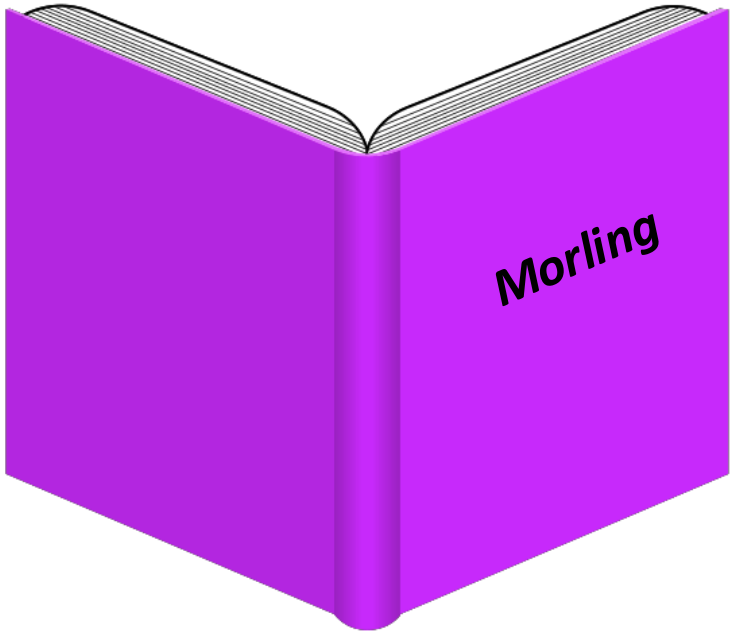
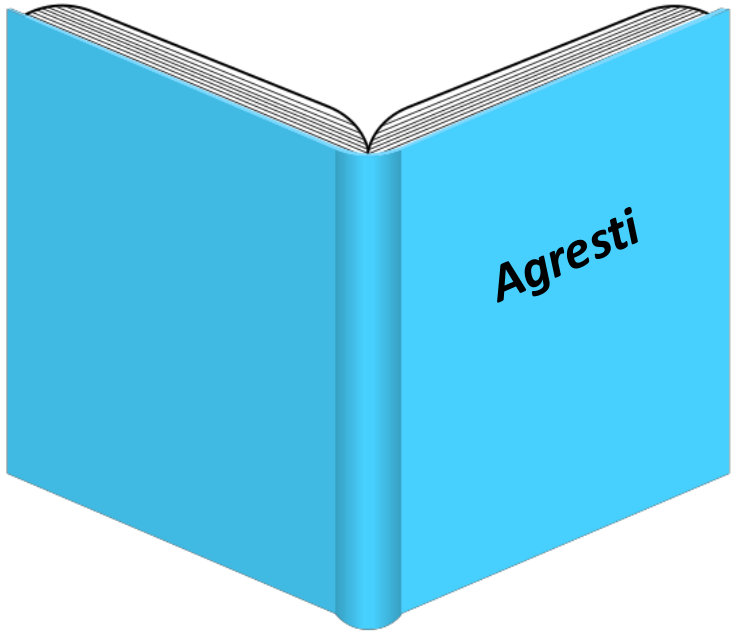




Research Methods and Statistics

Lecture 18: Association between categorical variables

Johnny van Doorn

Pictures source: pixabay

Exam note on comparing means

To simplify the procedure, we use the following rule at the exam:

- Comparing proportions: use z-distribution
- Comparing means: use t-distribution
- Independent samples: df = n1 + n2 – 2
- Dependent samples: df = n - 1

See also the flowchart

Association between two categorical variables

Source: One flew over the Cuckoo's nest (1975)

Overview of Today

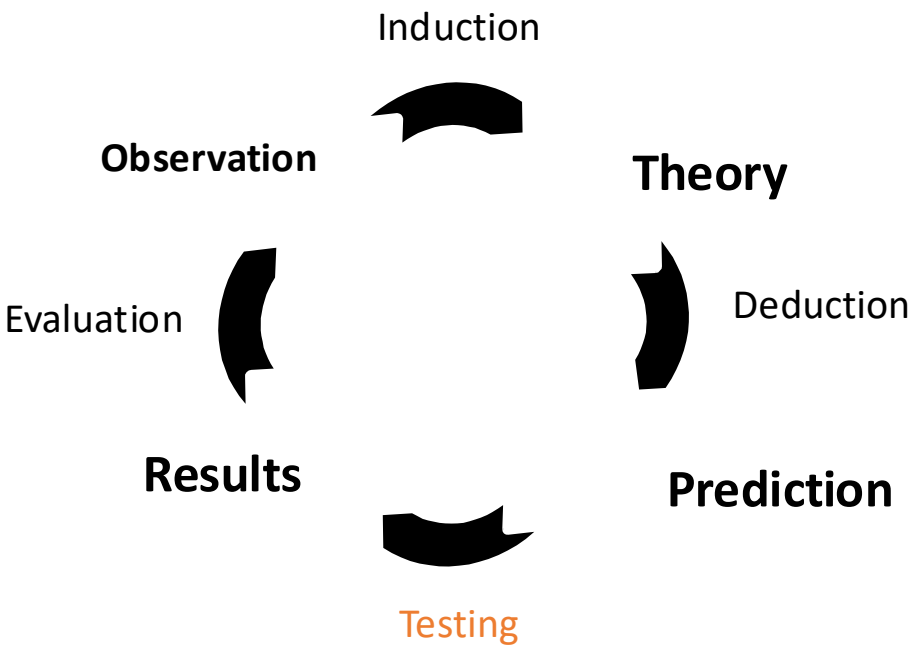
1. **Recap**
   - Association between two categorical variables
   - Conditional proportions
   - Association and (in)dependence

2. Hypothesis test for the association between two categorical variables
   - Strength of association
   - z-test

2. Recap
   - Next time
   - Example exam question

Role of variables in empirical cycle

Operationalize: determine how you will measure the conceptual variables from the prediction



Role of variables in empirical cycle

Explanatory variable

Response variabele	**Quantitative**	**Categorical**
Quantitative	Correlation Regression	t-test ANOVA
Categorical	Logistic regression	Contingency table

Association between two categorical variables

- Research question: Do people that are treated with electroshock therapy recover from a psychotic disorder?

- Both the response and explanatory variables are categorical (yes/no)
  - Contingency table!

Contingency Table

Counts

Treatment?	Recovered? No	Recovered? Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

Proportions

Treatment?	Recovered? No	Recovered? Yes	Total
No	0.3625	0.1375	0.5
Yes	0.1625	0.3375	0.5
Total	0.525	0.475	**80(= n)**

(e.g., 29 / 80 = 0.3625)

Conditional Proportions

Conditional proportion: The proportion of the response variable, for *one level* of the explanatory variable

Recovered?

Treatment?	No	Yes	Total
No	29	11	40
Yes	**13**	**27**	**40**
Total	42	38	**80 (= n)**

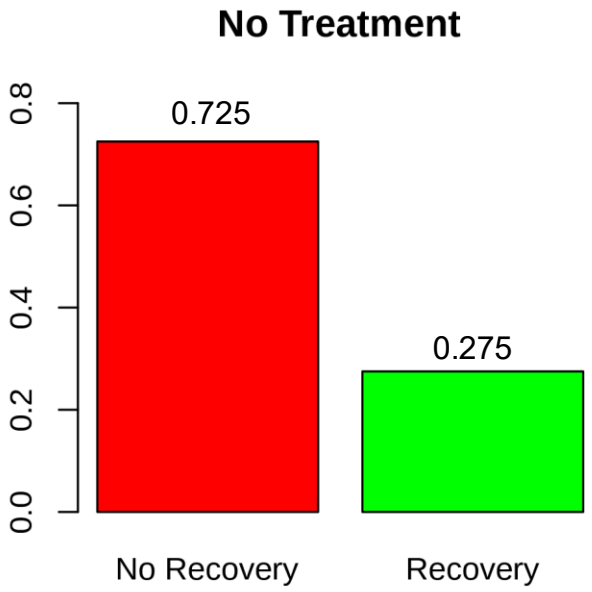
Recovered?

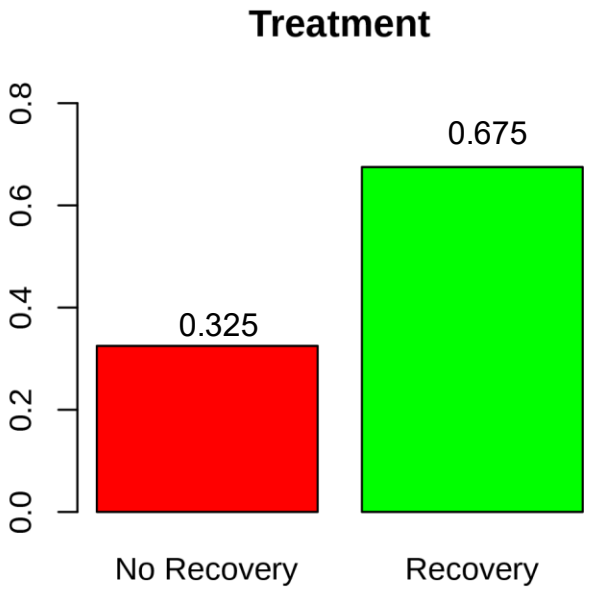
Treatment?	No	Yes	Total
No	0.725	0.275	1
Yes	0.325	0.675	1
Total			**80(= n)**

(e.g., 13 / 40 = 0.325)

Conditional Proportions

Conditional proportion: The proportion of the response variable, for *one level* of the explanatory variable





Recovered?

Treatment?	No	Yes	Total
No	0.725	0.275	1
Yes	0.325	0.675	1
Total			**80(= n)**

What can we now express?

- Electroshock therapy seems to affect the chance of recovery
  - Treatment and recovery are *associated*

(In)dependence

Independence: A response variable is **independent** of the explanatory variable, when the *conditional* proportions are the same for all categories of the explanatory variable
i.e. there is **no association**

Dependence: A response variable is **dependent** of the explanatory variable, when the *conditional* proportions differ for one category of the explanatory variable
i.e. there is an **association**

What can we now express?

- Electroshock therapy seems to affect the chance of recovery
  - Treatment and recovery are *associated*
  - In this sample → Descriptive statistics
  - In week 9 we will discuss inferential statistics


NEVER GIVE UP
JOHN CENA
THE TIME IS NOW
OFFICIAL THEME
HLR

What about the inference to the population?

We can test the null hypothesis that IV and DV are independent

Overview of Today

1. Recap
   - Association between two categorical variables
   - Conditional proportions
   - Associations and (in)dependence

2. **Hypothesis test for the association between two categorical variables**
   - Strength of association
   - z-test

2. Recap
   - Next time
   - Example exam question

The five steps of a significance test for categorical variables

Step 0: specify your alpha level!

1. **Assumptions**
2. Hypothesis
3. Test statistic
4. P-value
5. Conclusion

1. Assumptions

- Both response and explanatory variable (DV and IV) are categorical
- Random sample
- A large sample:
  - At least 5 counts per cell *are expected*
  - I will come back to “expected” later

Recovered?

Treatment?	No	Yes	Total
No	55	8	63
Yes	9	8	17
Total	64	16	**80(= n)**

Recovered?

Treatment?	No	Yes	Total
No	55	8	63
Yes	15	2	17
Total	70	10	**80(= n)**

The five steps of a significance test for categorical variables

1. Assumptions
2. **Hypothesis**
3. Test statistic
4. P-value
5. Conclusion

2. Hypothesis

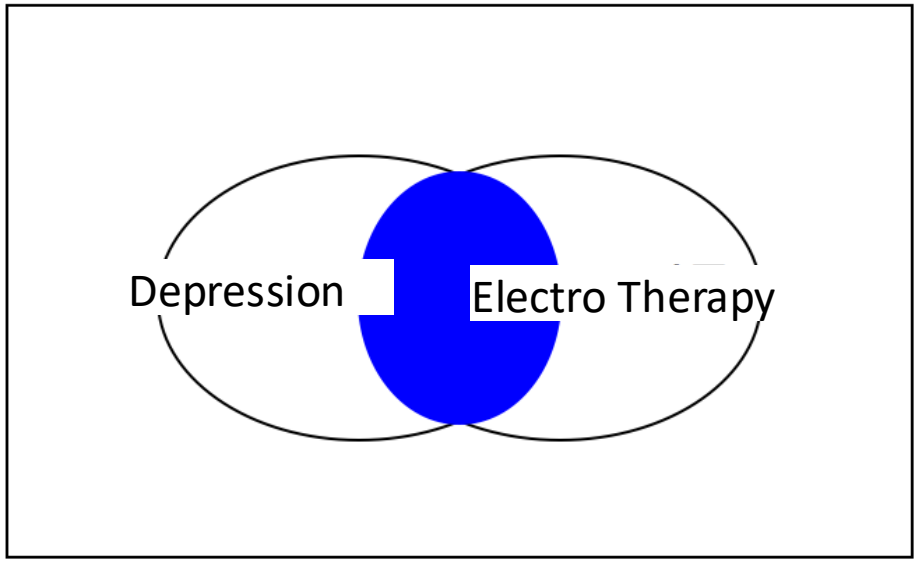
Hypothesis : “Electroshock therapy is associated with Psychosis Recovery”

- H_0: the variables are *independent*
- H_A: the variables are *dependent*

- What does H_0 imply?

- P(A and B) = P(A) * P(B)

Multiplication rule for ***independent*** events



2. Hypothesis

Counts

Treatment? \ Recovered?	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

Proportions

Treatment? \ Recovered?	No	Yes	Total
No	0.3625	0.1375	0.5
Yes	0.1625	0.3375	0.5
Total	0.525	0.475	**80(= n)**

(e.g., 29 / 80 = 0.3625)

2. Hypothesis

Recovered?

Treatment?	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

2. Hypothesis

Proportions:

Recovered?

Treatment?	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

Recovered?

Treatment?	No	Yes	Total
No			0.5
Yes			0.5
Total	0.525	0.475	**1**

(e.g., 42/80=0.525)

2. Hypothesis

Proportions:

Expected by H_0:

Treatment?

Recovered?

	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

Recovered?

	No	Yes	Total
No			0.5
Yes			0.5
Total	0.525	0.475	**1**

Recovered?

	No	Yes	Total
No			0.5
Yes			0.5
Total	0.525	0.475	**1**

- If H_0 is true, then P(A and B) = P(A) * P(B)
- P(Treatment AND Recovery) = P(Treatment) * P(Recovery)
- P(Treatment AND Recovery) = 0.5 * 0.475 = 0.2375

These are the probabilities of all possible outcomes ***if*** you assume independence (ie H_0)

2. Hypothesis

Proportions:

Expected by H_0:

Treatment? \ Recovered?	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

Treatment? \ Recovered?	No	Yes	Total
No			0.5
Yes			0.5
Total	0.525	0.475	**1**

Treatment? \ Recovered?	No	Yes	Total
No			0.5
Yes		0.2375	0.5
Total	0.525	0.475	**1**

- If H_0 is true, then P(A and B) = P(A) * P(B)
- P(Treatment AND Recovery) = P(Treatment) * P(Recovery)
- P(Treatment AND Recovery) = 0.5 * 0.475 = 0.2375

- P(Treatment AND Not Recovery) = 0.5 * 0.525 = 0.2625
- …

These are the probabilities of all possible outcomes ***if*** you assume independence (ie H_0)

2. Hypothesis

Proportions:

Expected by H_0:

Treatment?

Recovered?

	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

Recovered?

	No	Yes	Total
No			0.5
Yes			0.5
Total	0.525	0.475	**1**

Recovered?

	No	Yes	Total
No			0.5
Yes	0.2625	0.2375	0.5
Total	0.525	0.475	**1**

- If H_0 is true, then P(A and B) = P(A) * P(B)
- P(Treatment AND Recovery) = P(Treatment) * P(Recovery)
- P(Treatment AND Recovery) = 0.5 * 0.475 = 0.2375

- P(Treatment AND Not Recovery) = 0.5 * 0.525 = 0.2625
- …

These are the probabilities of all possible outcomes ***if*** you assume independence (ie H_0)

2. Hypothesis

Proportions:

Expected by H_0:

Treatment?

Recovered?

	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80(= n)**

Recovered?

	No	Yes	Total
No			0.5
Yes			0.5
Total	0.525	0.475	**1**

Recovered?

	No	Yes	Total
No	0.2625	0.2375	0.5
Yes	0.2625	0.2375	0.5
Total	0.525	0.475	**1**

- If H_0 is true, then P(A and B) = P(A) * P(B)
- P(Treatment AND Recovery) = P(Treatment) * P(Recovery)
- P(Treatment AND Recovery) = 0.5 * 0.475 = 0.2375

- P(Treatment AND Not Recovery) = 0.5 * 0.525 = 0.2625
- …

These are the probabilities of all possible outcomes ***if*** you assume independence (ie H_0)

The five steps of a significance test for categorical variables

1. Assumptions
2. Hypothesis
3. **Test statistic**
4. P-value
5. Conclusion

3. Test statistic χ^2

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

3. Test statistic

- H_0: the variables are independent: no association
- What are the frequencies in each cell ***under*** H_0?
- Multiply back to the *Expected contingency table*

Recovered?

Treatment?

	No	Yes	Total
No	0.2625	0.2375	0.5
Yes	0.2625	0.2375	0.5
Total	0.525	0.475	**1**

x 80 (*n*)

Recovered?

	No	Yes	Total
No	21	19	40
Yes	21	19	40
Total	42	38	**80**

Expected ***proportions*** under H_0

Expected ***frequencies*** under H_0

3. Test statistic

- H_0: the variables are independent
- What are the frequencies in each cell ***under*** H_0?
- Multiply back to the *Expected contingency table*

Recovered?

Treatment?	No	Yes	Total
No	0.2625	0.2375	0.5
Yes	0.2625	0.2375	0.5
Total	0.525	0.475	**1**

x 80 (*n*)

Recovered?

	No	Yes	Total
No	21	19	40
Yes	21	19	40
Total	42	38	**80**

Expected ***proportions*** under H_0

Expected ***frequencies*** under H_0

Recall the assumptions of a large sample: If any of these **expected** cell frequencies < 5, the X^2-test should not be done.

Expected cell frequency

$$\text{expected cell frequency} = \frac{(\text{row total}) \times (\text{column total})}{\text{sample size (n)}}$$

Recovered?

Treatment?	No	Yes	Total
No	21	19	40
Yes	21	19	40
Total	42	38	**80**

E.g.:
(40 * 42) / 80 = 21

Expected ***frequencies*** under H_0

Expected cell frequency

$$\text{expected cell frequency} = \frac{(\text{row total}) \times (\text{column total})}{\text{sample size (n)}}$$

Expected cell frequencies can be decimal numbers!! (e.g., 5.65)

3. Test statistic χ^2

Data (what is observed): O

Expected under H_0: E

Recovered?

Treatment?

	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80**

Observed frequencies

Recovered?

Treatment?

	No	Yes	Total
No	21	19	40
Yes	21	19	40
Total	42	38	**80**

Expected frequencies under H_0

3. Test statistic χ^2

Data (what is observed): O

Expected under H_0: E

Recovered?

Treatment?

	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80**

Recovered?

Treatment?

	No	Yes	Total
No	21	19	40
Yes	21	19	40
Total	42	38	**80**

Do the *observed data* deviate extremely from what is *expected under H_0*?
If yes, then reject H_0

How to quantify the difference between O and E?

3. Test statistic χ^2

χ^2 statistic ("chisquare") can be used to determined how much the observed and expected cell frequencies deviate (assuming the null hypothesis of independence).

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

3. Test statistic χ^2

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\chi^2 = \frac{(29-21)^2}{21} + \frac{(11-19)^2}{19} + \frac{(13-21)^2}{21} + \frac{(27-19)^2}{19}$$

$$\chi^2 = 3.0476 + 3.3684 + 3.0476 + 3.3684$$

$$\boxed{\chi^2 = 12.832}$$

Observed frequencies

	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80**

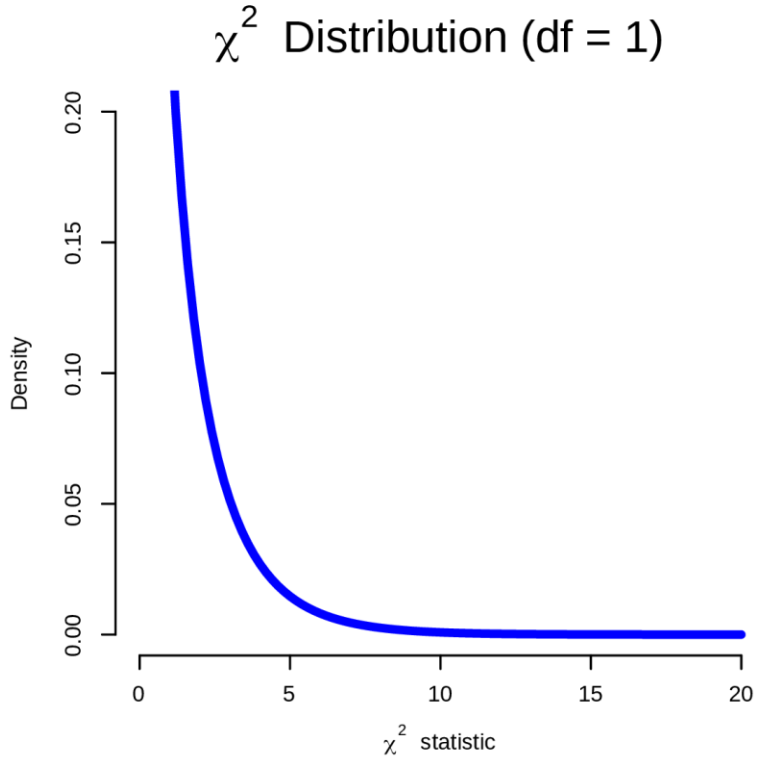
Expected frequencies under H_0

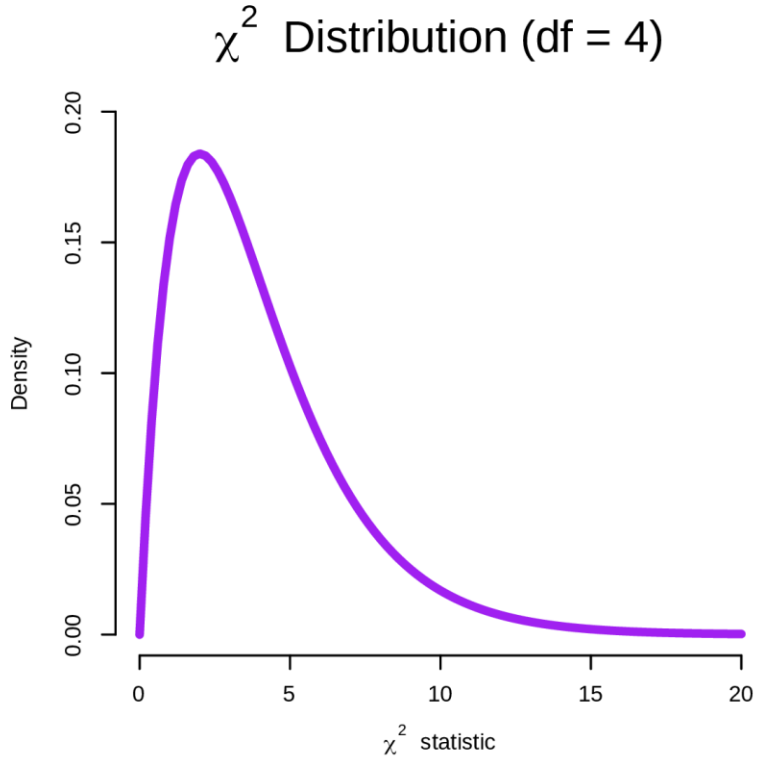
	No	Yes	Total
No	21	19	40
Yes	21	19	40
Total	42	38	**80**

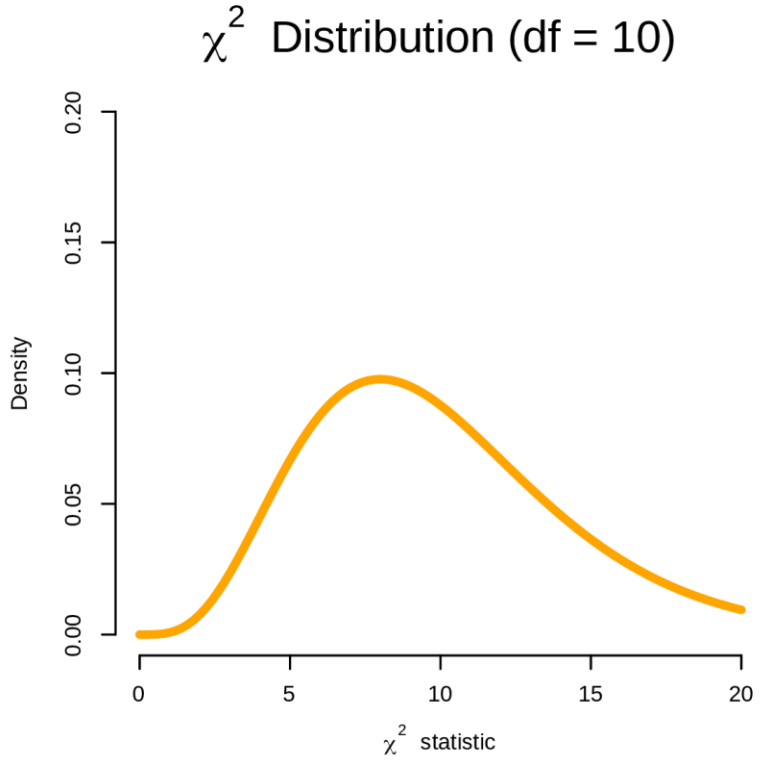
The five steps of a significance test for categorical variables

1. Assumptions
2. Hypothesis
3. Test statistic
4. **P-value**
5. Conclusion

Sampling distribution for χ^2: χ^2-distribution

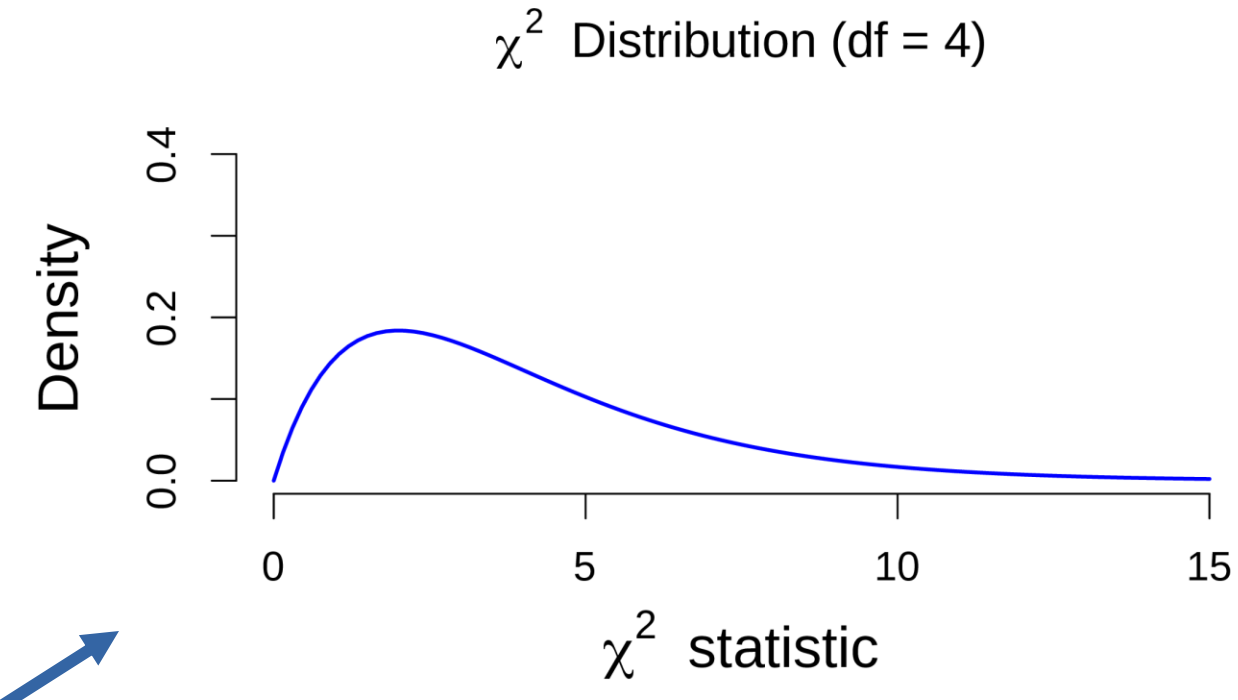






Sampling distribution for $\chi 2$

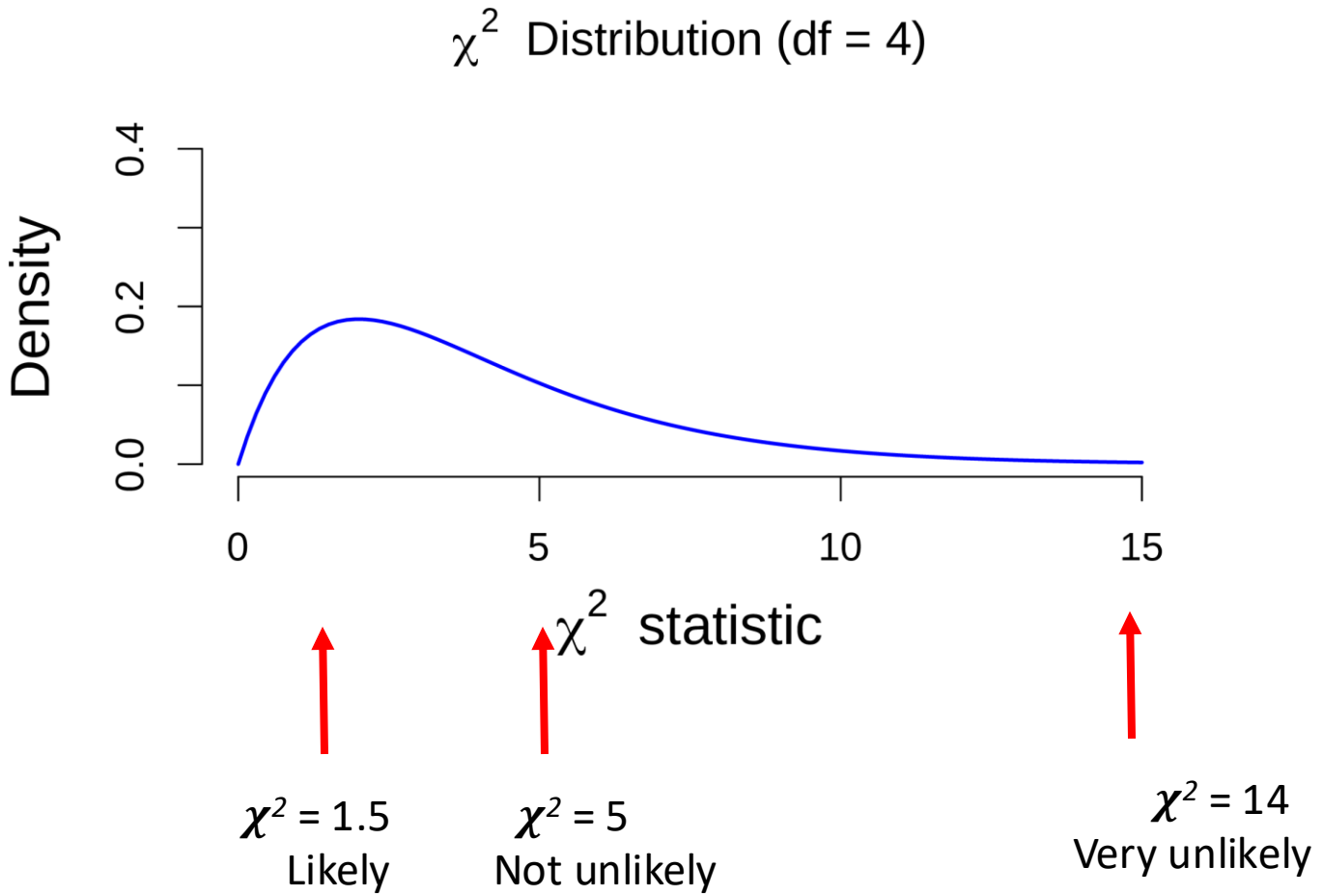
Test statistic: A measure of how far the point estimate falls from the parameter value, if the null hypothesis is true



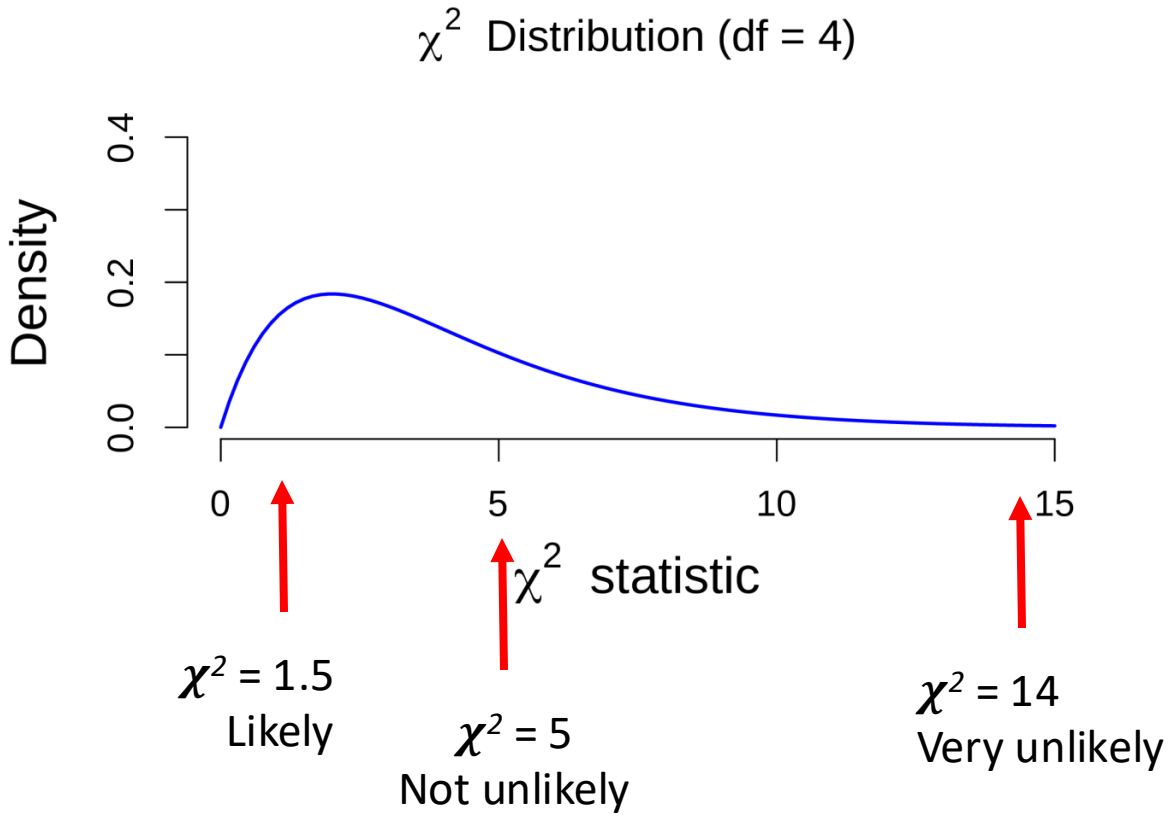
Sampling distribution of $\chi 2$, **if** the null hypothesis is true

Sampling distribution for $\chi 2$

Test statistic: A measure of how far the point estimate falls from the parameter value, if the null hypothesis is true



4. p-value

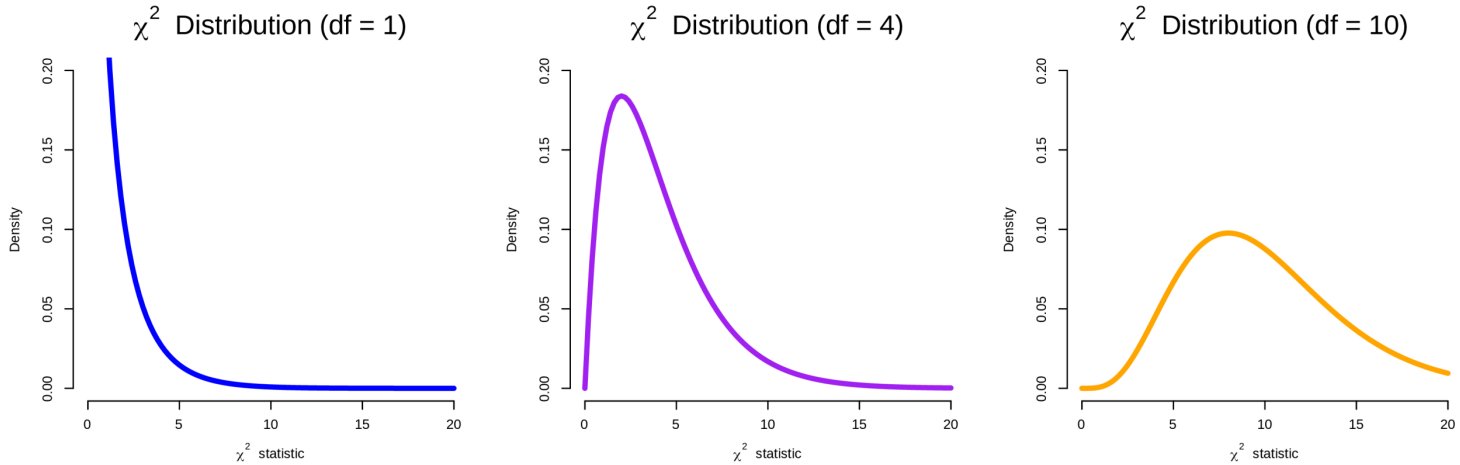


- How likely?

p-value: Probability that you observe this statistic *or more extreme*, if H_0 is true.

χ^2-distribution

- X^2 has a specific sampling distribution

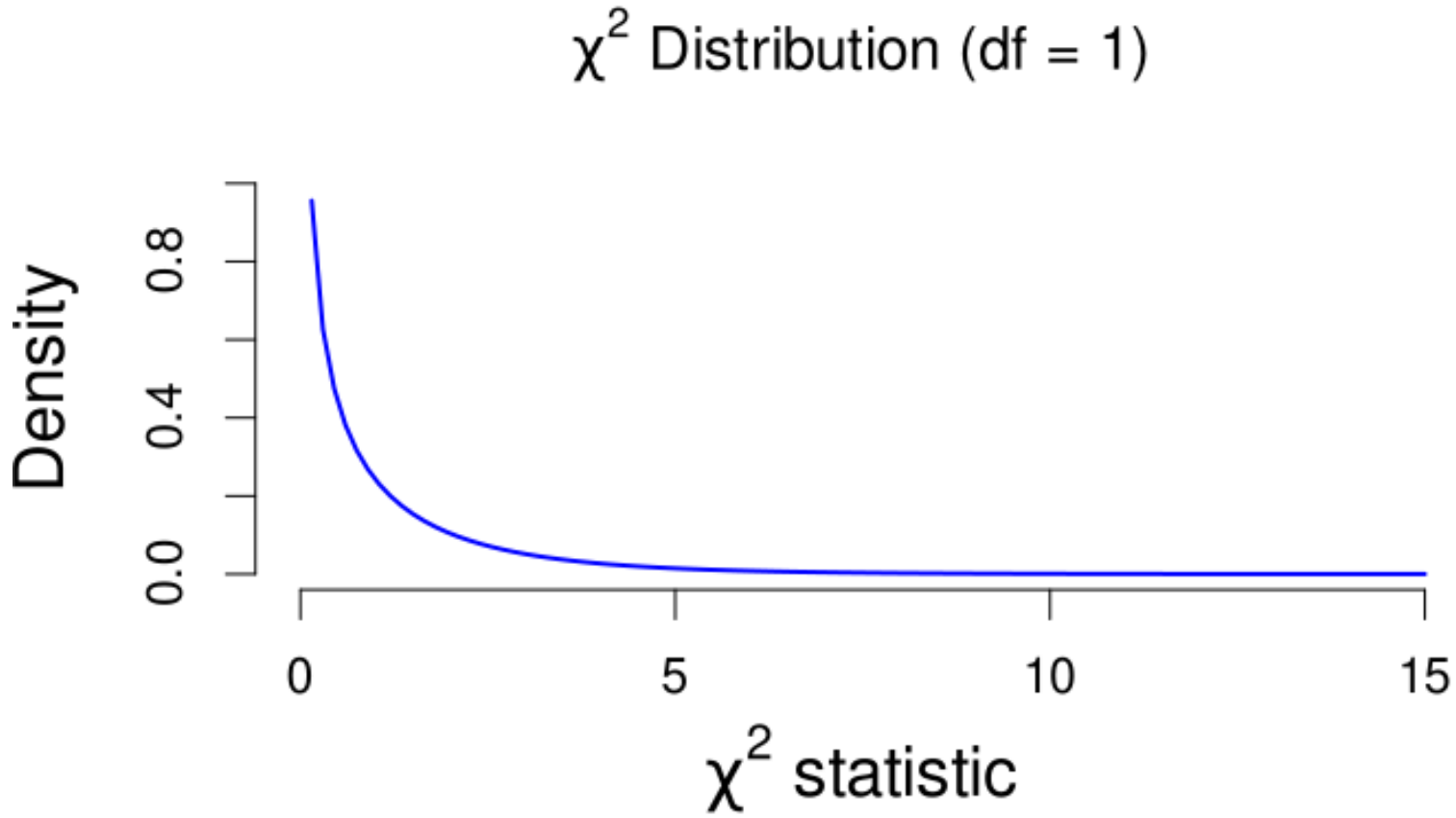


- Always positive (Because $\chi^2 = \sum \frac{(O-E)^2}{E}$)
- df = (#rows – 1) * (#columns - 1)
- Mean equals df
- Approaches bell-shape as df increases
- Large X^2 is evidence *against* independence

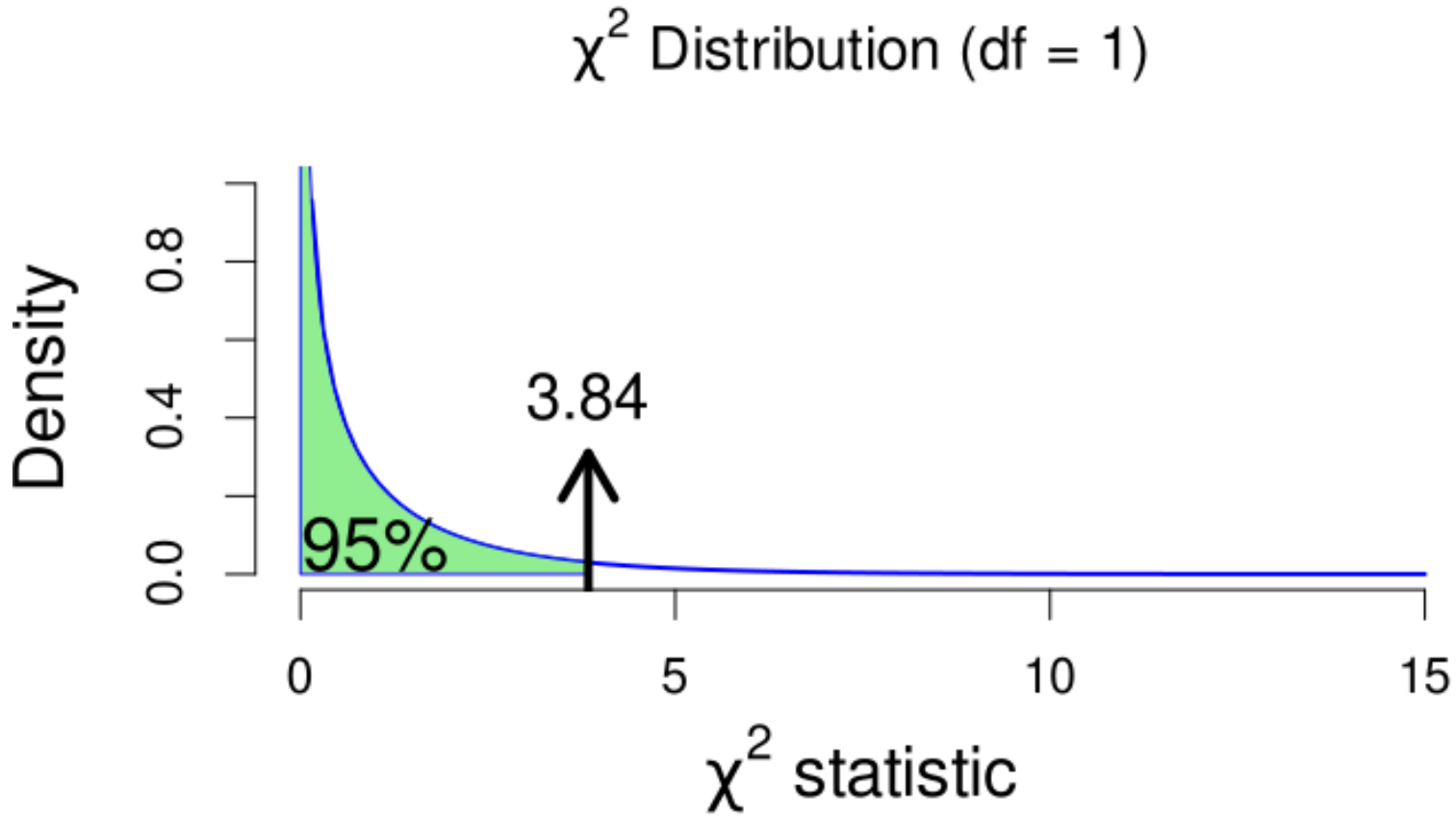
In our case, we have 2 rows, 2 columns:
df = (2 – 1) * (2 – 1) = 1

A X^2 test is always two-sided

4. p-value


χ^2 Distribution (df = 1)
Density
0.0
0.4
0.8
0
5
10
15
χ^2 statistic

4. p-value



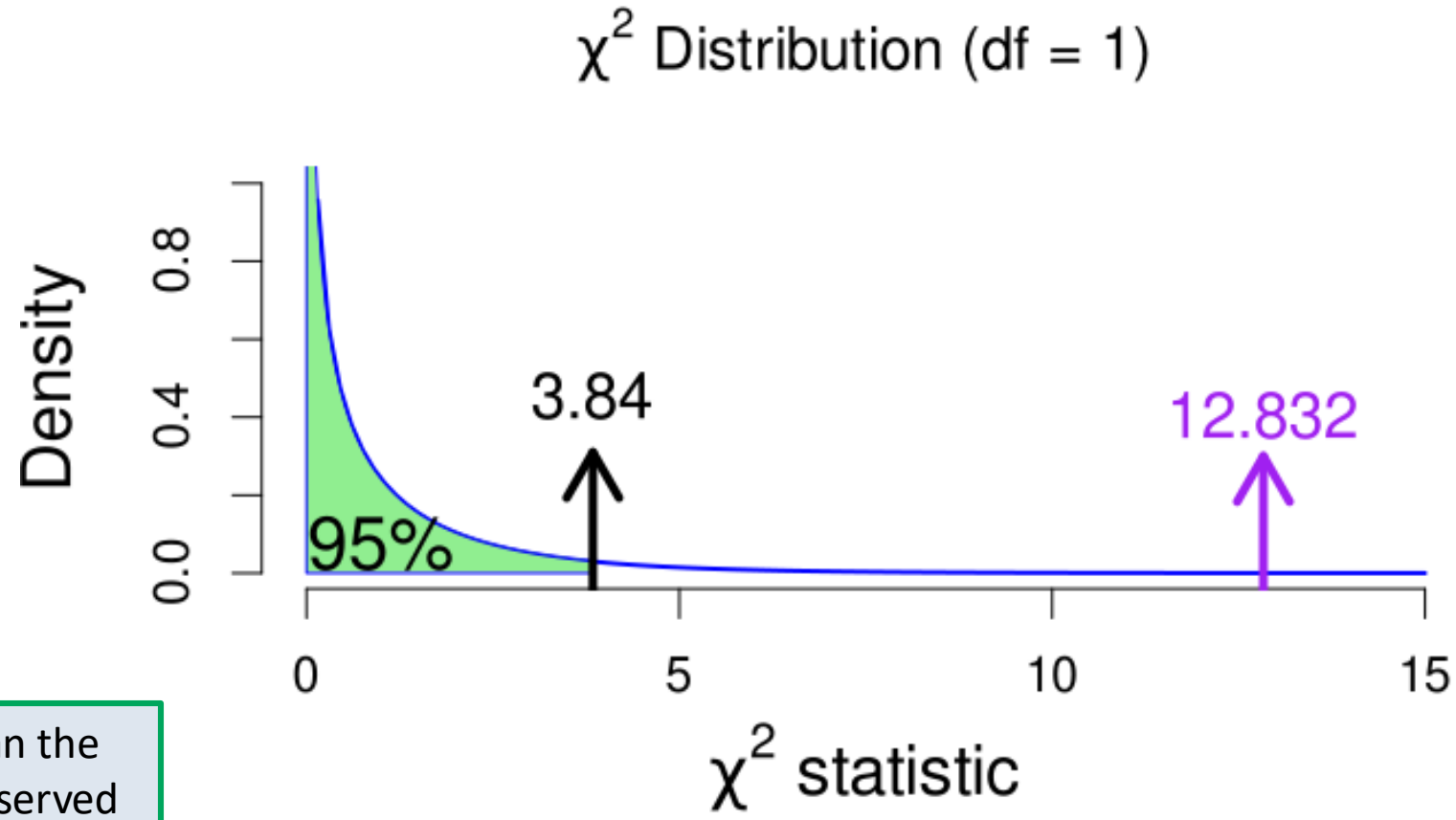
MS Excel: `= CHISQ.INV(0.95, 1)`
`= 3.84`

df

1 - alpha

= the critical X^2 value corresponding to our alpha of 0.05 and df = 1

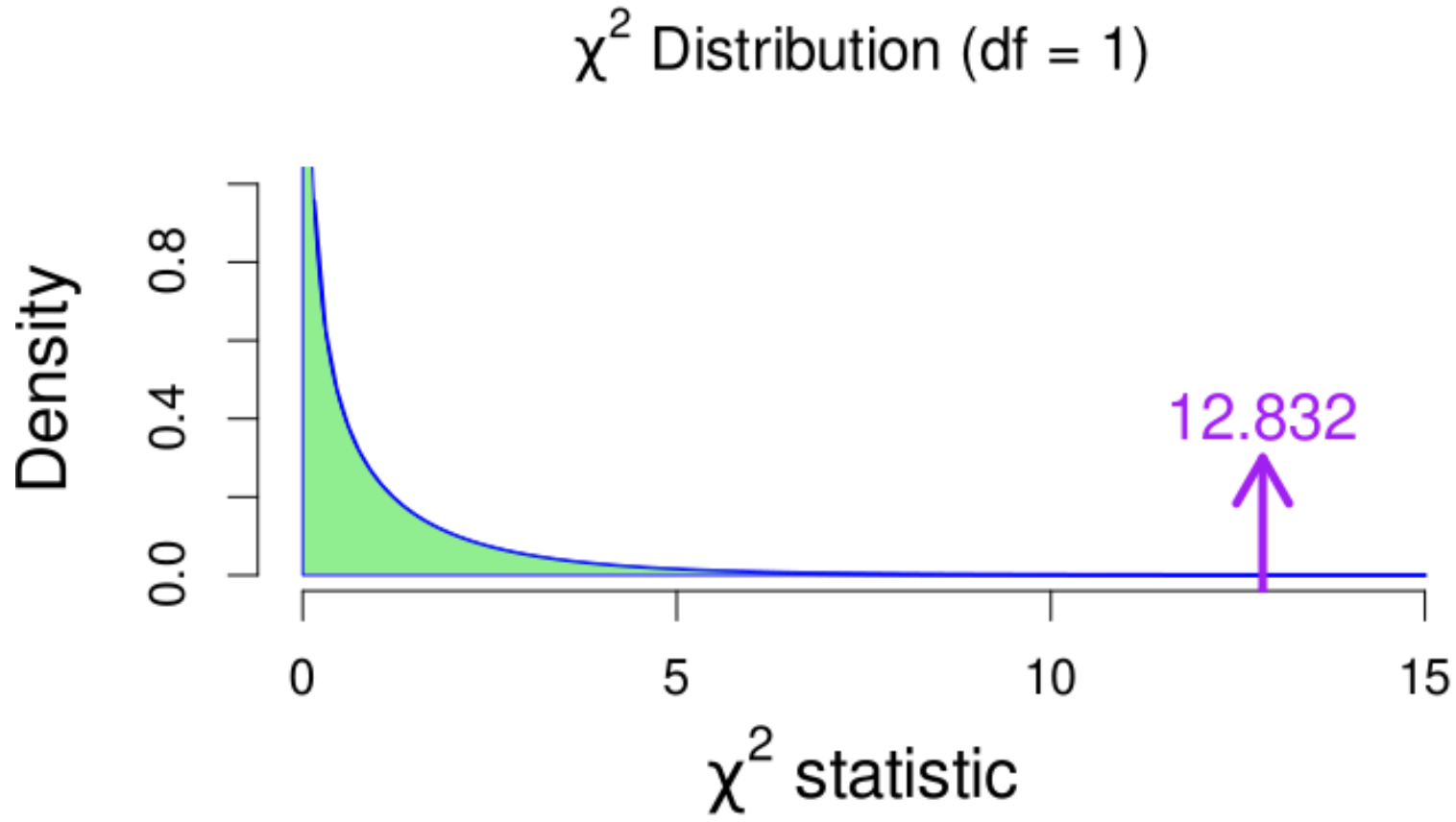
4. p-value



Observed X^2 -value is greater than the critical X^2 value of 3.84, so our observed result is significant at the 0.05 level (i.e., $p < 0.05$)

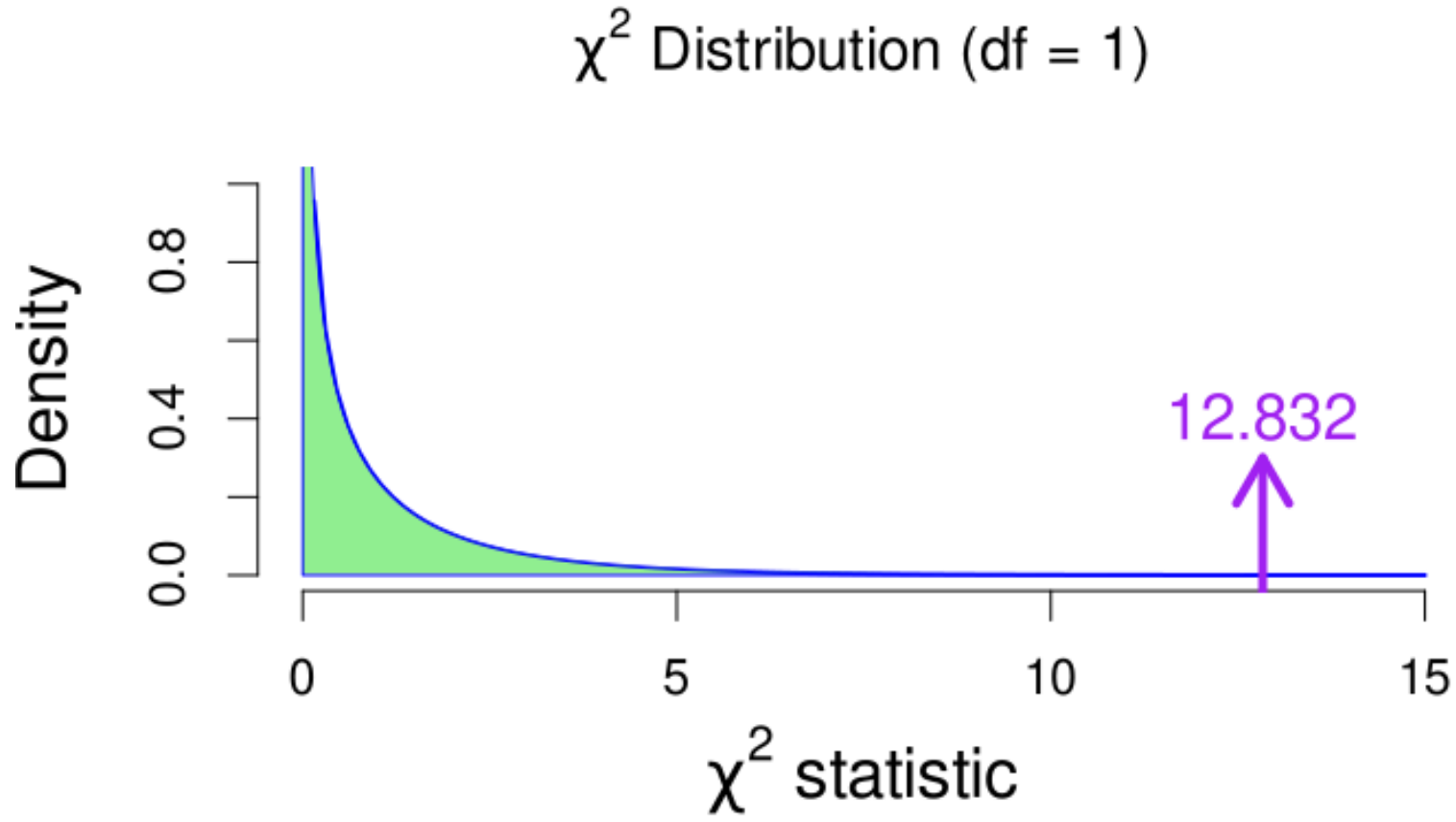
But how to get the specific p-value?

4. p-value



MS Excel: `= 1-CHISQ.DIST(12.832, 1,TRUE)`

4. p-value

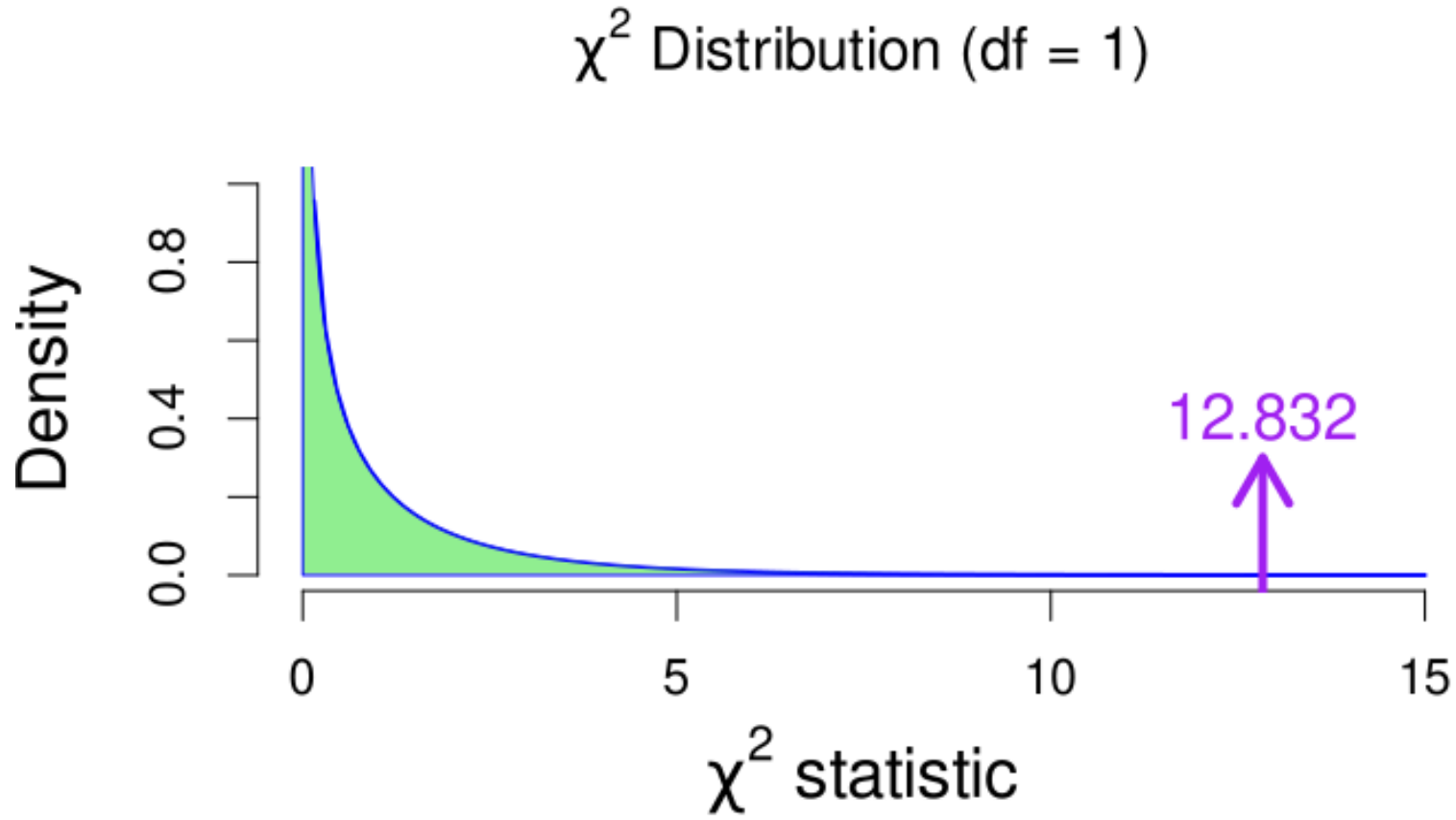


MS Excel: `= 1-CHISQ.DIST(12.832, 1,TRUE)`

Test statistic

df

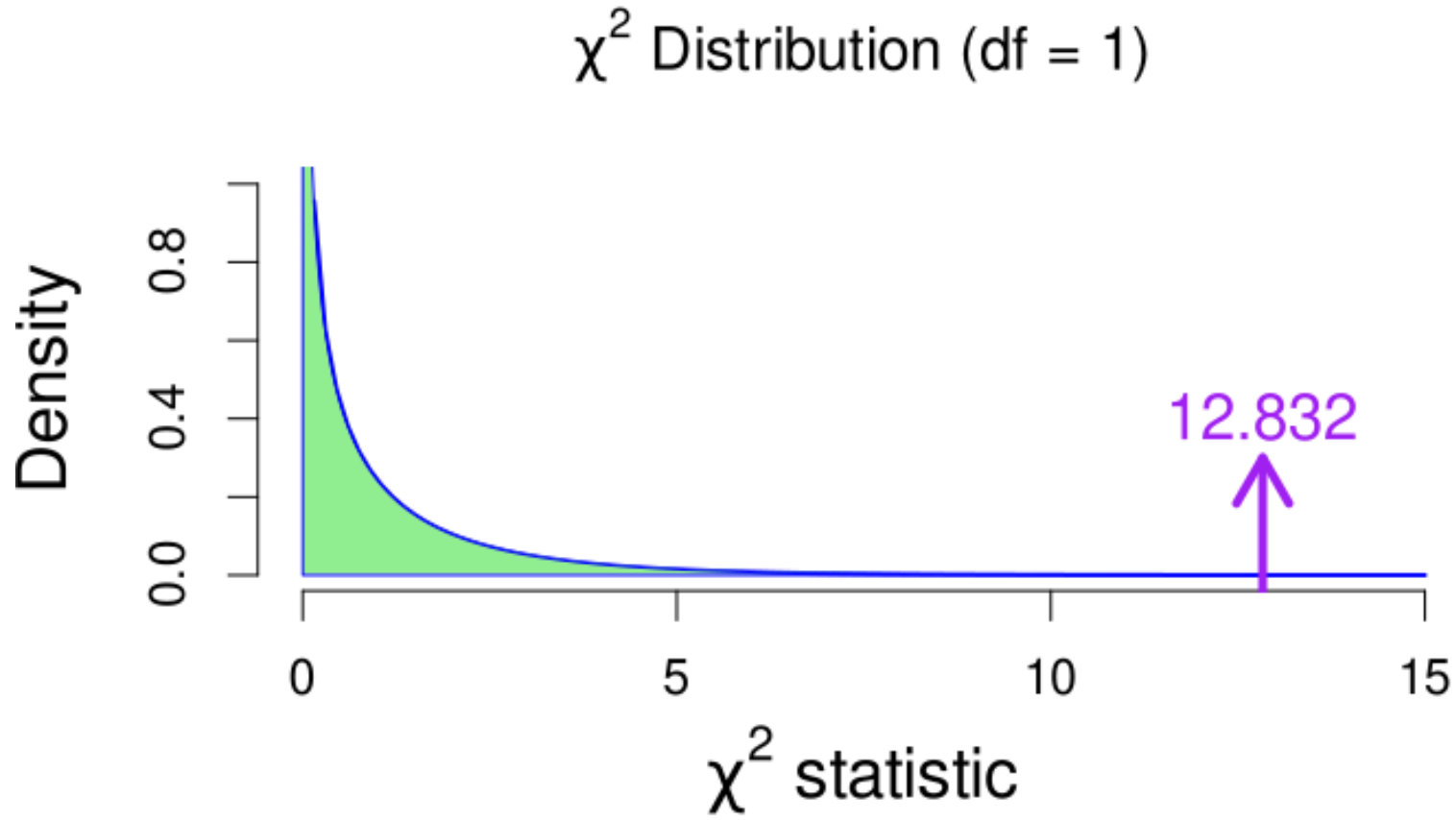
4. p-value



MS Excel: `= 1-CHISQ.DIST(12.832, 1,TRUE)`

Set to true for the area under the curve, to the *left* of the test statistic
(to get the area to the right, we preface the Excel command by "1 - ")

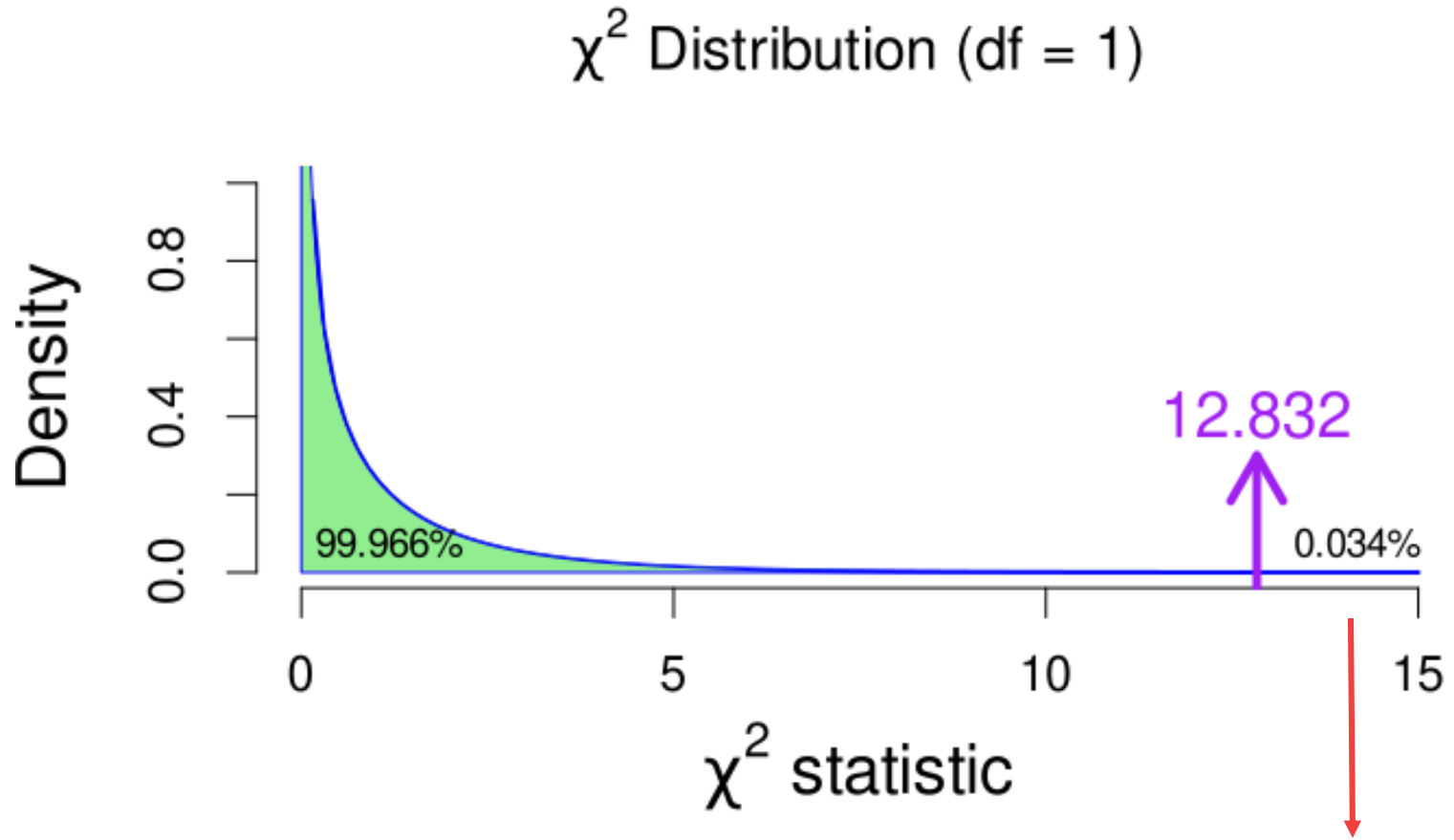
4. p-value



MS Excel: `= 1-CHISQ.DIST(12.832, 1,TRUE)`

`= 0.0003407`

4. p-value



$P(X^2 > 12.832) = 0.00034$

MS Excel: `= 1-CHISQ.DIST(12.832, 1,TRUE)`

`= 0.0003407`

p-value: Probability that you observe the observed test statistic *or more extreme*, ***if*** H_0 is true.

The five steps of a significance test for categorical variables

1. Assumptions
2. Hypothesis
3. Test statistic
4. P-value
5. **Conclusion**

5. Conclusion

- Because $p < 0.05$ we can reject H_0
  - (if that is the significance level alpha we chose beforehand)

- There is evidence for an association between Psychosis Recovery and Electroshock Therapy

- But X^2 and p indicate nothing about the strength of the assocation!

Overview of Today

1. Recap
   - Association between two categorical variables
   - Conditional proportions
   - Association and (in)dependence

2. Hypothesis test for the association between two categorical variables
   - **Strength of association**
   - z-test

2. Recap
   - Next time
   - Example exam question

54

X^2 vs strength of association

Observed ***conditional proportions***

Treatment?

Recovered?

	No	Yes	Total
No	0.51	0.49	100
Yes	0.49	0.51	100

n = 200

$X^2 = 0.08$

$p = 0.78$

Recovered?

	No	Yes	Total
No	0.51	0.49	200
Yes	0.49	0.51	200

n = 400

$X^2 = 0.16$

$p = 0.69$

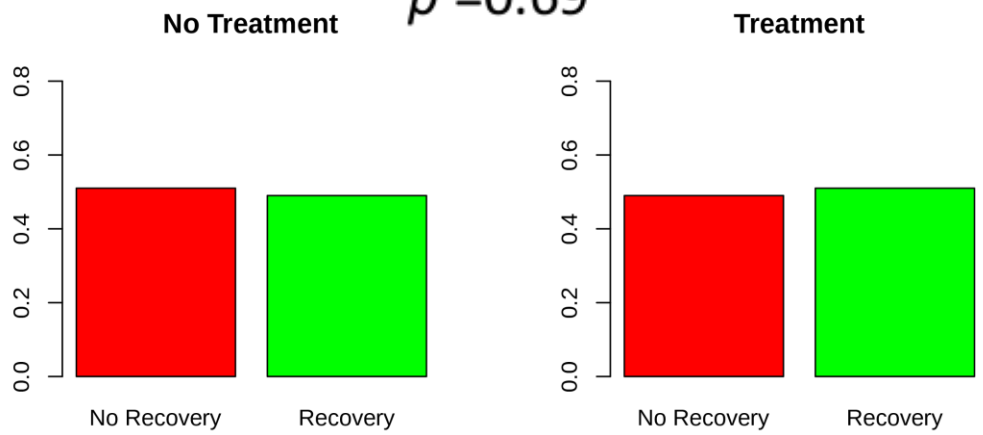
Recovered?

	No	Yes	Total
No	0.51	0.49	10000
Yes	0.49	0.51	10000

n = 20,000

$X^2 = 8.0$

$p = 0.005$



X^2 vs strength of association

High X^2 value ≠ strong association.

A high X^2 implies strong *evidence* against the null hypothesis

Statistical significance ≠ practical significance

For **all hypothesis tests**, if a test shows a statistically significant effect, this does not necessarily mean a big or relevant effect.

A significant effect can be due to a strong association, or due to the sample size

Two Measures for Strength of Association

- Difference of conditional proportions

We compare the percentage recovery in one condition to the other,
the further away from **0** the difference is, the stronger the association

0.275 – 0.675 = -0.4

Proportions:

Recovered?

Treatment?	No	Yes	Total
No	0.26	0.24	0.5
Yes	0.16	0.34	0.5
Total	0.42	0.58	**80(= n)**

Conditional proportions, per level of treatment:

Recovered?

Treatment?	No	Yes	Total
No	0.725	0.275	1
Yes	0.325	0.675	1
Total			**80(= n)**

Two Measures for Strength of Association

- Difference of conditional proportions

Much weaker association!

0.49 - 0.51 = -0.02

Independent of sample size!

Conditional proportions, per level of treatment:

Recovered?

Treatment?	No	Yes	Total
No	0.51	0.49	10000
Yes	0.49	0.51	10000

n = 20,000

Two Measures for Strength of Association

- Ratio of conditional proportions → Relative risk (RR)

We compare the percentage recovery in one condition to the other,
the further away from **1** the ratio is, the stronger the association

0.275 / 0.675 = 0.4074

Proportions:

Recovered?

Treatment?	No	Yes	Total
No	0.26	0.24	0.5
Yes	0.16	0.34	0.5
Total	0.42	0.58	**80(= n)**

Conditional proportions, per level of treatment:

Recovered?

Treatment?	No	Yes	Total
No	0.725	0.275	1
Yes	0.325	0.675	1
Total			**80(= n)**

Two Measures for Strength of Association

- Ratio of conditional proportions

Much weaker association!

0.49 / 0.51 = 0.961

Independent of sample size!

Conditional proportions, per level of treatment:

Recovered?

Treatment?	No	Yes	Total
No	0.51	0.49	10000
Yes	0.49	0.51	10000

n = 20,000

Overview of Today

1. Recap
   - Association between two categorical variables
   - Conditional proportions
   - Association and (in)dependence

2. Hypothesis test for the association between two categorical variables
   - Strength of association
   - **z-test**

2. Recap
   - Next time
   - Example exam question

On Tuesday, Riet said we could do a *z*-test??

Riet is right as well

IF you only have 2 columns and 2 rows, then yes

If more, then you need to use the X^2 test!

From Riet
On first page formula sheet

Comparing two proportions: Significance test

Step 3: Test statistic

- $z = \frac{estimate - null\ hypothesis value}{standard\ error} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{se_0}$

overall proportion if you take both groups together

- with $se_0 = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$

Observed frequencies

Treatment?	Recovered? No	Recovered? Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80**

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{se}_0}$$

$$\text{se}_0 = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

$\hat{p}$= 38/80 = 0.475

$$\text{se}_0 = \sqrt{\frac{0.475(1-0.475)}{40} + \frac{0.475(1-0.475)}{40}} = 0.1117$$

$\hat{p}_1$ = 11/40 = 0.275

$\hat{p}_2$ = 27/40 = 0.675

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{se}_0} = \frac{0.275 - 0.675}{0.1117} = -3.5822$$

We look at the proportion of recovery p, so then for p_1 and p_2, we look at the proportions of recovery for each level of the explanatory variable

Observed frequencies

Recovered?

Treatment?	No	Yes	Total
No	29	11	40
Yes	13	27	40
Total	42	38	**80**

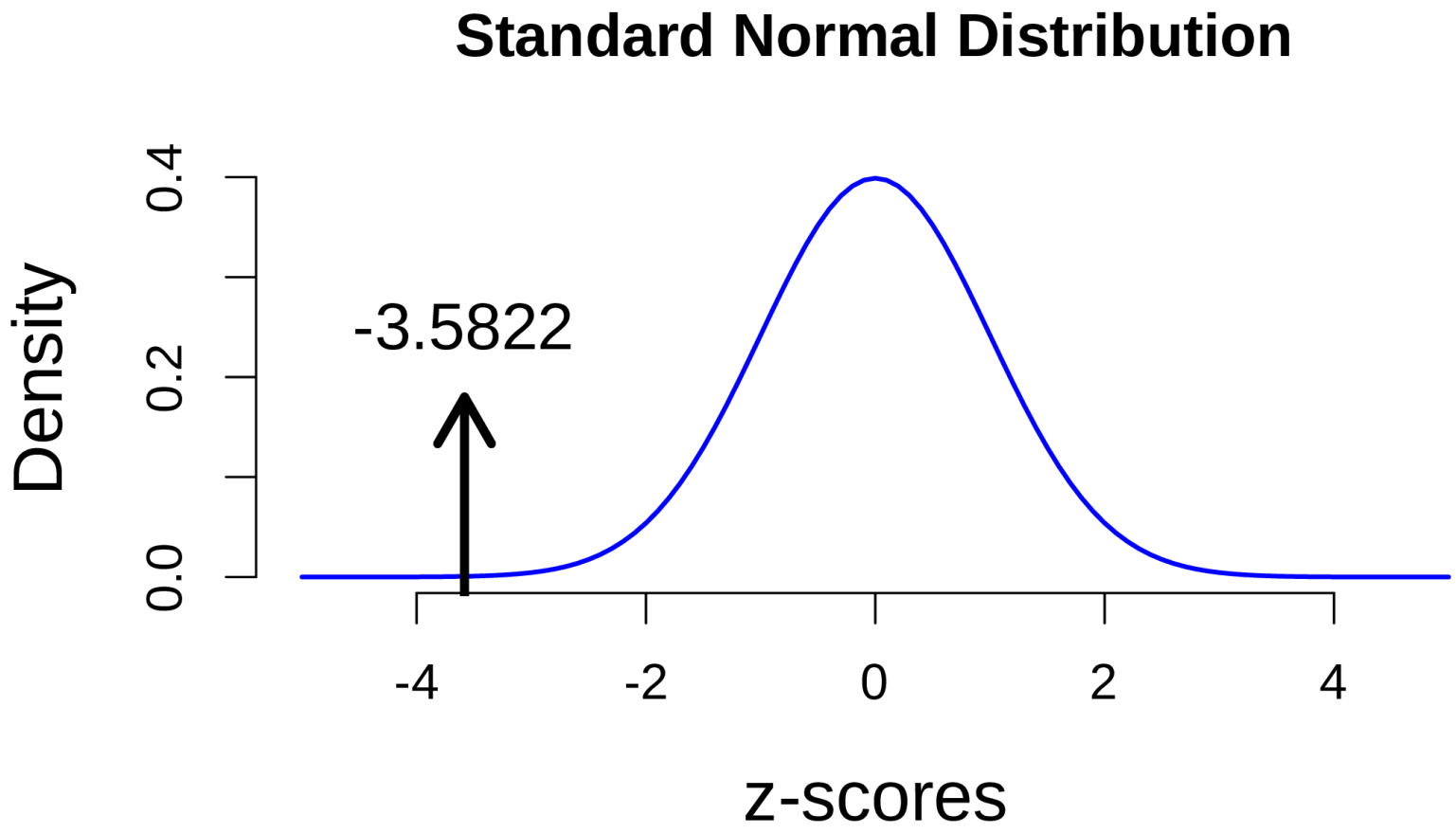
$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\text{se}_0}$$

$$\text{se}_0 = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$$

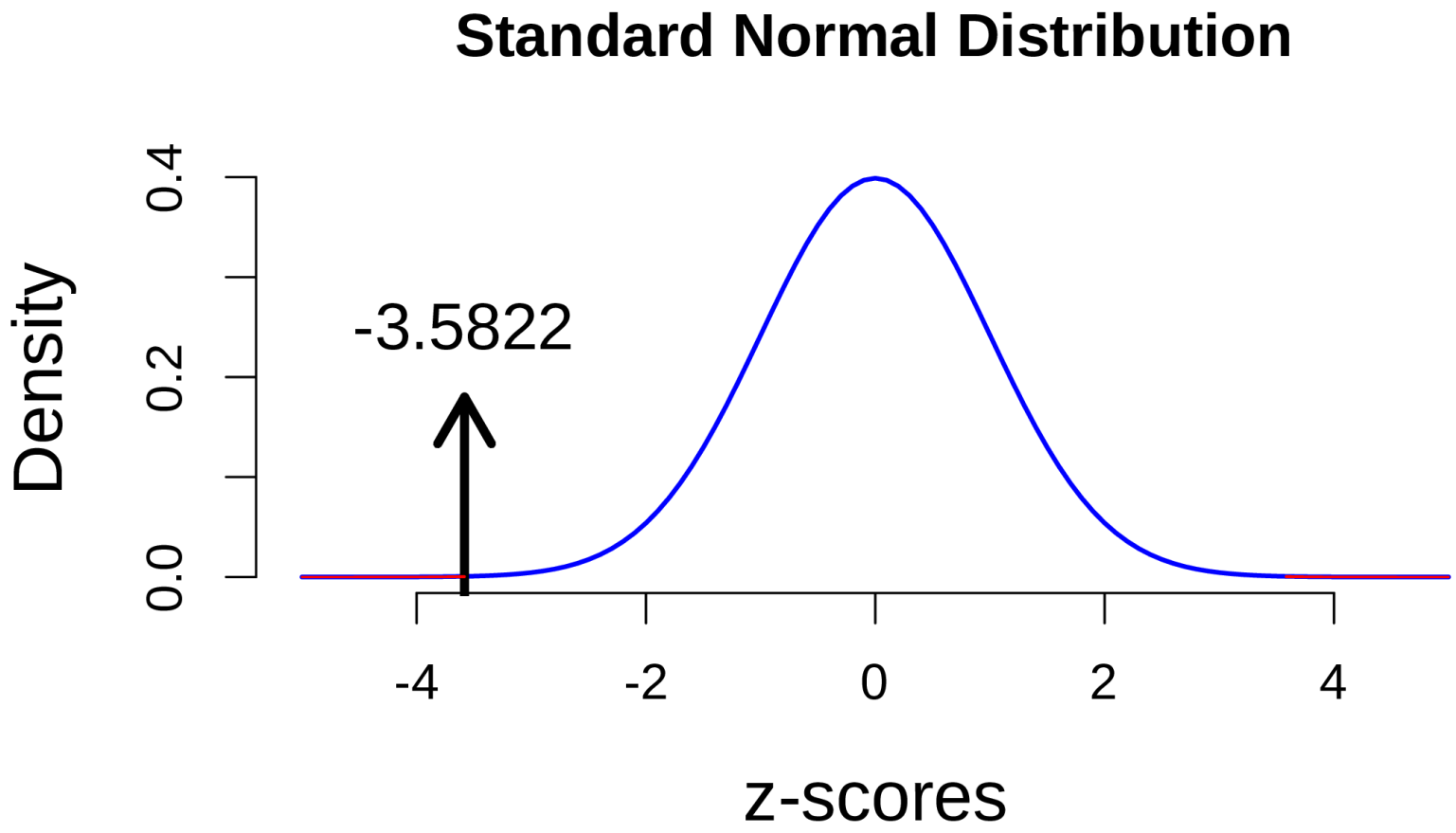
This works for all other combinations of p, p1, and p2.

It all depends on how you define p. In the previous slide, p was proportion recovery, so then p1 and p2 were the proportions of recovery for each level of treatment.
We could also define p as the proportion of **non**-recovery – in that case p1 and p2 should be conditional proportions of non-recovery for each level of treatment (p = 42/80, p1 = 29/40, p2 = 13/40). This leads to exactly the same test statistic (positive instead of negative, but does not matter because we always test two-sided here)

Or, we could define p in terms of the treatment proportion, so then p1 and p2 should be conditional proportions of treatment for each level of recovery (i.e., recovery and non-recovery) → (p = 40/80, p1 = 29/42, p2 = 11/38). Again, we will obtain exactly the same test statistic – crazy!



p-value: Probability that you observe the observed test statistic *or more extreme*, ***if*** H_0 is true.

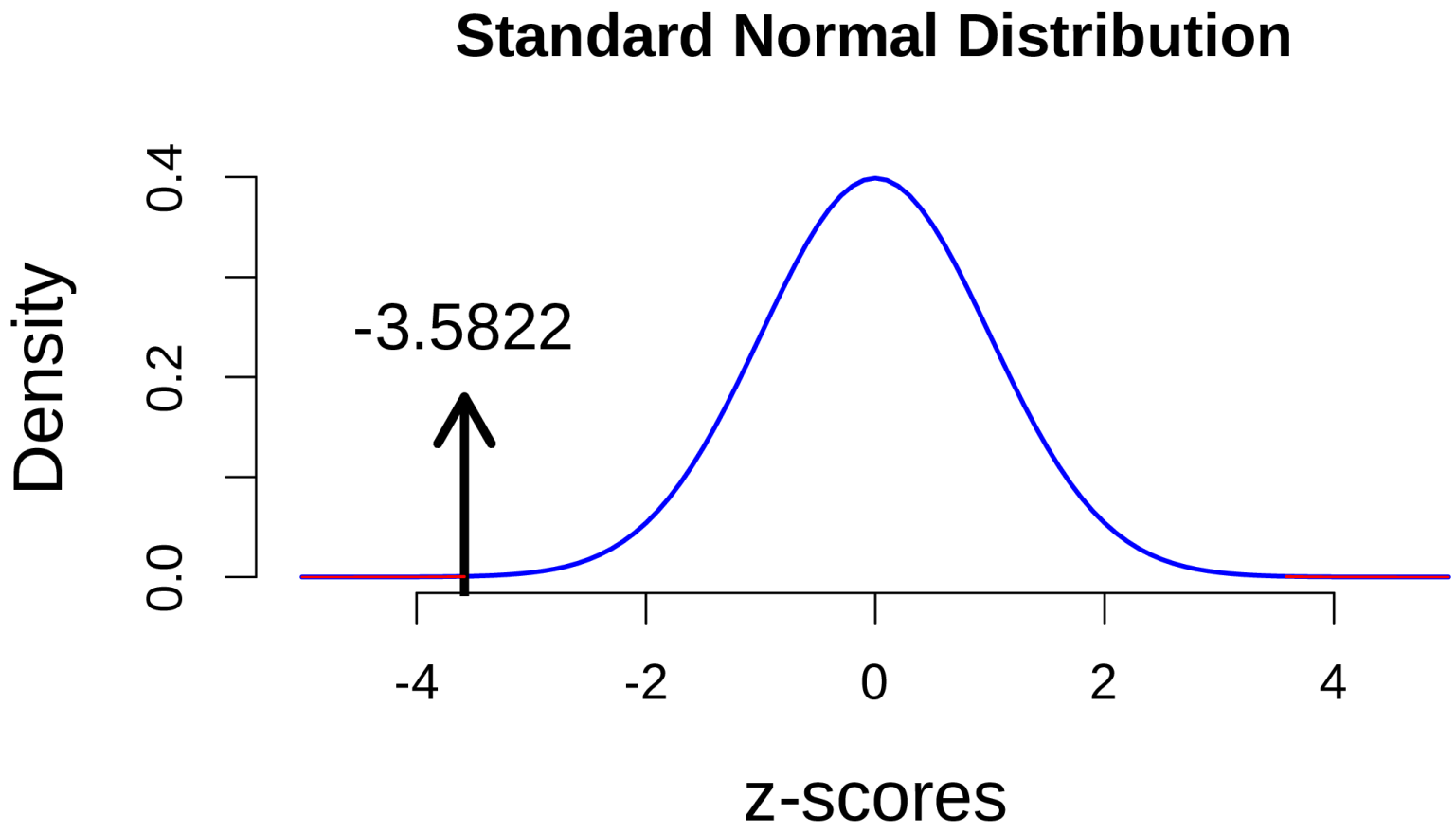


MS Excel: = NORM.DIST(-3.5822, 0 , 1, TRUE)
= 0.00017

mean

Test statistic

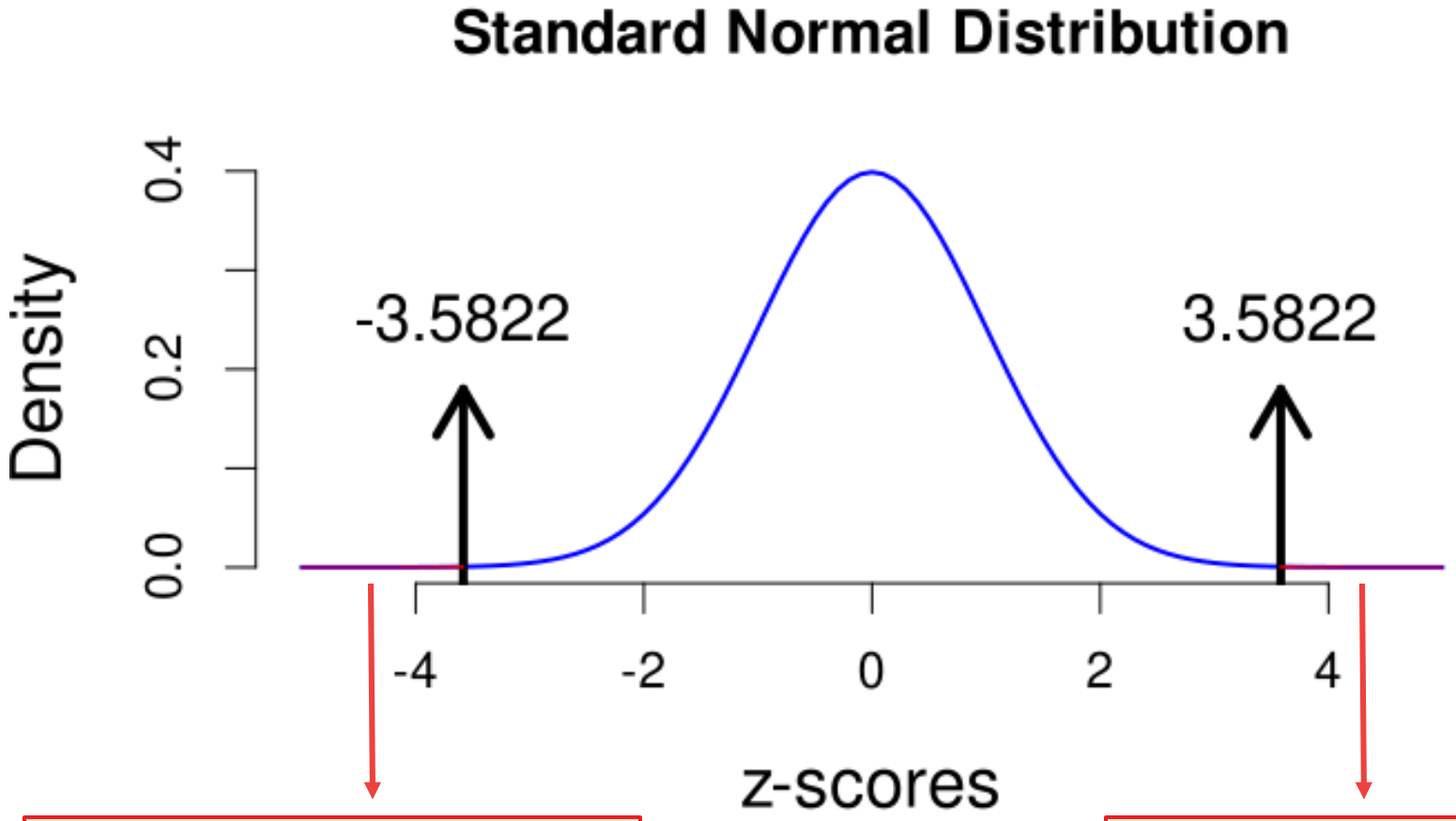
sd



MS Excel: `= NORM.DIST(-3.5822, 0 , 1, TRUE)`

`= 0.00017`

Set to TRUE for the area under the curve (to the *left* of the test statistic)



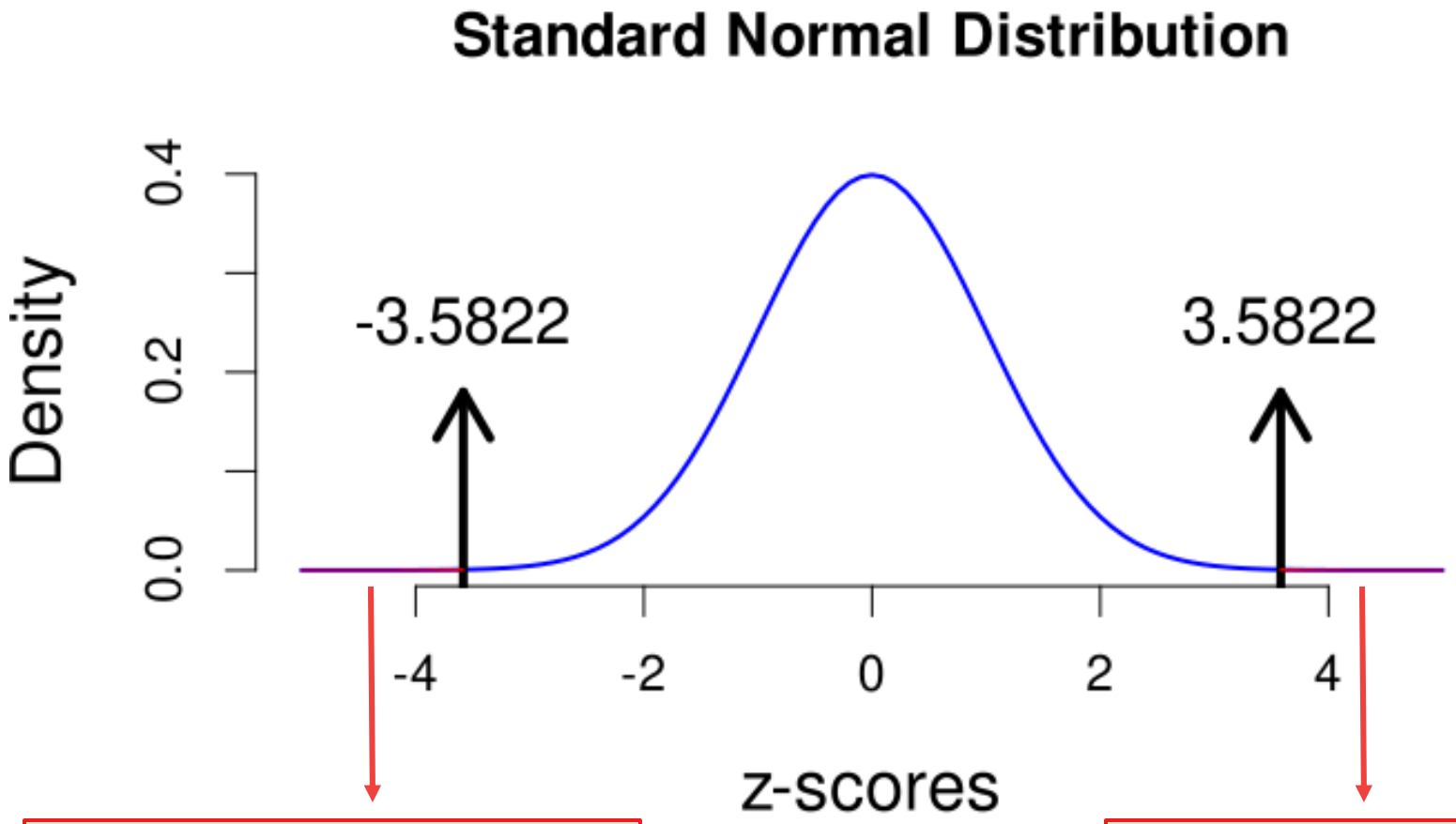
P(z < -3.5822) = 0.00017

P(z > 3.5822) = 0.00017

Because X^2 is always two-sided, we also get two-sided p-value for z = 0.00017*2 = 0.00034

We have obtained exactly the same p-value as with the X^2 test!

MS Excel: = 1-CHISQ.DIST(12.832, 1,TRUE)
= 0.0003407



P(z < -3.5822) = 0.00017

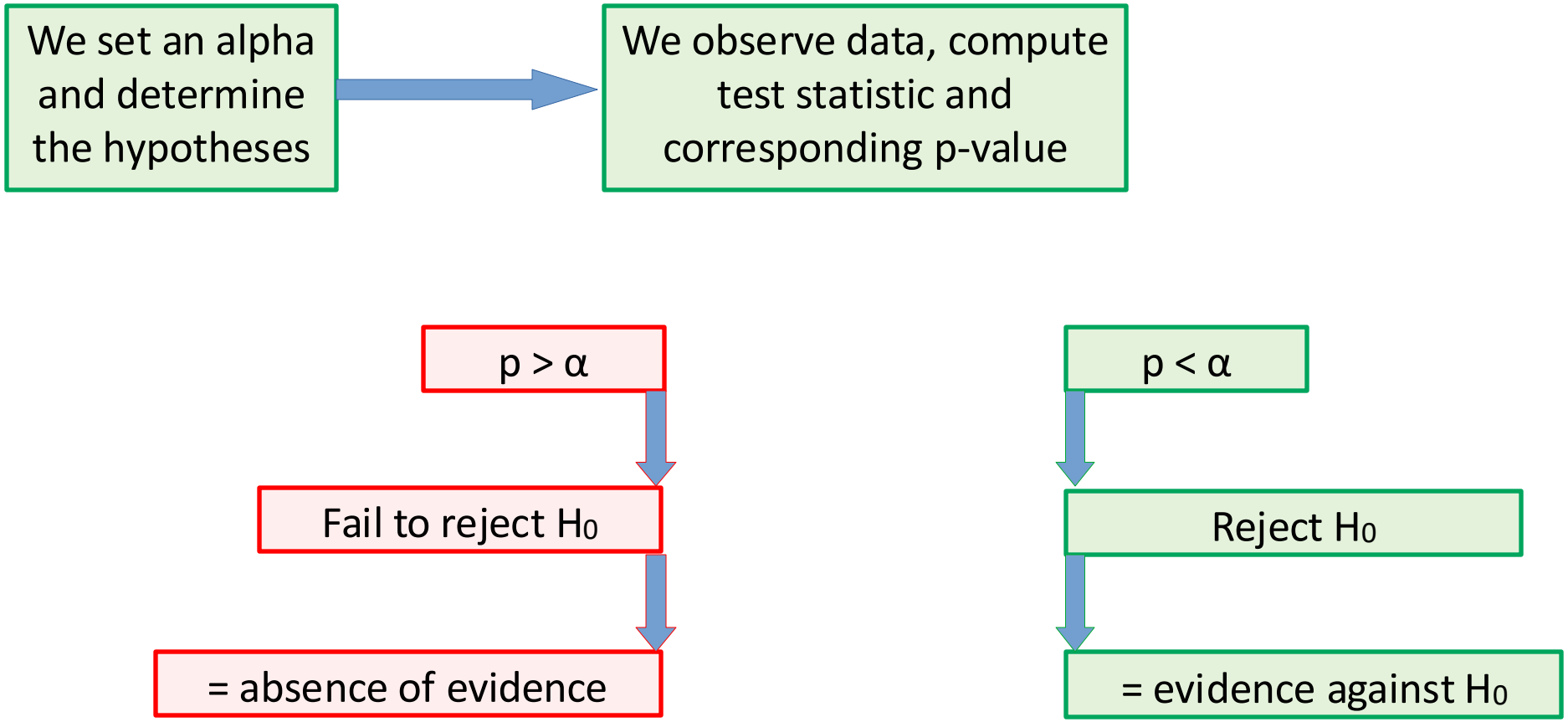
P(z > 3.5822) = 0.00017

Because X^2 is always two-sided: p-value = 0.00017*2 = 0.00034

Note that the z-test can be one-sided!

MS Excel: = 1-CHISQ.DIST(12.832, 1,TRUE)
= 0.0003407

What Are We Doing?



- Black/white reasoning is dangerous and arbitrary (see bonus slide of lecture 16)
- We cannot gain evidence **for** H_0 (i.e., evidence of absence)

Intermezzo – Meaning of the p-value

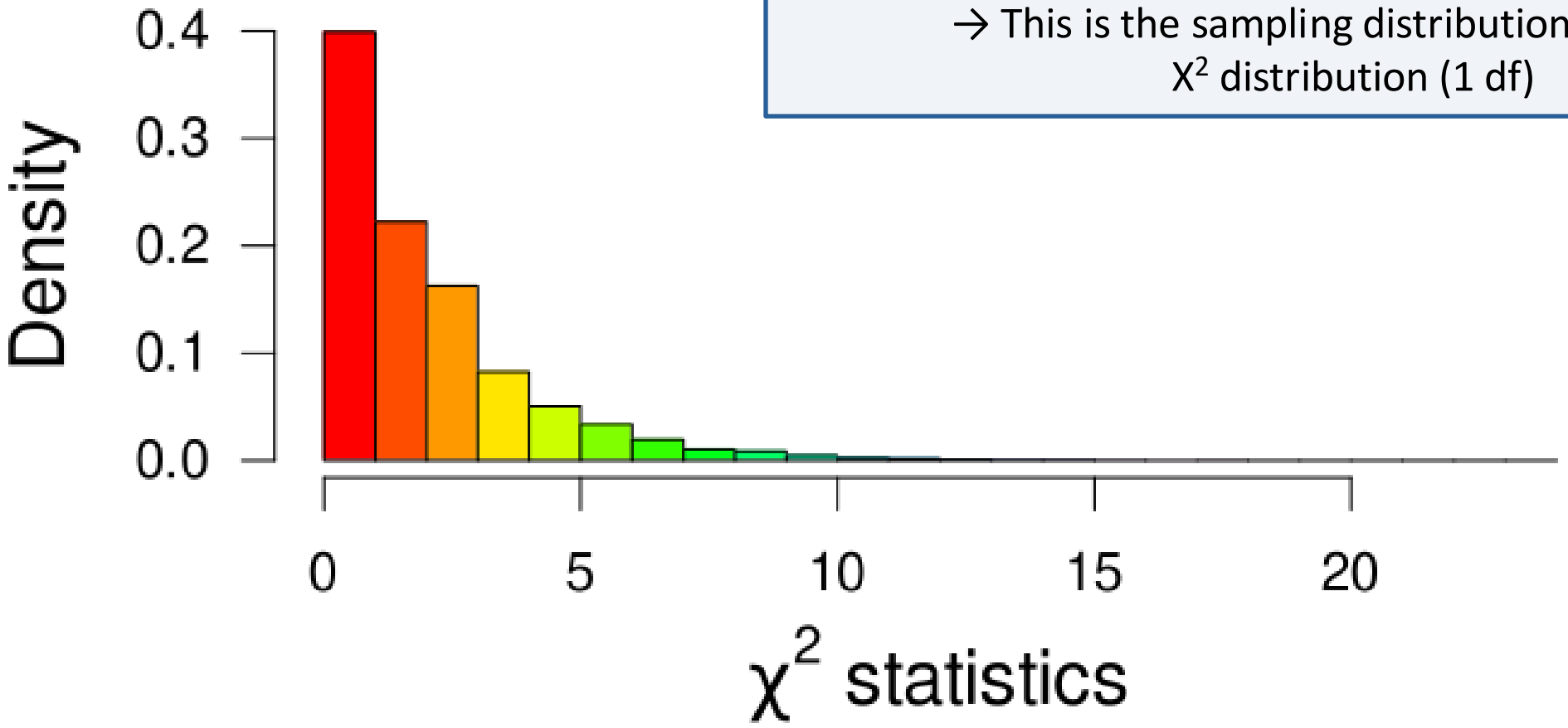
- In frequentist hypothesis testing, we want to minimize the probability of falsely rejecting H_0 (i.e., type I error)
- Alpha (α) determines this error rate
- If we set α to relatively high value (e.g., 0.05), we more often reject H_0 (= often the goal for publication)
- If we set α to relatively low value (e.g., 0.001), we need a more extreme result before we reject H_0
- Why?!

Why is α the Type I error rate?

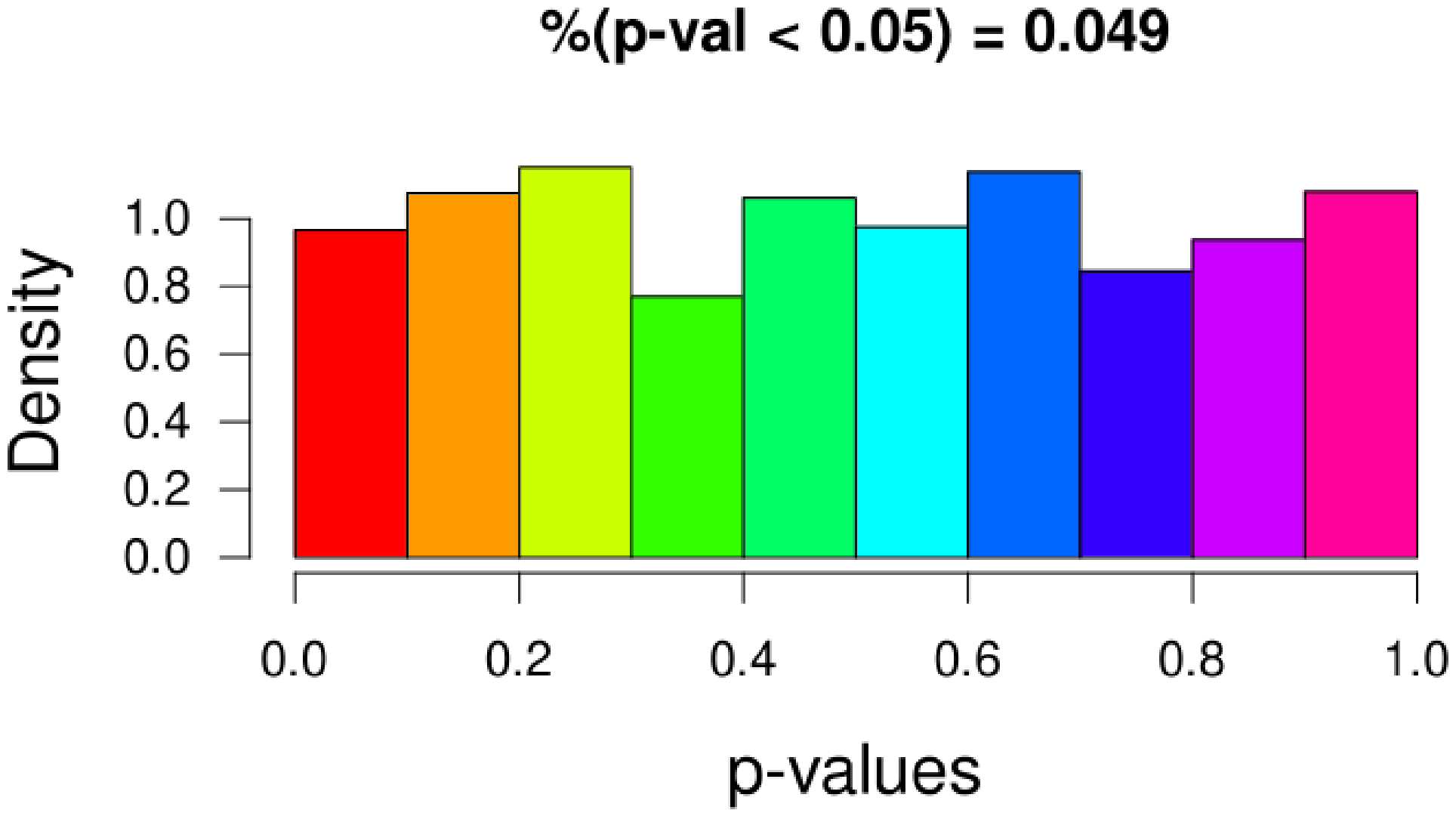
- Imagine we conduct 100,000 identical studies (sampling from the population each time).
- H_0 is true: there is no association between our two categorical variables
- Each time, we compute the X2 statistic and corresponding p-value

Distribution of those 100,000 X^2 Statistics

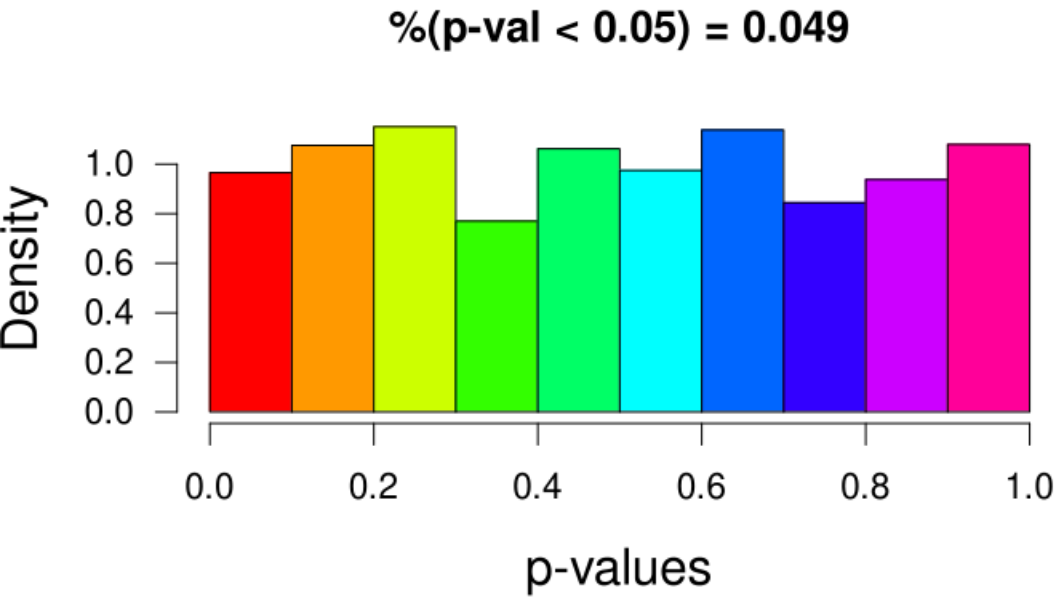
→ This is the sampling distribution under H_0!
X^2 distribution (1 df)



Distribution of the corresponding p-values



Distribution of the corresponding p-values



In about 5% of the cases, we would have wrongfully rejected H0 (i.e., make type I error) if we used $\alpha = 0.05$, just because we observed a significant deviation from H_0 due to chance.

If we would have used $\alpha = 0.001$, we would have made a type I error in about 0.1% of the cases

If we would have chosen our alpha "after the fact", we are not able to make any statement about our risk of a type I error

How do you feel about these snacks?

edu.nl/4twmf

Overview of Today

1. Recap
   - Association between two categorical variables
   - Conditional proportions
   - Association and (in)dependence

2. Hypothesis test for the association between two categorical variables
   - Strength of association
   - z-test

2. Recap
   - Next time
   - Example exam question

Recap

1. In order to test if there is an association between two categorical variables, we can perform a X^2 test
2. This test compares the *observed* cell frequencies to the *expected* cell frequencies under H_0
3. If these frequencies differ a lot (i.e., the data are a poor match with H_0), or if n is high, this will lead to a high X^2 statistic
4. To measure the magnitude/strength of the association, we can look at the ratio or difference between the conditional proportions

Recap – basic flow of conducting X^2-test:

1. Make a frequency table
2. Compute the expected cell frequencies under H_0
   - This can be done using the formula from slide 31
   - Or you can get the marginal proportions (e.g., proportion Recovered/Not-recovered, proportion Treatment/Not-treatment) and use the multiplication rule for independent events
3. Compute X^2
4. Look up the probability of X^2, or more extreme, in Excel
5. Compare to the pre-specified alpha level

Overview of Today

1. Recap
   - Association between two categorical variables
   - Conditional proportions
   - Association and (in)dependence

2. Hypothesis test for the association between two categorical variables
   - Strength of association
   - z-test

2. Recap
   - Next time
   - **Example exam question**

Example exam question

- The table indicates for a sample of 1000 children when they were born, and what high school level they attend.

School?

Born?	vmbo	havo	vwo	total
Oct - Jan	180	120	100	400
Feb - May	150	90	60	300
Jun -Sep	160	90	50	300
total	490	300	210	**1000 (= n)**

- Is there an association between birth month and high school level? Assume a significance level of 0.05

Answer

1. Assumptions → Categorical variable, random sample, large sample → Chisquare test

2. Hypothesis → H_0: the variables are independent

3. Test statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$

4. P-value

5. Conclusion

Answer

$$\text{expected cell frequency} = \frac{(\text{row total}) \times (\text{column total})}{\text{sample size (n)}}$$

- Expected cell frequencies under H_0?

School?

Born?

	vmbo	havo	vwo	total
Oct - Jan	400*490/1000	400*300/1000	400*210/1000	400
Feb - May	300*490/1000	…		300
Jun -Sep				300
total	490	300	210	**1000 (= n)**

Answer

$$\text{expected cell frequency} = \frac{(\text{row total}) \times (\text{column total})}{\text{sample size (n)}}$$

- Expected cell frequencies under H_0 – Observed frequencies?

School?

Born?

	vmbo	havo	vwo	total
Oct - Jan	180 - 196	120 - 120	100 - 84	400
Feb - May	150 - 147	90 - 90	60 - 63	300
Jun -Sep	160 - 147	90 - 90	50 - 63	300
total	490	300	210	**1000 (= n)**

Answer

- Each cell: $\frac{(O-E)^2}{E}$

School?

Born?

	vmbo	havo	vwo	total
Oct - Jan	$(180 - 196)^2/196$	$(120 - 120)^2/120$	$(100 - 84)^2/84$	400
Feb - May	$(150 - 147)^2/147$	$(90 - 90)^2/90$	$(60 - 63)^2/63$	300
Jun -Sep	$(160 - 147)^2/147$	$(90 - 90)^2/90$	$(50 - 63)^2/63$	300
total	490	300	210	**1000 (= n)**

Answer

School?

Born?

	vmbo	havo	vwo	total
Oct - Jan	1.306	0	3.048	
Feb - May	0.061	0	0.143	
Jun -Sep	1.150	0	2.683	
total				

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 1.306 + 0.061 + 1.150 + 0 + 0 + 0 + 3.048 + 0.143 + 2.683 = 8.39$$

87

Answer

1. Assumptions → Categorical variable, random sample, lrage sample → Chisquare test

2. Hypothesis → H_0: the variables are independent

3. Test statistic: $\chi^2 = \sum \frac{(O-E)^2}{E} = 8.39$

4. P-value

5. Conclusion

88

Answer

- P-value: Probability that you observe this statistic *or more extreme*, if H_0 is true.
- We need to know the df
- df = (r-1)(c-1) (r: row; c: column)
  - df = (3-1)(3-1) = 4
- Look up in Table C:
  - $0.050 < P(X^2 \geq 8.39) < 0.100$
- MS Excel: `= 1-CHISQ.DIST(8.39,4,TRUE)`
  - p = 0.078

Answer

1. Assumptions → Categorical variable, random sample, lrage sample → Chisquare test

2. Hypothesis → H_0: the variables are independent

3. Test statistic: $\chi^2 = \sum \frac{(O-E)^2}{E} = 8.39$

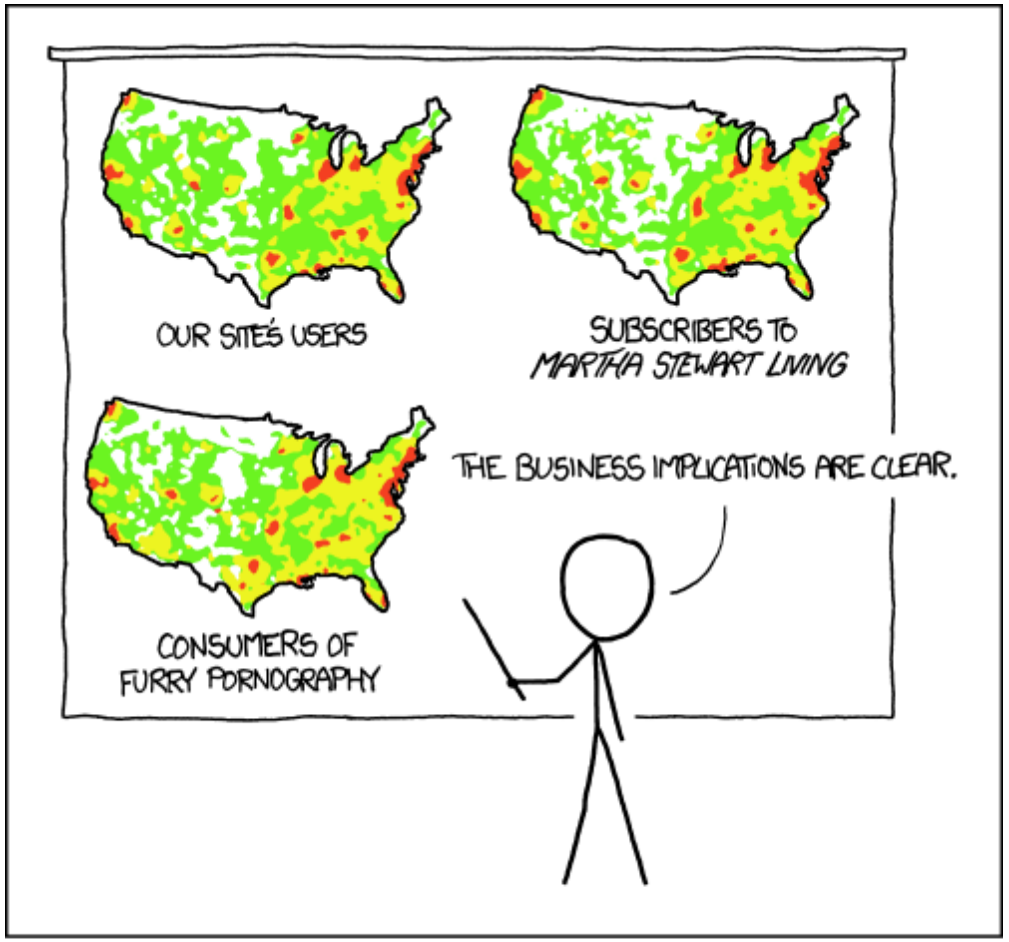
4. P-value = 0.078

5. Conclusion: Because significance level was set at 0.05, we cannot reject H_0

90

Questions?

Thank you for your attention



PET PEEVE #208:
GEOGRAPHIC PROFILE MAPS WHICH ARE BASICALLY JUST POPULATION MAPS

https://xkcd.com/1138/

Bonus Movie

The screenshot from the electroshock therapy was taken form the 1975 movie "*One flew over the cuckoo's nest*", starring Jack Nickolson.

It's an amazing movie (imo), and features some interesting and disturbing insights into the mental healthcare in the 60's

Trailer

Excel commands for calculations:

Regular numeric operations:
= 2.3 + 3 - 7 * SQRT(10) / 4^2

For distributions → surface area:
= NORM.DIST(2, 0, 1, TRUE)
= T.DIST(2, 14, TRUE)
= CHISQ.DIST(8, 1, TRUE)

You start with inputting the observed test statistic. Then you give the df (for X^2 and t distributions), or the mean and sd (for the normal/z distribution). Then you add TRUE to get the area under the curve, to the LEFT of the test statistic.

For distributions-→ critical value:
= NORM.INV(0.025, 0, 1)
= T.INV(0.025, 20)
= CHISQ.INV(0.95, 1)

You start with inputting the desired probability. For instance, 0.025 gives you the value of the distribution, where to the left of that value, 2.5% of the distribution is situated. This is useful when constructing the 95% CI (not relevant for X^2).
Then you give the df (for X^2 and t distributions), or the mean and sd (for the normal/z distribution).

→ Paste these into any cell and press enter!

Highlighted exercises from the book

(sections 11.1, 11.2, 11.3 most important)

- 11.10 (“what gives p-value = 0.05”)
- 11.18 (“z test for heart attack study”)
- 11.32 (“Marital happiness ”)

→ try yourself first, then check next slides for answers

11.10

Df = (#rows-1) * (#columns-1)

a) Df = 1, X^2 = 3.84
b) Df = 2, X^2 = 5.99
c) Df = 4, X^2 = 9.49
d) Df = 16, X^2= 26.3
e) Df = 16, X^2 = 26.3

So, for instance, a X^2 value of 4 would not lead to a significant effect if you have 2 df (e.g., 2 rows and 3 columns), but does lead to a significant effect when you have 4 df (e.g., 3 rows and 3 columns)!

Use Excel: CHISQ.INV(0.95, 1)

11.18

Proportion of heart attacks in aspirin condition: 18/676 = 0.0266 = p_1

Proportion of heart attacks in placebo condition: 28/684 = 0.0409 = p_2

H_0: $p_1 = p_2$
H_A: $p_1 =/= p_2$

The previous result was a p-value of 0.144 and X^2 value of 2.1,
In the z-test, we can square the z value to get X^2: $1.46^2 = 2.1$ (rounded to 1 decimal). This is because of the connect between the z-test and the X^2 test. We can transform one to the other, and they have exactly the same p-value (df = 1 for X^2 test)! For the z-test this is the *two-sided* p-value.

11.32

A

- $\chi^2 = 214$
- Degrees of freedom = (3−1)×(3−1) = 4
- Critical value for df=4 at α=0.05 ≈ 9.49

Because 214 >> 9.49, we reject H_0.
Interpretation:
There is strong evidence that marital happiness and general happiness are associated — in other words, general happiness depends on marital happiness. However, a significant chi-square result does not indicate a strong effect

B

A large χ^2 value indicates a statistically significant relationship, but not the strength of the association. The χ^2 statistic is heavily influenced by sample size — with 894 people, even moderate differences can yield a large χ^2.

Difference of the conditional proportions:

For completeness' sake, see next slide for the X2 result, if you want to practice it yourself first

11.32

C

Difference in proportions:

P(Not too happy | Marital not too happy) = 11 / 26 = 0.423
P(Not too happy | Marital very happy) = 20 / 588 = 0.034

Difference = 0.423 − 0.034 = 0.389

People who are not too happy in their marriage are about 39 percentage points more likely to be generally not too happy compared to those who are very happy in their marriage.

D

Relative risk:

P(Not too happy | Marital not too happy) = 11 / 26 = 0.423
P(Not too happy | Marital very happy) = 20 / 588 = 0.034

RR = (0.423) / (0.034) = 12.4

Those who are not too happy in their marriage are about 12 times more likely to be generally not too happy than those who are very happy in their marriage.