   76   89
[4,]   20    3
> chisq.test(x)

	Pearson's Chi-squared test

data:  x
X-squared = 117.31, df = 3, p-value < 2.2e-16
```

My take

Much of what was presented today forms the basis of logically thinking about your experimental results. Similarly these topics are key to your ability to evaluate the evidence presented in the papers that you read. I will definitely ask you multiple questions on your midterm and final that will require you to think clearly about these topics.