Chi-squared tests: Goodness of fit 2

Research Methods for Human Inquiry
Andrew Perfors

Remember how to build a statistical test

1) A diagnostic test statistic, T
2) Sampling distribution of T if the null is true
3) The observed T in your data
4) A rule that maps every value of T onto a decision (accept or reject H0)

Let's construct a test statistic

- Last lecture we talked about using mean, but mean doesn't make much sense here…

Leave (B)	Attack (D)	Rescue (G)	Analyse (S)
2	55	36	7

= ??

BUNNY	DOGGIE	GLADLY	SHADOW
0.125	0.455	0.334	0.086

- Intuitively what we want is some measure of how closely the two of these match…

$\boxed{\text{The Goodness of Fit statistic}}$

Our test statistic: Goodness of fit (GOF)

- The <u>expected frequencies</u>... What would we expect the observed frequencies to be if the null hypothesis were true?

$$E_i = N \times \theta_i$$

The number of people we would "expect" to say they voted for each person i if the null hypothesis is true...

... multiplied by the total number of people in our study

100

... is equal to the probability that the null hypothesis predicts (i.e., the probability in the electoral data)

Leave (B)	Attack (D)	Rescue (G)	Analyse (S)
12.5	45.5	33.4	8.6

BUNNY	DOGGIE	GLADLY	SHADOW
0.125	0.455	0.334	0.086

Our test statistic: Goodness of fit (GOF)

	Expected, E_i	Observed, O_i
Leave (B)	12.5	2
Attack (D)	45.5	55
Rescue (G)	33.4	36
Analyse (S)	8.6	7

Maybe our test statistic should "compare" these?

Our test statistic: Goodness of fit (GOF)

	Expected, E_i	Observed, O_i	O_i - E_i
Leave (B)	12.5	2	10.5
Attack (D)	45.5	55	-9.5
Rescue (G)	33.4	36	-2.6
Analyse (S)	8.6	7	1.6

Deviations from what the
null hypothesis "expects"

Our test statistic: Goodness of fit (GOF)

	Expected, E_i	Observed, O_i	$(O_i - E_i)^2$
Leave (B)	12.5	2	110.25
Attack (D)	45.5	55	90.25
Rescue (G)	33.4	36	6.76
Analyse (S)	8.6	7	2.56

Just as we did with standard deviation, we'll make sure these are non-negative by squaring

Our test statistic: Goodness of fit (GOF)

	Expected, E_i	**Observed, O_i**	$\frac{(O_i - E_i)^2}{E_i}$
Leave (B)	12.5	2	8.82
Attack (D)	45.5	55	1.983
Rescue (G)	33.4	36	0.202
Analyse (S)	8.6	7	0.298

Then divide by expected frequencies.

(Technical reasons, but basically it makes the numbers smaller whilst squaring made them very large)

Our test statistic: Goodness of fit (GOF)

	Expected, E_i	Observed, O_i	$\frac{(O_i - E_i)^2}{E_i}$
Leave (B)	12.5	2	8.82
Attack (D)	45.5	55	1.983
Rescue (G)	33.4	36	0.202
Analyse (S)	8.6	7	0.298

Remember a test statistic is just a single number, so let's add these together

11.303

Our test statistic: Goodness of fit (GOF)

The equation (where k is the number of categories - here $k=4$)

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

	Expected, E_i	Observed, O_i	$\frac{(O_i - E_i)^2}{E_i}$
Leave (B)	12.5	2	8.82
Attack (D)	45.5	55	1.983
Rescue (G)	33.4	36	0.202
Analyse (S)	8.6	7	0.298

Remember a test statistic is just a single number, so let's add these together

11.303

Larger values of the X^2 statistic mean a worse fit to the data

Note: you do not need to memorise this equation for the exam

Doing it in R

```
> ed
 bunny doggie gladly shadow
 0.125  0.455  0.334  0.086
```

our workspace

```
> votingTable <- table(d$vote)
> votingTable

 bunny doggie gladly shadow
     2     55     36      7
```

```
> O <- votingTable
> E <- ed * 100
```

N

calculating O and E from this

```
> Xsquared <- sum( (O-E)^2 / E )
> Xsquared
[1] 11.30359
```

the X^2 value

So…

We need a few things…

X^2 ✔ 1) A diagnostic test statistic, *T*

2) Sampling distribution of *T* if the null is true

11.303 ✔ 3) The observed *T* in your data

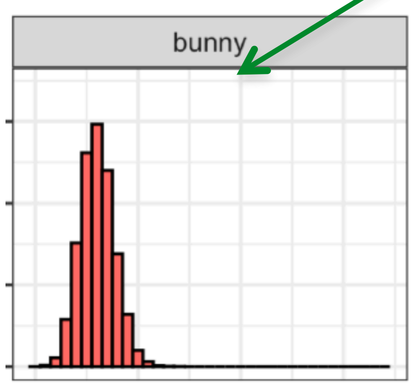
4) A rule that maps every value of *T* onto a decision (accept or reject H0)

Sampling distribution of the test statistic (X^2) if the null hypothesis is true

Simulate what you'd expect if the null were true

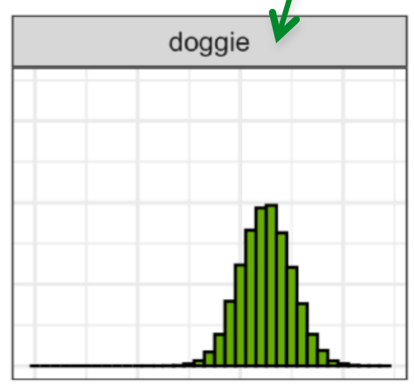
$H_0 : \boldsymbol{\theta} = (0.125, 0.455, 0.334, 0.086)$

O_1



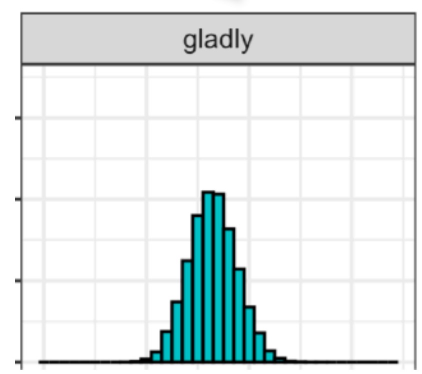
`rbinom(n=25000, size=100,prob=0.125))`

O_2



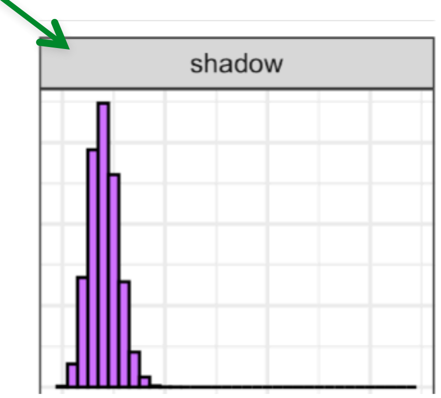
`rbinom(n=25000, size=100,prob=0.455)`

O_3



`rbinom(n=25000, size=100,prob=0.334)`

O_4



`rbinom(n=25000, size=100,prob=0.086)`

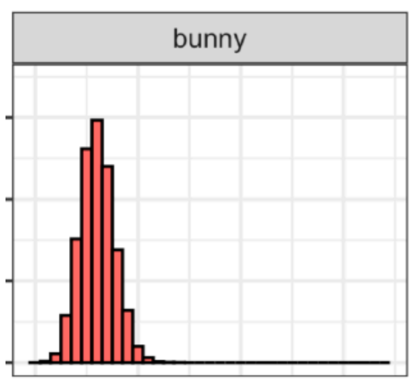
distributions of outcomes under the null hypothesis

Predicts that you'd generate each observation O with a binomial distribution in which θ_i is the probability

Sampling distribution of the test statistic (X^2) if the null hypothesis is true

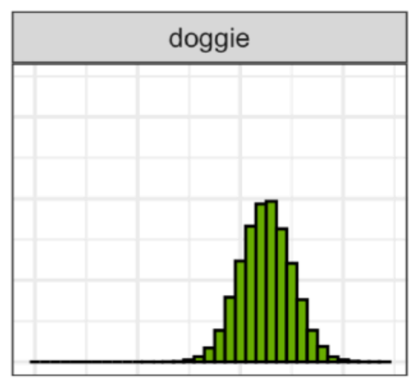
As sample size grows large enough, binomial distributions are normal.
So with large enough samples, this is a bunch of normal distributions

O_1



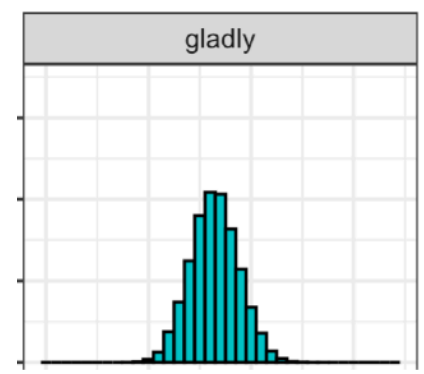
`rbinom(n=25000, size=100,prob=0.125))`

O_2



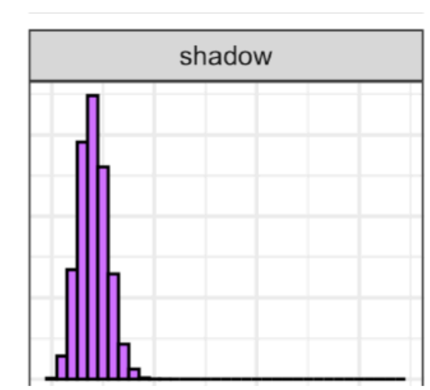
`rbinom(n=25000, size=100,prob=0.455)`

O_3



`rbinom(n=25000, size=100,prob=0.334)`

O_4



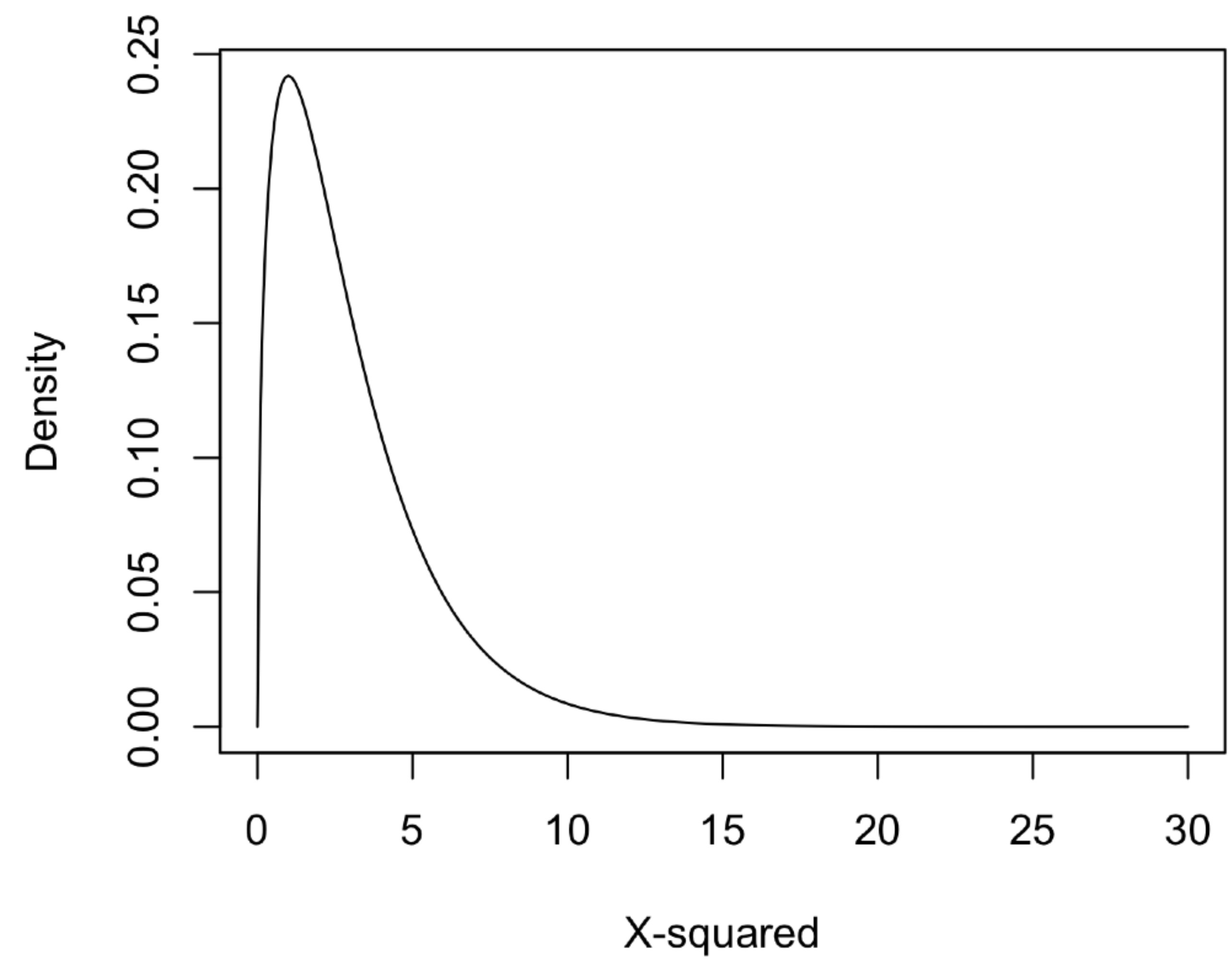
`rbinom(n=25000, size=100,prob=0.086)`

X^2 just takes these, squares them, and adds them

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

Sampling distribution of the test statistic (X^2) if the null hypothesis is true

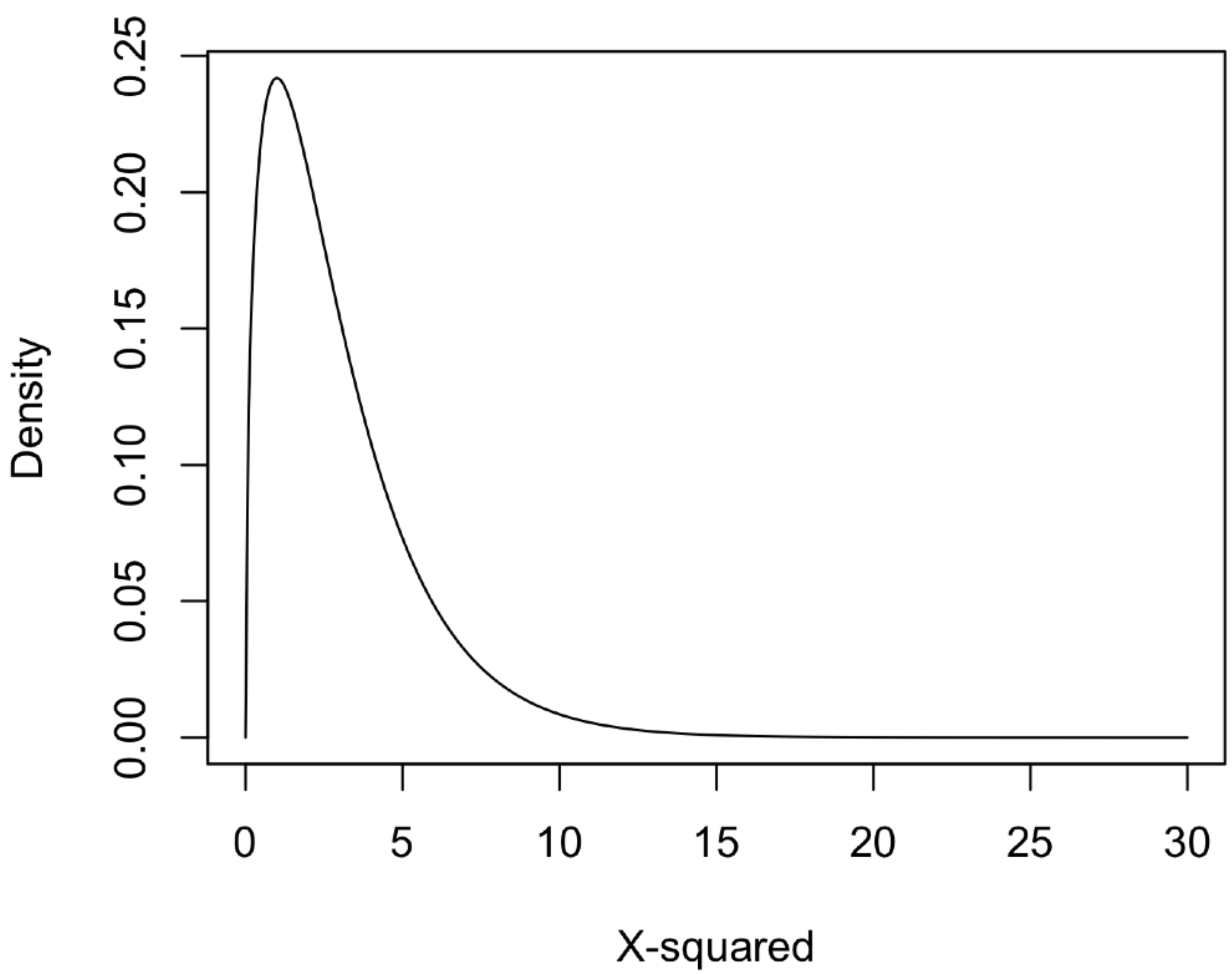
Karl Pearson: pointed out that the **chi-squared distribution** (χ^2) is what you get when you take normally distributed data, square it, and add it



X^2 just takes these, squares them, and adds them

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

The chi-square (χ^2) distribution

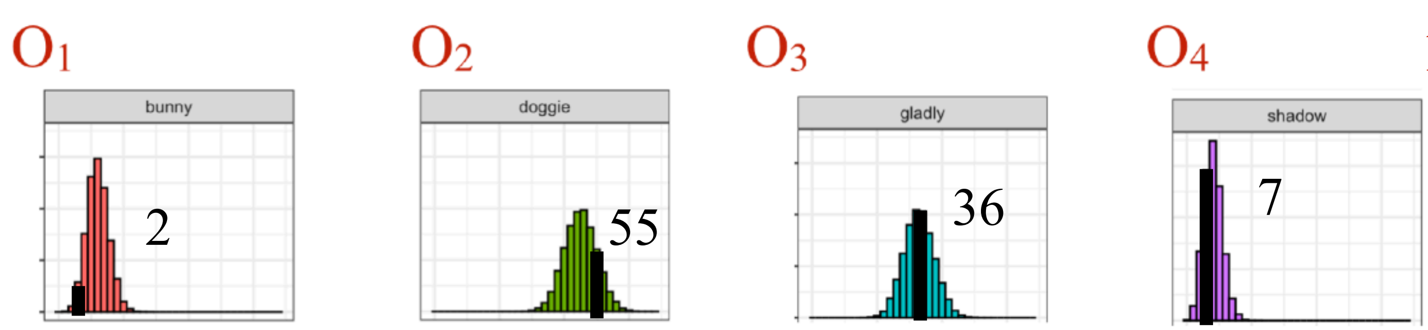


- Continuous distribution
- Has a noticeable positive skew to it
- The shape of the distribution depends on the "**degrees of freedom**"

What is "degrees of freedom"?

- A simple definition...
  - The number of "degrees of freedom" (*df*) in your data are the total number of "things" you're interested in minus the number of known constraints on those "things"

Four observations, so four “things” in this data



has to be 7, since 100-36-55-2=7

Our sample size is 100, so the # of total observations must sum to 100. This is **one** constraint, so df=3

Why does this matter? Think about what we’re assuming about how observations are generated when we calculate the X^2 statistic

Example: the *df* for our voting data

Option voted for	Observed frequency
Leave (B)	2
Attack (D)	55
Rescue (G)	36
Analyse (S)	7
total	100

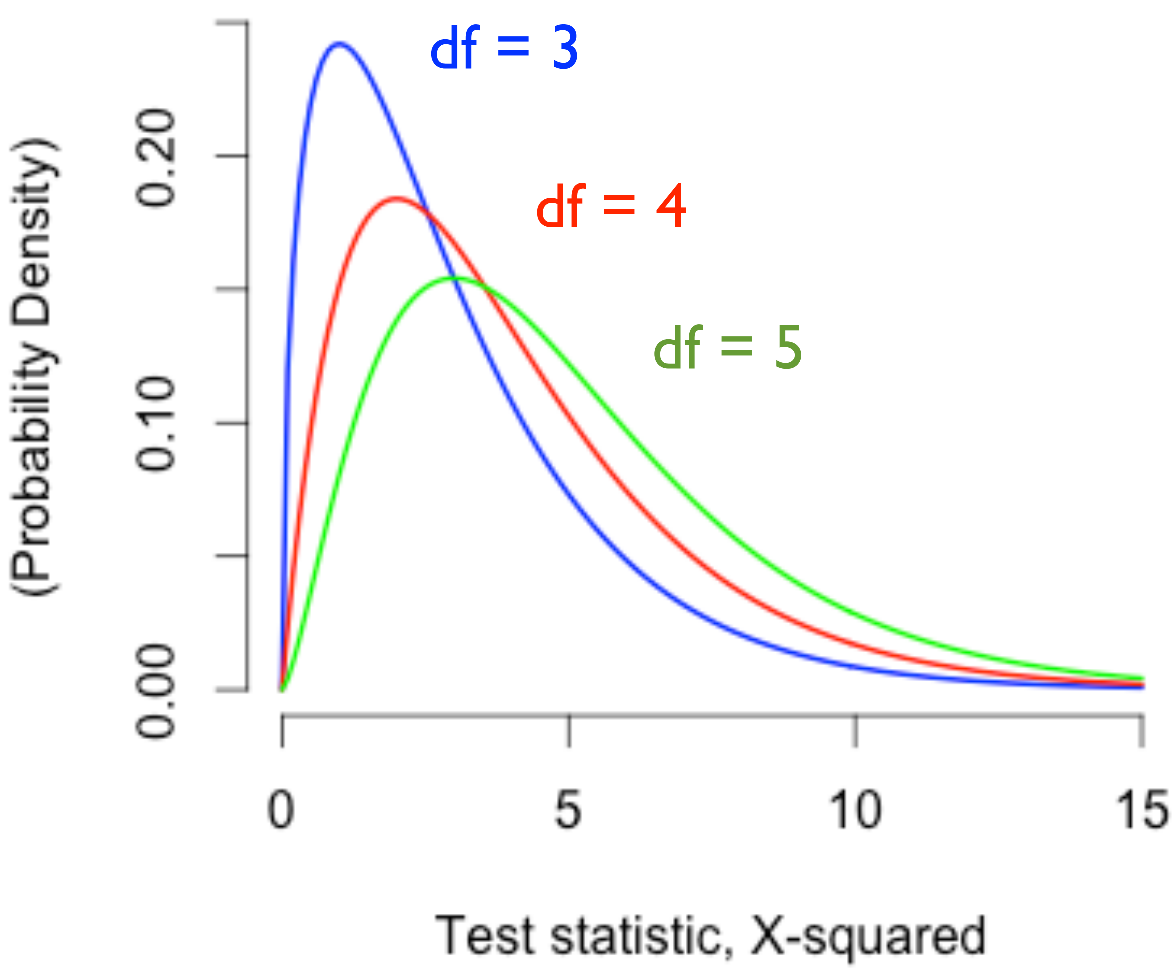
Four quantities of interest in the data

One constraint on those quantities

= Three degrees of freedom

More precisely

- For a chi-square goodness of fit test involving k categories, the degrees of freedom is equal to k-1
- Here's how the chi-square distribution changes as the degrees of freedom increases...



More precisely

- We can manually demonstrate this in R just to satisfy ourselves that I’m not making stuff up

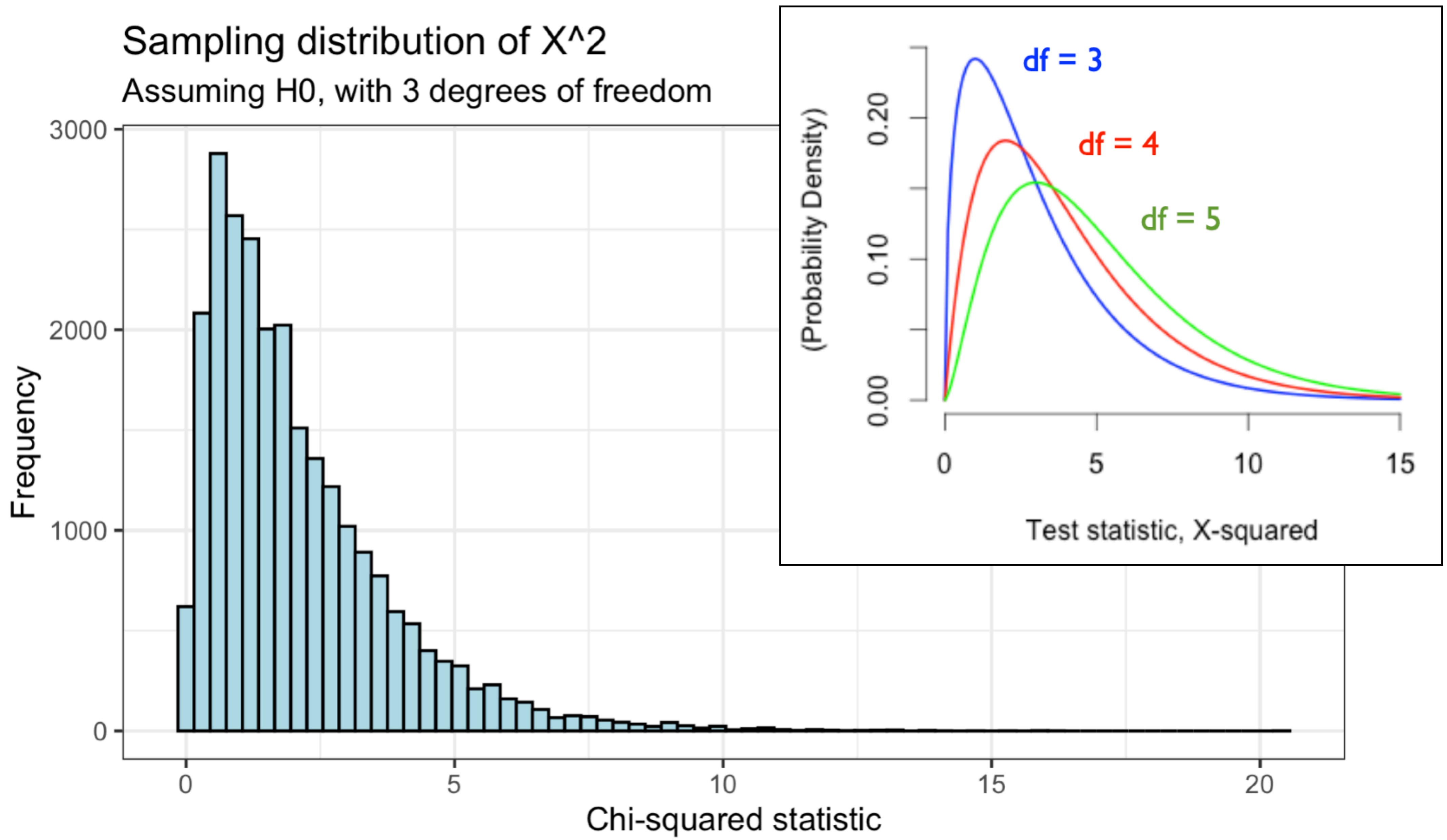
```
# now calculate chi-squared sampling distribution under null
N <- 100
longdb <- tibble(bunny,doggie,gladly,shadow)
# calculates (O-E)^2/E for each of the simulated draws
longdb <- longdb %>%
  mutate(chsBunny = (bunny-N*ed[["bunny"]])^2/(N*ed[["bunny"]]),
         chsDoggie = (doggie-N*ed[["doggie"]])^2/(N*ed[["doggie"]]),
         chsGladly = (gladly-N*ed[["gladly"]])^2/(N*ed[["gladly"]]),
         chsShadow = (shadow-N*ed[["shadow"]])^2/(N*ed[["shadow"]]))
# sums them up for each of the stimulated draws
# this is the sampling distribution for the chi-squared
# statistic under the null hypothesis
longdb <- longdb %>%
  mutate(chisq = chsBunny+chsDoggie+chsGladly)

longdb %>%
  ggplot(mapping = aes(x=chisq)) +
  geom_histogram(colour="black",binwidth=0.3,fill="lightblue") +
  theme_bw() +
  labs(title = "Sampling distribution of X^2",
       subtitle = "Assuming H0, with 3 degrees of freedom",
    x = "Chi-squared statistic",
    y = "Frequency")
```

* This will not be assessed

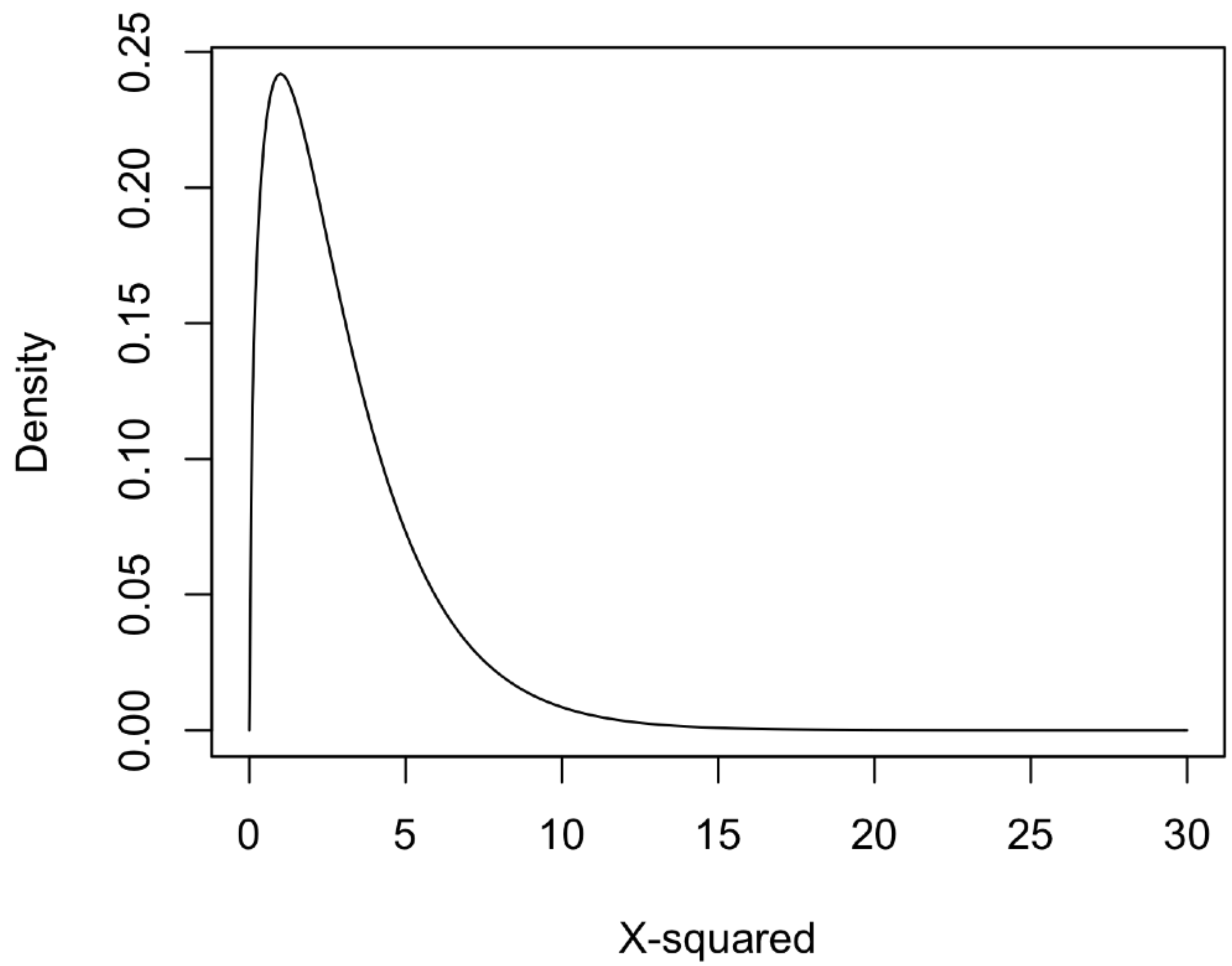
More precisely

- We can manually demonstrate this in R just to satisfy ourselves that I'm not making stuff up



Back to our hypothesis test

If the null hypothesis is true, then the sampling distribution for our X^2 statistic is a chi-square distribution with 3 (i.e., $k-1$) degrees of freedom



(That is, if the null hypothesis is true, these are the X^2 statistics we'd expect to see over many repeated experiments)

Back to our hypothesis test

We need a few things…

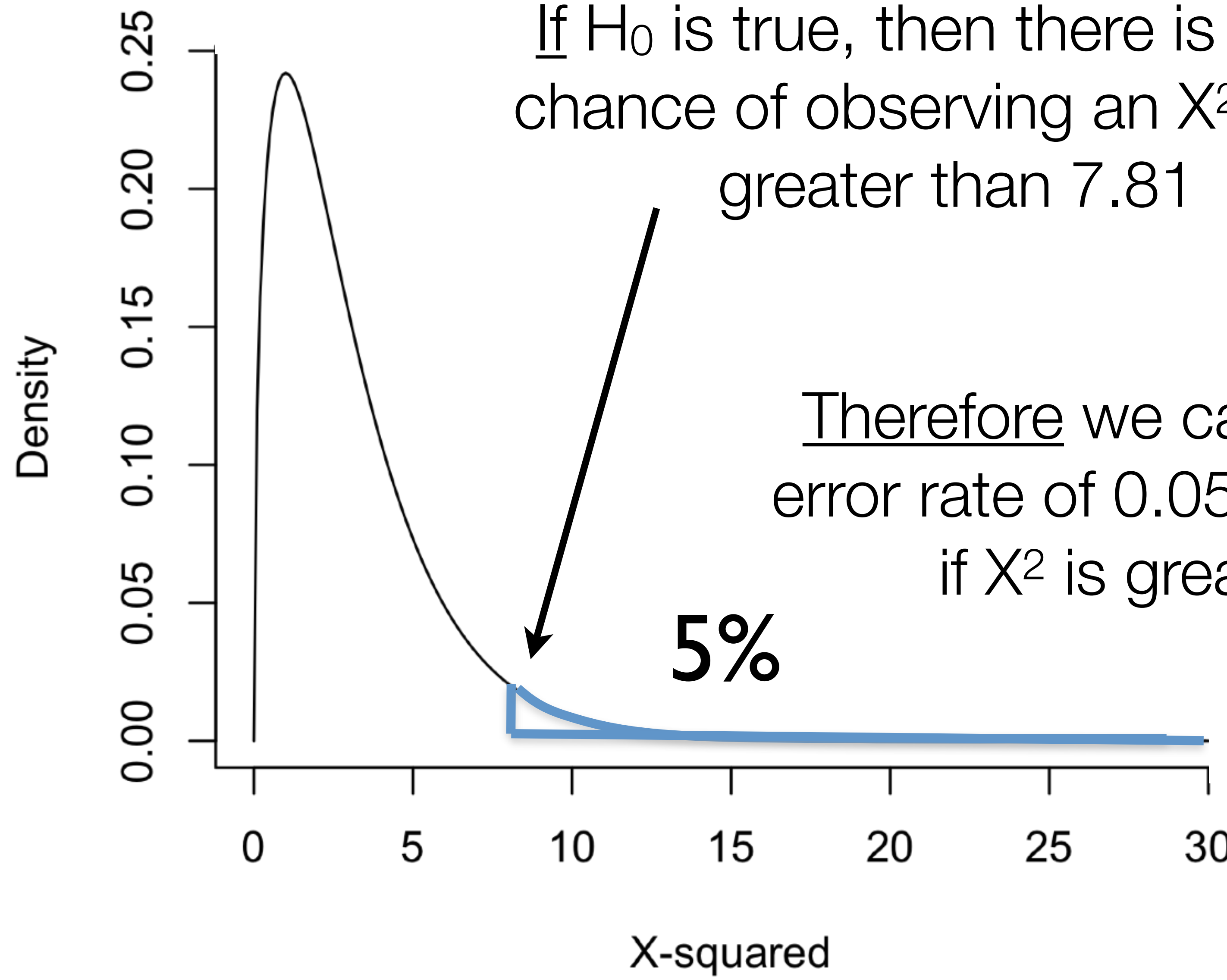
X^2	✔	1) A diagnostic test statistic, *T*
χ^2, df=3	✔	2) Sampling distribution of *T* if the null is true
11.303	✔	3) The observed *T* in your data
		4) A rule that maps every value of *T* onto a decision (accept or reject H0)

A rule that maps onto a decision

- We know that large values of X^2 imply that the null hypothesis is doing a bad job of explaining the data.

- So we will reject the null hypothesis if X^2 is <u>bigger</u> than some critical value...

The rejection region (critical region)



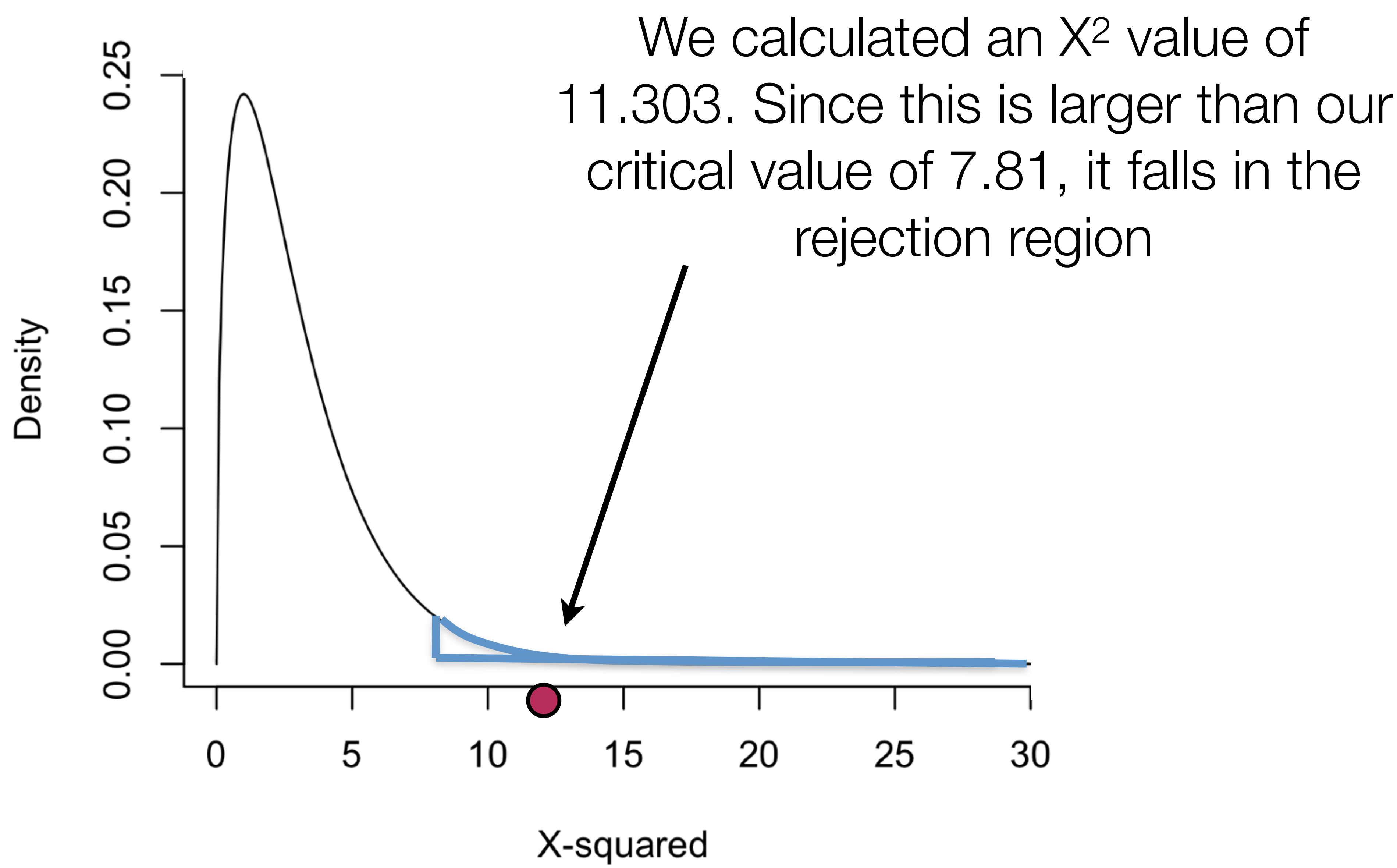
If H_0 is true, then there is a 5% chance of observing an X^2 value greater than 7.81

Therefore we can ensure a Type 1 error rate of 0.05 if we reject H_0 only if X^2 is greater than 7.81

Finds the **95%** **quantile** of the **chi-squared** distribution with **3** degrees of freedom

```
> qchisq(.95,df = 3)
[1] 7.814728
```

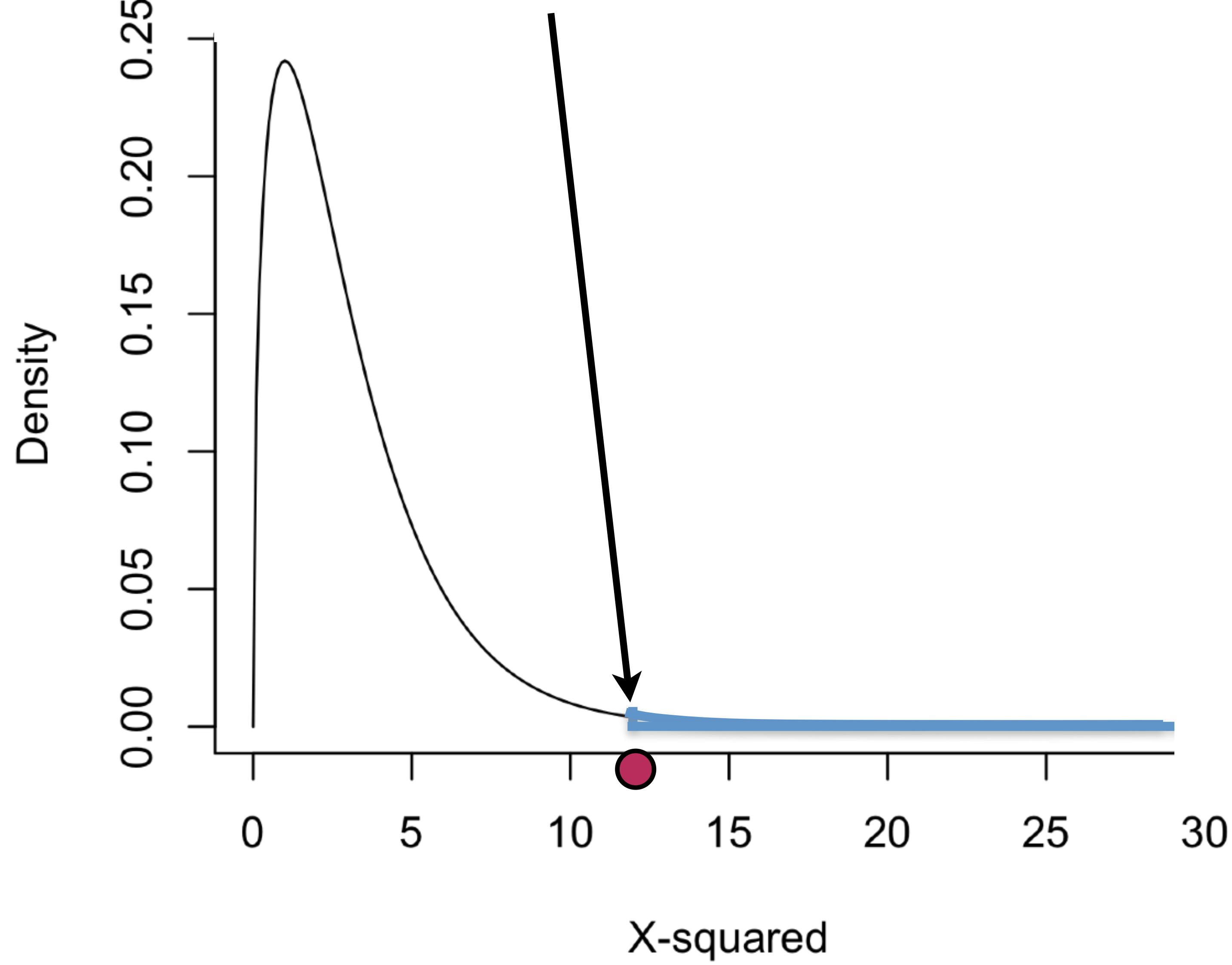
Reject the null



Therefore, for a significance level of 0.05, we reject the null hypothesis. (i.e., *p* < *.05*)

Can we calculate the exact p-value?

This is smallest possible rejection region that still includes our data set!

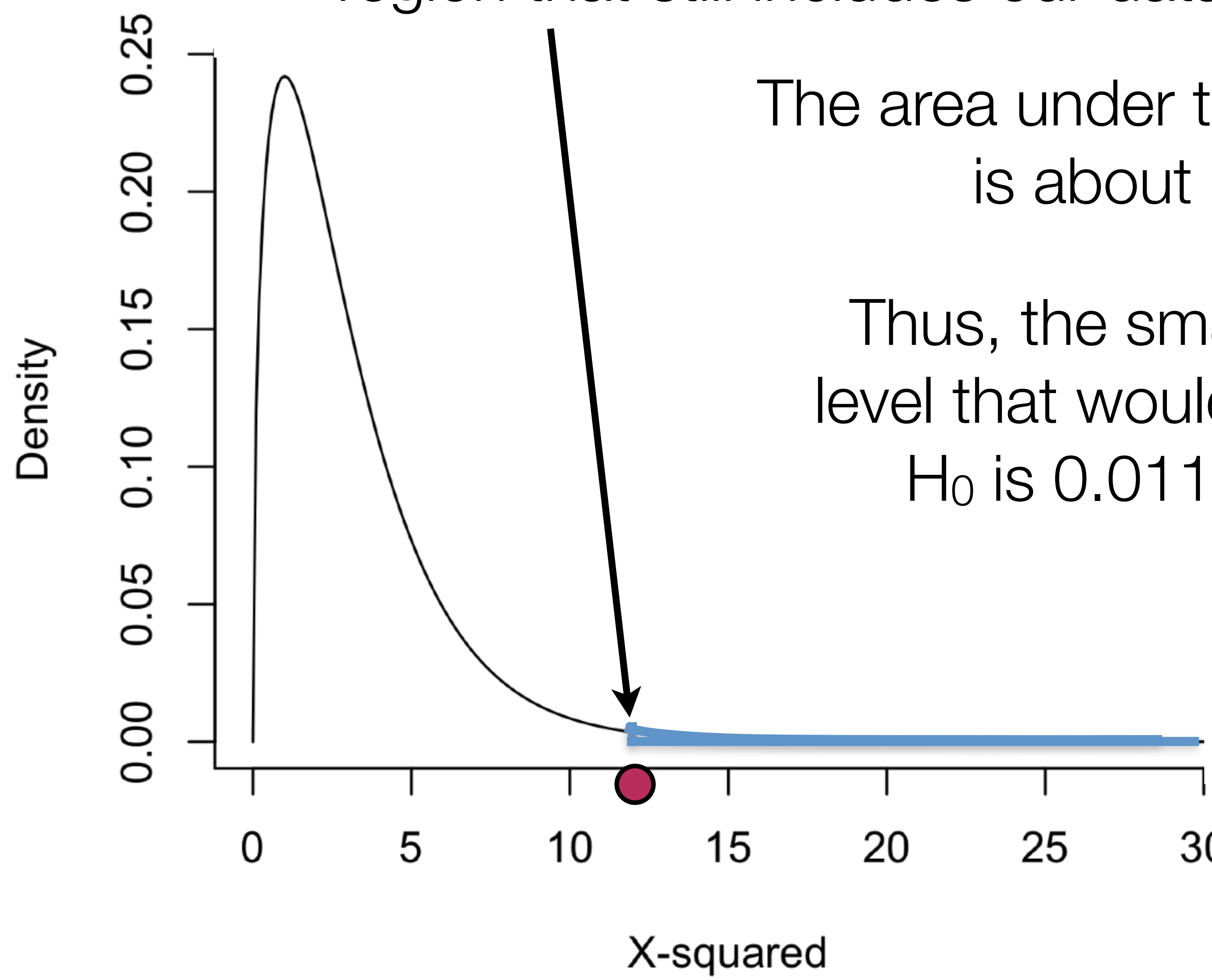


Finds the **probability** of obtaining a statistic of **11.303 or less** given a **chi-squared** distribution with **3** degrees of freedom

```
> pchisq(11.303,df = 3)
[1] 0.9898046
```

Can we calculate the exact p-value?

This is smallest possible rejection region that still includes our data set!



The area under the curve here is about 0.011

Thus, the smallest significance level that would allow us to reject H_0 is 0.011 (i.e. $p = 0.011$)

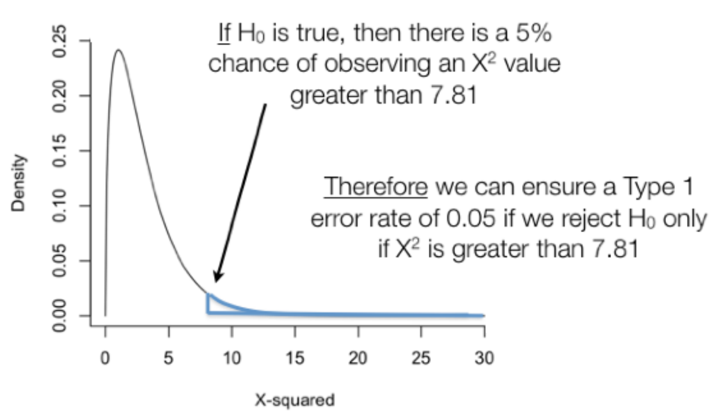
Finds the **probability** of obtaining a statistic of **11.303 or more** given a **chi-squared** distribution with **3** degrees of freedom

```
> 1-pchisq(11.303,df = 3)
[1] 0.01019535
```

Recap

Chi-square goodness of fit test is used for categorical data when you want to compare observed frequencies against some hypothesis about the true probabilities.

1) A diagnostic test statistic, *T*
   goodness of fit (X^2): larger value = more evidence against the null
2) Sampling distribution of *T* if the null is true
   χ^2 distribution. degrees of freedom=k-1, where k=number of categories
3) The observed *T* in your data
   11.303 in our example
4) A rule that maps every value of *T* onto a decision (accept or reject H0)



How do we calculate this in R?

Remember our data…

```
> ed
 bunny doggie gladly shadow 
 0.125  0.455  0.334  0.086 

> votingTable <- table(d$vote)
> votingTable

 bunny doggie gladly shadow 
     2     55     36      7 
```

How do we calculate this in R?

- The key arguments:
  - x - specifies the observed frequencies
  - p - specifies the probabilities for the null hypothesis

```
> chisq.test(x=votingTable,p=ed)

    Chi-squared test for given probabilities

data:  votingTable
X-squared = 11.304, df = 3, p-value = 0.01019
```

Understanding the output

What data is being analysed?

What kind of test did we run?

```
Chi-squared test for given probabilities

data:  votingTable
X-squared = 11.304, df = 3, p-value = 0.01019
```

The test statistic

The p value

The degrees of freedom for the test

How to write up the results

General formula

1) Report the relevant descriptive statistics

2) Specify the null hypothesis and the statistical test run

3) Give the result of the test

4) Where possible, interpret the results in terms of your research hypothesis.

Use this for all statistical tests! Even if you haven't been given a specific template, you can't go wrong with this

An example using the self report data

Pretty good

Of the 100 people in our sample, 36 voted to rescue LFB (Gladly's option), 55 voted to attack the Others (Doggie's option), 7 voted to analyse things further (Shadow's option), and 2 voted to leave (Bunny's option). When compared to the voting rates to each person in a previous election (33.4%, 45.5%, 8.6% and 12.5% respectively), using a chi-squared goodness of fit test, we found significant deviations, $\chi^2 = 11.30$, $df = 3$, $p = .0102$. This suggests that the votes this time did not simply reflect the popularity of each person.

An example using the self report data

Better

Table 1 compares the votes for each option (3rd column) to the votes in a previous election for each person (4th column). Using a chi-squared goodness of fit test, we found significant deviations, $\chi^2 = 11.30$, $df = 3$, $p = .0102$. This suggests that the votes this time did not simply reflect the popularity of each person.

Person	Option endorsed	Option votes	Election votes
Bunny	leave	2	12.5%
Doggie	attack	55	45.5%
Gladly	rescue	36	33.4%
Shadow	analyse	7	8.6%

An example using the self report data

1) Report the relevant descriptive statistics

Of the 100 people in our sample, 36 voted to rescue LFB (Gladly's option), 55 voted to attack the Others (Doggie's option), 7 voted to analyse things further (Shadow's option), and 2 voted to leave (Bunny's option). When compared to the voting rates to each person in a previous election (33.4%, 45.5%, 8.6% and 12.5% respectively), using a chi-squared goodness of fit test, we found significant deviations, $\chi^2 = 11.30$, $df = 3$, $p = .0102$. This suggests that the votes this time did not simply reflect the popularity of each person.

An example using the self report data

1) Report the relevant descriptive statistics

Table 1 compares the votes for each option (3rd column) to the votes in a previous election for each person (4th column). Using a chi-squared goodness of fit test, we found significant deviations, $\chi^2 = 11.30$, $df = 3$, $p = .0102$. This suggests that the votes this time did not simply reflect the popularity of each person.

Person	**Option endorsed**	**Option votes**	**Election votes**
Bunny	leave	2	12.5%
Doggie	attack	55	45.5%
Gladly	rescue	36	33.4%
Shadow	analyse	7	8.6%

An example using the self report data

1) Report the relevant descriptive statistics

Table 1 compares the votes for each option (3rd column) to the votes in a previous election for each person (4th column). Using a chi-squared goodness of fit test, we found significant deviations, $\chi^2 = 11.30$, $df = 3$, $p = .0102$. This suggests that the votes this time did not simply reflect the popularity of each person.

2) Specify the null hypothesis and the statistical test run

An example using the self report data

1) Report the relevant descriptive statistics

Table 1 compares the votes for each option (3rd column) to the votes in a previous election for each person (4th column). Using a chi-squared goodness of fit test, we found significant deviations, $\chi^2 = 11.30$, $df = 3$, $p = .0102$. This suggests that the votes this time did not simply reflect the popularity of each person.

2) Specify the null hypothesis and the statistical test run

3) Give the result of the test

An example using the self report data

1) Report the relevant descriptive statistics

Table 1 compares the votes for each option (3rd column) to the votes in a previous election for each person (4th column). Using a chi-squared goodness of fit test, we found significant deviations, $\chi^2 = 11.30$, $df = 3$, $p = .0102$. This suggests that the votes this time did not simply reflect the popularity of each person.

2) Specify the null hypothesis and the statistical test run

3) Give the result of the test

4) Where possible, interpret the results in terms of your research hypothesis.

The "stat reference”, version 1

$$\chi^2 = 11.30,\ df = 3,\ p = .0102$$

the p-value

the test statistic has a chi-square distribution under H_0

the value of the test statistic

the degrees of freedom for the relevant sampling distribution

A more compact (and more common) version

$$\chi^2(3) = 11.30,\ p = .0102$$

the test statistic has a chi-square distribution under H_0

the degrees of freedom for the relevant sampling distribution

the value of the test statistic

the p-value

Exercises are in w6day1exercises.Rmd