chi-square test

Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

What is a chi-square test?

- It helps us understand the **relationship** between **two categorical variables**; gender, smoke status, etc.
- Chi-square involves the **frequency** of events; count
- It helps us **compare** what we actually **observed** with what we **expected**
- We use the **chi-square distribution** and critical values to accept or reject our null hypothesis

When to use it

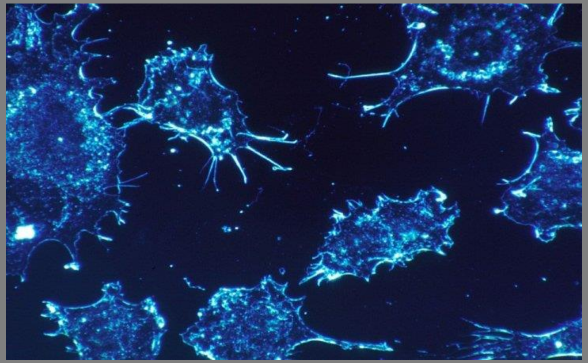
- The chi-square test is used to examine the **relationship** between two **categorical variables** that can be arranged in a 2x2 frequency (contingency) table
- The chi-square test is equivalent to evaluating the **difference between two percentages**

Assumptions

It's assumed that

- the sample is **randomly** selected from the population
- both variables are **categorical**
- every observation in the dataset is **independent**. That is, the value of one observation in the dataset does not affect the value of any other observation
- individuals can only **belong to one cell** in the contingency table. That is, an individual cannot belong to more than one cell
- the **expected value** of cells in the contingency table should be 5 or greater in at least 80% of cells and that no cell should have an expected value less than 1

Effectiveness of a New Anticancer Method



Comparative Effectiveness of a New Anticancer Method Compared to an Old One

Example

In a clinical study evaluating the effectiveness of a new anticancer method (A) compared to an old one (B), 501 patients were randomly divided into two groups corresponding to the two methods. Of the 257 patients who underwent method A, 41 died, while 64 of the 244 patients who underwent method B died.

Questions:

Is there a difference in mortality rates between the two methods, or is there a relationship between the method used and the outcome?

2 x 2 contingency table

	Outcome		
Method	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

We can examine whether the difference between the death rates is significant or, equivalently, whether there is a relationship between the treatment method and the outcome using the **chi-square test**

Comparing observed values with expected values

	Outcome		
Method	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

The chi-square test compares the differences between the 4 observed values in the table above and the corresponding expected values, **assuming that the effectiveness of the two treatment methods is exactly the same** (i.e., they have the same death rate)

Common mortality

	Outcome		
Method	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

Assuming the two methods are the same, the mortality rates for A and B would also be the same.

What is the **common mortality rate** for A and B?

Total mortality index

	Outcome		
Method	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

The common mortality rate can be estimated from the total mortality index, which is 105/501.

Expected values calculation

	Outcome		
Method	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

- Since there are a total of 257 patients in group A, the expected number of deaths for A (assuming the death rates for A and B are the same) is: **E_1 (expected) = 257 x (105 / 501) = 53.86**
- Similarly, the expected number of deaths for B (assuming the death rates for A and B are the same) is: **E_2 (expected) = 244 x (105 / 501) = 51.14**

Expected values calculation

	Outcome		
Method	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

- In the same way, assuming the two methods are the same, the survival rates for A and B would also be the same.
- Therefore, the combined survival rate for A and B can be estimated from the overall survival rate, which is 396/501

Expected values calculation

	Outcome		
Method	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

- Since there are a total of 257 patients in group A, the expected number of survivors for A (assuming the survival rates for A and B are the same) is: **E_3 (expected) = 257 x (396 / 501) = 203.14**
- Similarly, the expected number of survivors for B (assuming the survival rates for A and B are the same) is: **E_4 (expected) = 244 x (396 / 501) = 192.86**

Comparing Observed and Expected Outcomes of Methods A and B

Method	**Outcome**		
	Death (1)	Survival (2)	**Total**
A (1)	O_1=41 (E_1=54)	O_3=216 (E_3=203)	**257**
B (2)	O_2=64 (E_2=51)	O_4=180 (E_4=193)	**244**
Total	**105**	**396**	**501**

- If the observed numbers (observations) **do not differ** from the corresponding expected numbers (expectations), then the two methods would have the same effectiveness
- In this case, all the differences O (observations) – E (expectations) **would be zero**
- The greater the difference between the observed numbers and the corresponding expected numbers, the more the two methods A and B differ

Chi square - test

Method	Outcome		
	Death (1)	Survival (2)	**Total**
A (1)	O_1=41 (E_1=54)	O_3=216 (E_3=203)	**257**
B (2)	O_2=64 (E_2=51)	O_4=180 (E_4=193)	**244**
Total	**105**	**396**	**501**

- Then, the comparison of observed and expected values is given by the relation :

  $(O_1-E_1)+(O_2-E_2)+(O_3-E_3)+(O_4-E_4)=(41-54)+(64-51)+(216-203)+(180-193)$

- To eliminate the negative signs, we square the differences

  $(O_1-E_1)^2+(O_2-E_2)^2+(O_3-E_3)^2+(O_4-E_4)^2=(41-54)^2+(64-51)^2+(216-203)^2+(180-193)^2$

  and calculate the differences

Chi square – test calculation

Method	Outcome		
	Death (1)	Survival (2)	**Total**
A (1)	O_1=41 (E_1=54)	O_3=216 (E_3=203)	**257**
B (2)	O_2=64 (E_2=51)	O_4=180 (E_4=193)	**244**
Total	**105**	**396**	**501**

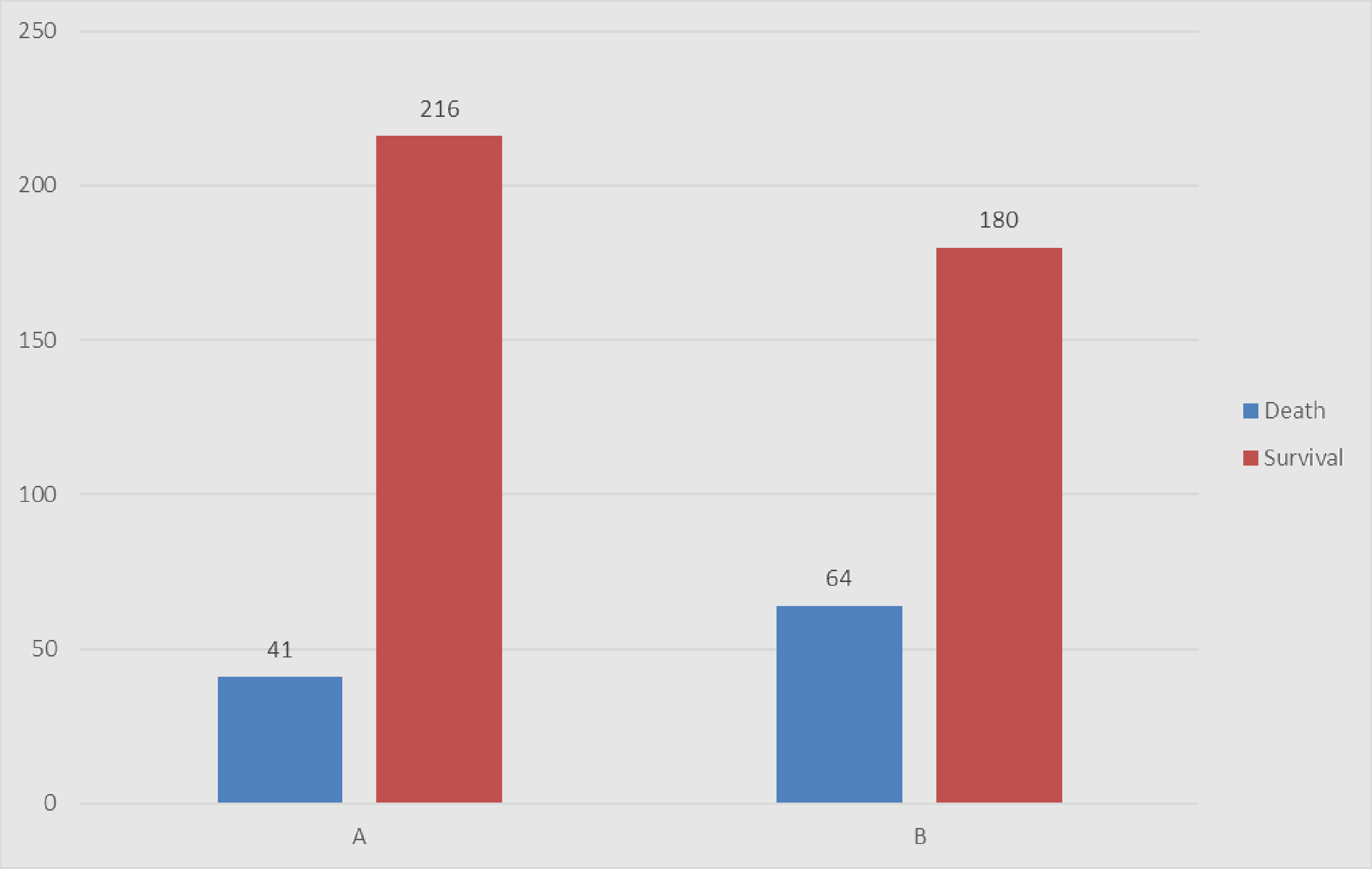
$$\chi^2 = \frac{\sum (O_i - E_i)^2}{E_i}$$

$$\chi^2 = \frac{(O_1-E_1)^2}{E_1} + \frac{(O_2-E_2)^2}{E_2} + \frac{(O_3-E_3)^2}{E_3} + \frac{(O_4-E_4)^2}{E_4} = \frac{(41-54)^2}{54} + \frac{(64-51)^2}{51} + \frac{(216-203)^2}{203} + \frac{(180-193)^2}{193} = 7.98$$

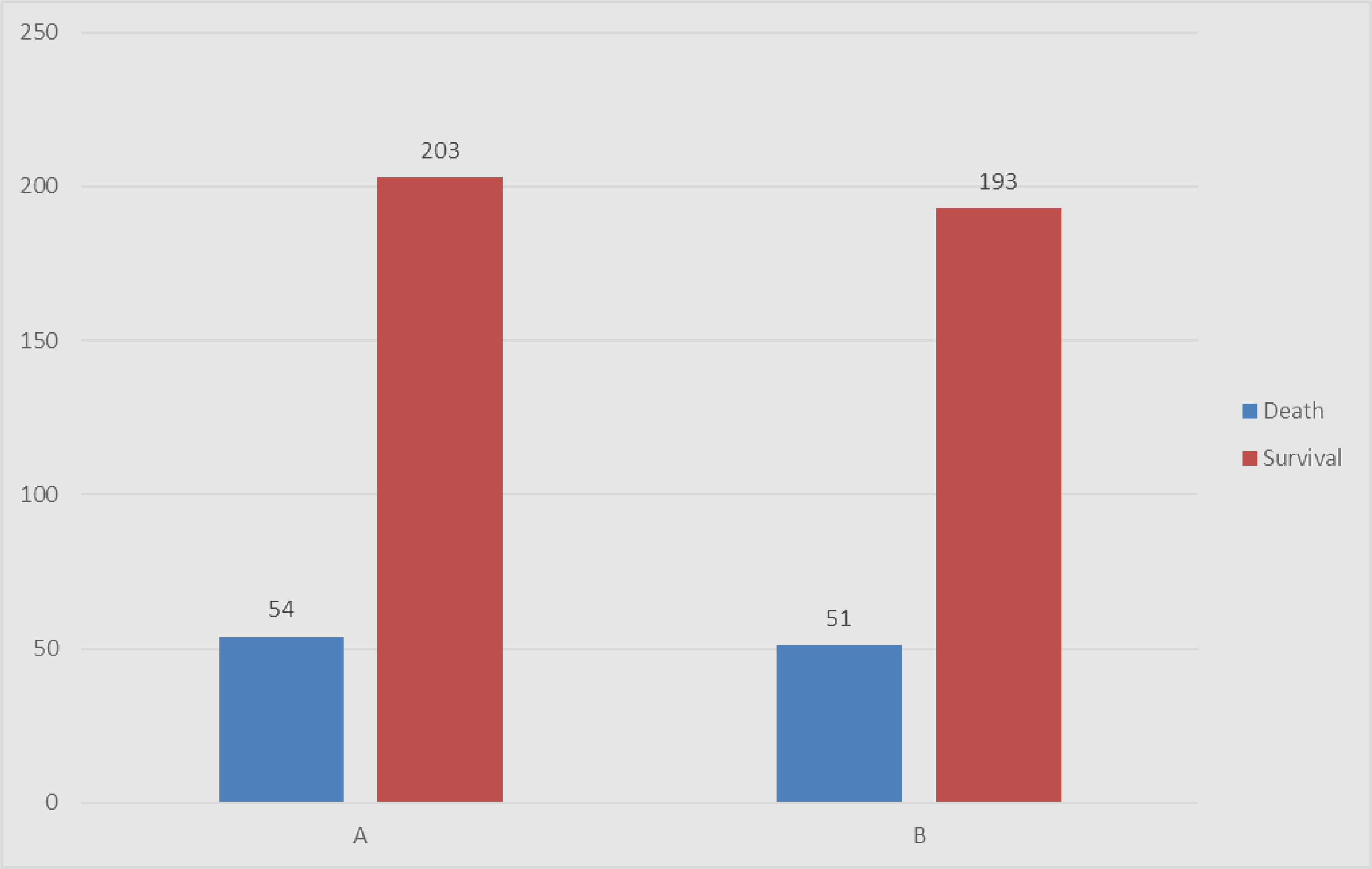
$(O - E)^2/E$

Method	Outcome		
	Death (1)	Survival (2)	**Total**
A (1)	3.070546	0.814116	3.884662233
B (2)	3.31373	0.857511	4.09137114
Total	6.304406	1.671628	**7.976033373**

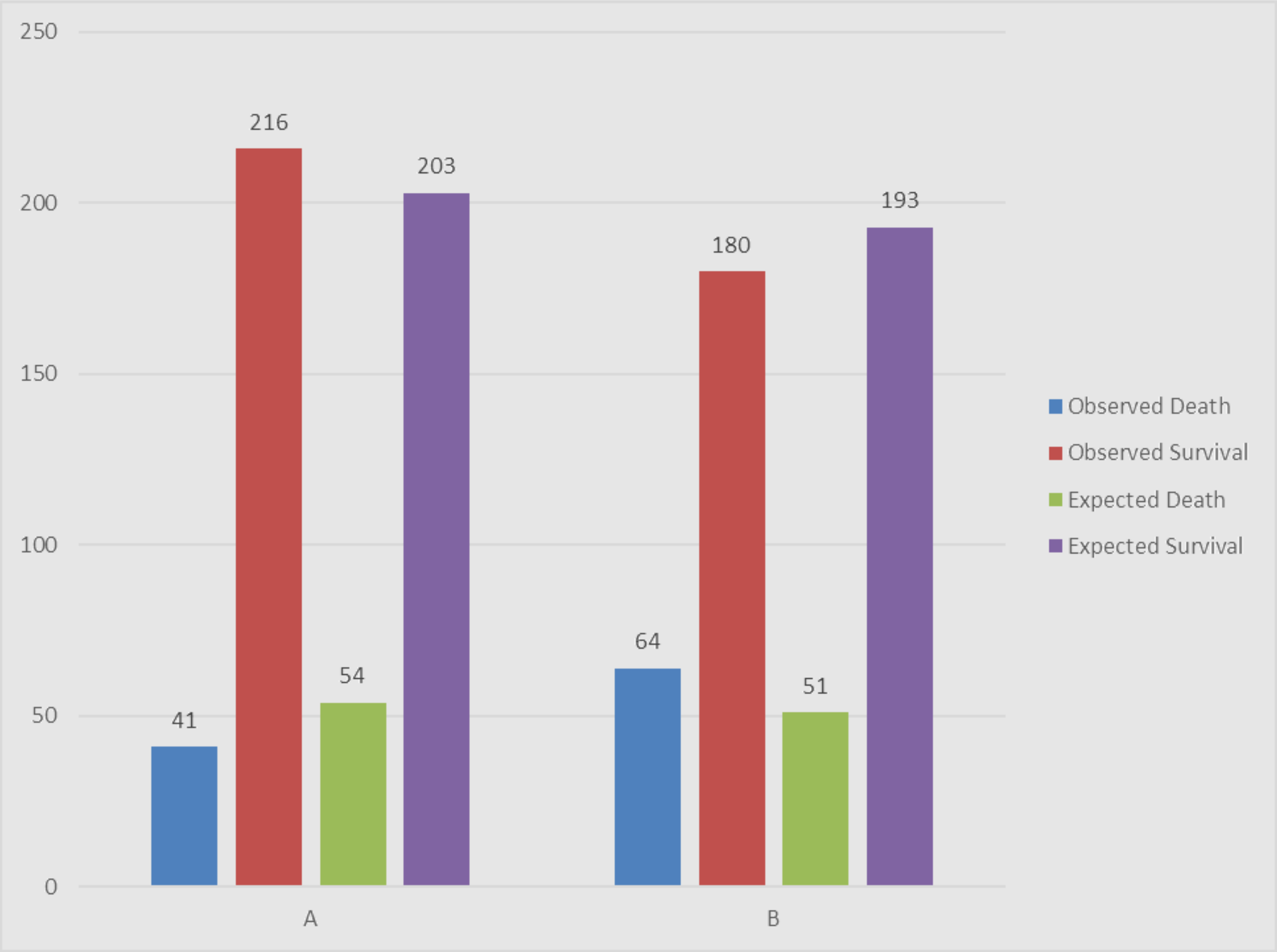
Bar chart of the observed values



Bar chart of the expected values



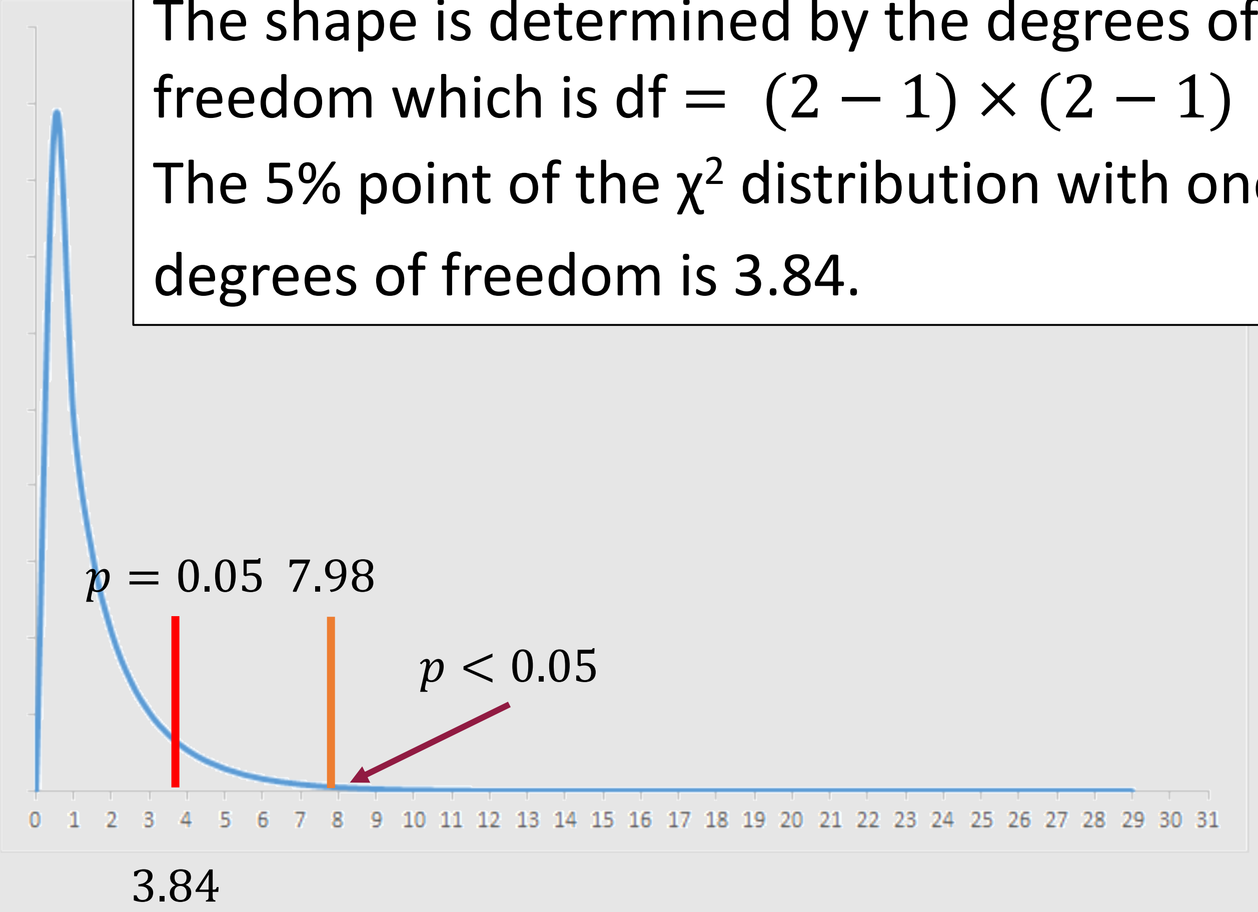
Observed versus expected values



Chi-square distribution

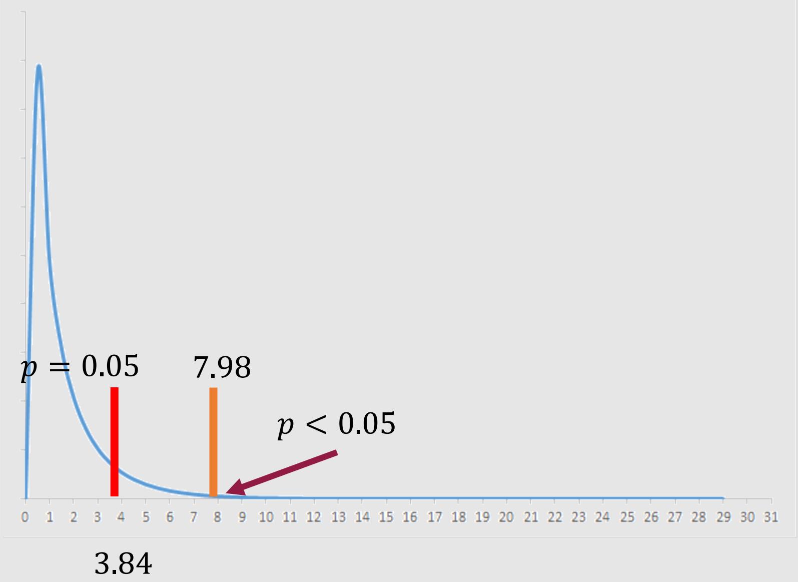
If we repeat (simulate) the study many times (assuming the variables are not related), calculate the χ^2 value each time, and construct the χ^2 distribution, then it will have the shape shown on the right

The shape is determined by the degrees of freedom which is df $= (2-1) \times (2-1) = 1$.
The 5% point of the χ^2 distribution with one degrees of freedom is 3.84.



Results

- The χ^2 produced the a value of 7.98
- We will use the 5% point of the χ^2 distribution
- The degrees of freedom are calculated as $(2-1)\times(2-1)=1$
- The 5% point of the χ^2 distribution, which is 3.84, is calculated in EXCEL using the function $= CHISQ.INV.RT(0.05, 1)$
- We conclude that the value 7.98 is not due to chance (i.e. is not random) because $\chi^2 > critical, 7.98 > 3.84$
- So, we expect we expect a percentage $p < 0.05$ $(p = 0.005)$ of the χ^2 – tests we simulated to be greater than 7.98



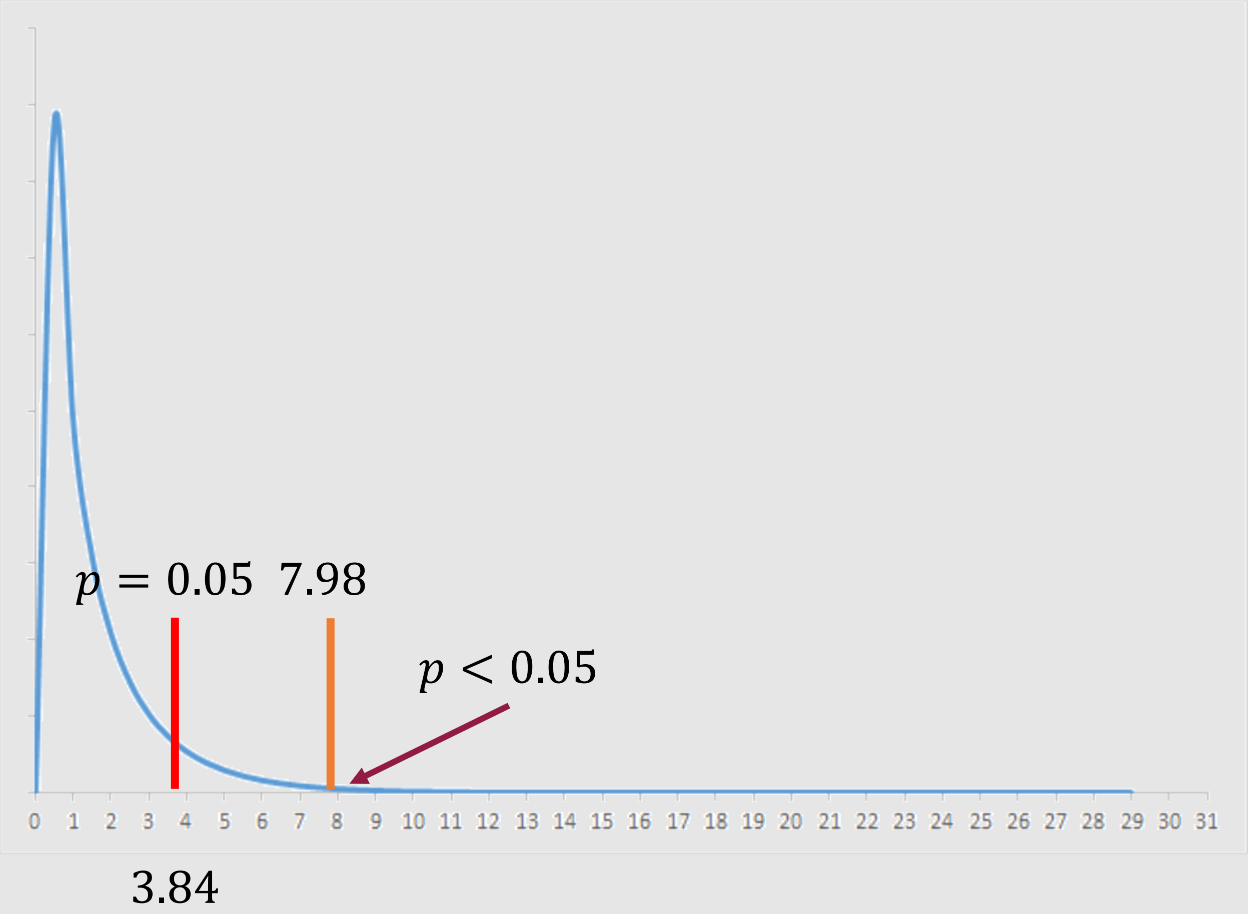
Finding the 5% point of the χ^2 distribution with one degrees of freedom in EXCEL

3.841459		
=CHISQ.INV.RT(0.05, 1)		

Interpretation

Therefore, with a small probability error $p < 0.05$ $(p = 0.005)$ we conclude that

- the two methods A and B are different or
- the outcome is related to the method or
- The mortality rates between the two methods are different



Chi-square distribution

	Percentace points of the chi-square distribution			
	p-value			
df	0.1	0.05	0.025	0.01
1	2.71	3.84	5.02	6.63
2	4.61	5.99	7.38	9.21
3	6.25	7.81	9.35	11.34
4	7.78	9.49	11.14	13.28
5	9.24	11.07	12.83	15.09
6	10.64	12.59	14.45	16.81
7	12.02	14.07	16.01	18.48
8	13.36	15.51	17.53	20.09
9	14.68	16.92	19.02	21.67
10	15.99	18.31	20.48	23.21
11	17.28	19.68	21.92	24.72
12	18.55	21.03	23.34	26.22
13	19.81	22.36	24.74	27.69
14	21.06	23.68	26.12	29.14
15	22.31	25	27.49	30.58
16	23.54	26.3	28.85	32
17	24.77	27.59	30.19	33.41
18	25.99	28.87	31.53	34.81
19	27.2	30.14	32.85	36.19
20	28.41	31.41	34.17	37.57
21	29.62	32.67	35.48	38.93
22	30.81	33.92	36.78	40.29
23	32.01	35.17	38.08	41.64
24	33.2	36.42	39.36	42.98
25	34.38	37.65	40.65	44.31
26	35.56	38.89	41.92	45.64
27	36.74	40.11	43.19	46.96
28	37.92	41.34	44.46	48.28
29	39.09	42.56	45.72	49.59
30	40.26	43.77	46.98	50.89

Odds Ratio

Probability review

$$p = \frac{outcomes\ of\ interest}{all\ possible\ outcomes}$$

Fair coin flip

$$p(heads) = \frac{1}{2} = 0.5$$

Fair die roll

$$p(1\ or\ 2) = \frac{2}{6} = \frac{1}{3} = 0.333$$

Deck of playing cards

$$p(hearts) = \frac{13}{52} = \frac{1}{4} = 0.25$$

What are the odds?

$$odds = \frac{P(occuring)}{P(not\ occuring)}$$

$$odds = \frac{p}{1 - p}$$

Fair coin

$$odds(heads) = \frac{0.5}{0.5} = 1\ or\ 1{:}1$$

Fair die roll

$$odds(1\ or\ 2) = \frac{0.333}{0.666} = \frac{1}{2} = 0.5\ or\ 1{:}2$$

Deck of playing cards

$$odds(hearts) = \frac{0.25}{0.75} = \frac{1}{3} = 0.333\ or\ 1{:}3$$

Odds Ratio

The odds ratio is exactly what it says it is, a ration of two odds

Fair coin flip

$$p(heads) = \frac{1}{2} = 0.5$$

$$odds(heads) = \frac{0.5}{0.5} = 1 \; or \; 1{:}1$$

Loaded coin flip

$$p(heads) = \frac{6}{10} = 0.7$$

$$odds(heads) = \frac{0.6}{0.4} = 1.5$$

$$odds\ ratio = \frac{odds_1}{odds_0}$$

$$odds\ ratio = \frac{\frac{p_1}{1 - p_1}}{\frac{p_0}{1 - p_0}}$$

$$odds\ ratio = \frac{\frac{0.6}{0.4}}{\frac{0.5}{0.5}} = \frac{0.6}{0.4} \times \frac{0.5}{0.5} = 1.5$$

The odds of getting **heads** on a loaded coin are 1.5x greater than the fair coin

Odds ratio

Method	Outcome		
	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

- We have already demonstrated that there is a relationship between treatment and outcome, but we are also interested in determining the magnitude of this relationship
- The magnitude is measured by the **Odds Ratio (OR)**

Odds Ratio

Method	Outcome		
	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

$$OR = \frac{\text{Odds Ratio of death by method } A}{\text{Odds Ratio of death by method } B}$$

$$OR = \frac{\frac{41}{216}}{\frac{64}{180}} = 0.53$$

Therefore, the chance of dying with method B is approximately 1/0.53 = 1.88 times greater than with method A

95% confidence interval for the odds ratio

The significance of the odds ratio (OR) is determined by the 95% confidence interval (CI)

Since the standard error (SE) of the odds ratio (OR) is not known, we first calculate the 95% confidence interval (CI) of ln(OR) :

(ln(OR)-1.96*SE, ln(OR)+1.96*SE)

Method	**Outcome**		
	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

[1] The odds ratio (OR) has a bounded lower limit, as it cannot be negative, but an unbounded upper limit, which results in a non-symmetric distribution. In contrast, the natural logarithm of the odds ratio (ln OR) can take any value and thus follows an approximately normal distribution

95% confidence interval for the odds ratio

95% CI of ln(OR)

(ln(OR)-1.96*SE, ln(OR)+1.96*SE)

Method	Outcome		
	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

ln(OR) = ln(0.53) = -0.63

$$\mathrm{SE}\big(\ln(\mathrm{OR})\big) = \sqrt{\frac{1}{216} + \frac{1}{41} + \frac{1}{180} + \frac{1}{64}} = 0.224$$

Thus, the 95% CI of ln(OR) is: (-0.63-1.96*0.224, -0.63+1.96*0.224) = (-1.07, -0.19)

Therefore, to find the 95% confidence interval (CI) of the odds ratio (OR), we take the antilog of the limits of the 95% CI of ln(OR)

95% confidence interval for the odds ratio

- So, the 95% CI of the OR is $(e^{-1.07}, e^{-0.19}) = (0.34, 0.83)$
- Since this interval **does include 1**, it suggests that the odds ratio is non statistically significant
- Thus, there is an increased risk of death with method B, approximately double.

Method	Outcome		
	Death (1)	Survival (2)	**Total**
A (1)	41	216	**257**
B (2)	64	180	**244**
Total	**105**	**396**	**501**

Quick note: If we use 1/OR, i.e., 1/0.53 = 1.886, we should also adjust the 95% confidence interval (CI) accordingly. In this case, the lower limit becomes the upper limit, with a value of 1/0.34 = 2.906, and the upper limit becomes the lower limit, with a value of 1/0.83 = 1.207. Thus, the 95% CI is (1.207, 2.906)