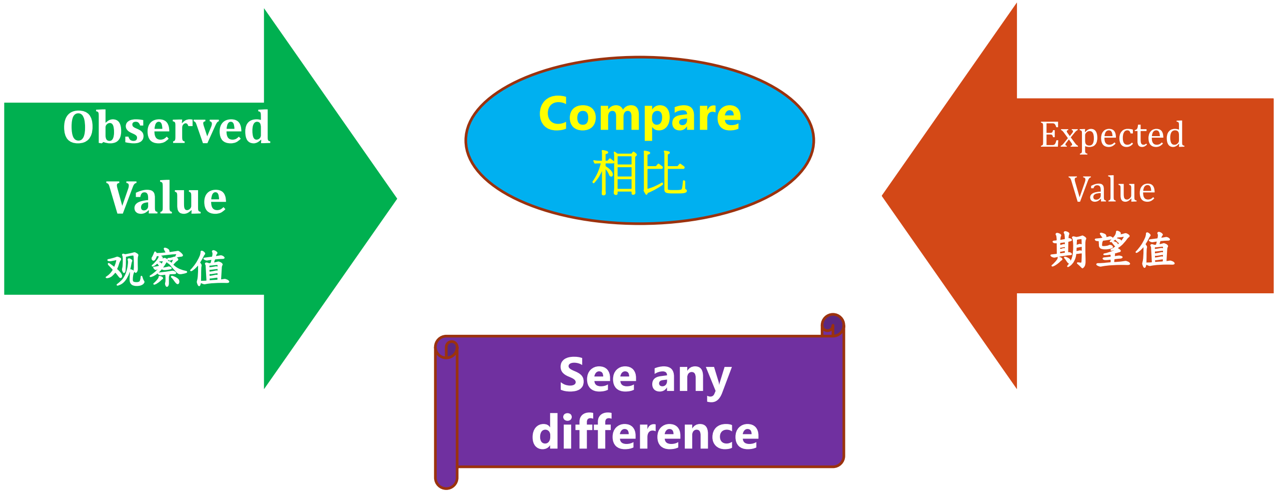
The Chi Squared Test
卡方检验

Ms Ivy Cheng

Fellow of the Hong Kong Academy of Nursing (Mental Health)
International faculty of Asia Pacific EBMN workshop and conference, Singapore

Chi Square Test 卡方检验

- **Test for Proportion 概率测试**
- Understand and analyze the **relationship on Frequency/ counts** between two categorical variables
  理解和分析两个分类变量之间频率/计数的关系



Observed Value 观察值

Compare 相比

Expected Value **期望值**

See any difference

Test of proportion 概率测试

- If Drug A and Drug B have same effect
- 如果药物A和药物B具有相同的效果
- We can get the expected values for the 4 boxes by the following methods
- 我们可以通过以下方法得到4个框的期望值

	Death	Alive	
Drug A	a	b	a+b
Drug B	c	d	c+d
	a+c	b+d	a+b+c+d (all)

$$\textbf{Box expected value} = \frac{\textbf{Row total x column total}}{\textbf{Overall total}}$$

e.g. expected value for the box with observed value a = (a+b) (a+c) / all

Chi Square Test 卡方检验

- Measure **the association** between two categorical variables

  检定两组类别变量的关联性

- Examples of categorical variables with only two categories: Gender (Female and Male), Dead or Alive, Age group
- 只有两个类别的分类变量示例：性别（女性和男性）、死或生、年龄组

  Dichotomous 二分法

- Use the Chi Square distributions and critical value to **accept or reject our hypothesis**
- 使用卡方分布和临界值来接受或拒绝我们的假设

  假设检定:

H_0: A 变项与 B 变项之间没有关联性

H_1: A 变项与 B 变项之间具有关联性

Assumption 前提假设

1. 所有的变项为**类别变项(categorical variable)**

2. 样本须为**独立变项(Independent variable)**

即是：第一组的样本不影响第二组的样本；
第二组的样本也不影响第一组

- 3. 每一检定细格(cell)内的数据应该设为频率(Frequency)或次数(count)，而不是百分比或是经过转换之数据。

Using 2x2 contingency table to explain

Observed value 观察值

	Dead	Alive	
Drug A	a 30	b 40	70 (Row Total) a+b
Drug B	C 60	d 70	130 (Row Total) c+d
	90 (Column Total) a+c	110 (Column Total) b+d	200 Overall Total (a+b+c+d)

How to calculate the expected value 期望值

For each group a/b/c/d

$$\frac{\text{Row total x Column Total}}{\text{Overall total}}$$

For Box A

a

$$\frac{(a+b) \text{ x } (a+c)}{a+b+c+d}$$

$= \underline{(30+40) \text{ x } (30+60)} \;/\; (30+40+60+70)$

$= 31.5$

For Box B

b

$$\frac{(a+b) \text{ x } (b+d)}{a+b+c+d}$$

$$= \frac{(30+40) \text{ x } (40+70)}{200}$$

$= 38.5$

	Dead	Alive	
Drug A	a **30**	b **40**	**70 (Row Total)** **a+b**
Drug B	C **60**	d **70**	**130 (Row Total)** **c+d**
	90 (Column Total) a+c	110 (Column Total) b+d	**200** **Overall Total** **(a+b+c+d)**

How to calculate the expected value 预期值

For each group a/b/c/d

$$\frac{\text{Row total x Column Total}}{\text{Overall total}}$$

	Dead	Alive	
Drug A	a 30	b 40	70 (Row Total) a+b
Drug B	C 60	d 70	130 (Row Total) c+d
	90 (Column Total) a+c	110 (Column Total) b+d	200 Overall Total (a+b+c+d)

For Box C

c

$$\frac{(c+d) \times (a+c)}{a+b+c+d} = \frac{130 \times 90}{200} = 58.5$$

For Box D

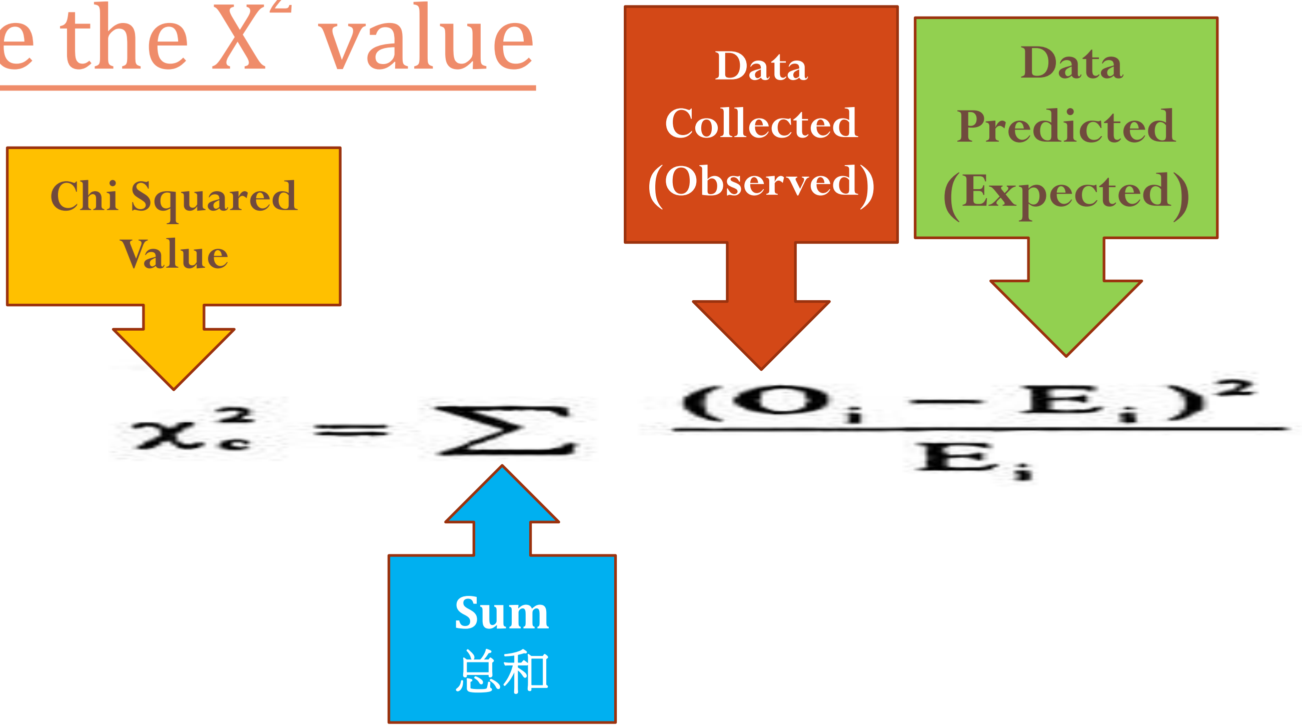
d

$$\frac{(c+d) \times (b+d)}{a+b+c+d} = \frac{130 \times 110}{200} = 71.5$$

See the difference between the observed values and the expected values
查看观察值和期望值之间的差异

	Observed观察值	Expected预期值	Difference差异
Drug A Dead	a **30**	31.5	**- 1.5**
Drug A Alive	b **40**	38.5	**1.5**
Drug B Dead	**C** **60**	58.5	**1.5**
Drug B Alive	**D** **70**	71.5	**-1.5**

Calculate the X^2 value



Chi Squared Value

Data Collected (Observed)

Data Predicted (Expected)

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

Sum 总和

X^2 = (2.25/31.5) +(2.25/38.5) + (2.25/58.5) + (2.25/71.5)

= 0.071 + 0.058+ 0.038 + 0.031

= 0.198

degree of freedom (df) 自由度

- 统计学上的**自由度**，是指当以样本的统计量来估计母体的参数时，样本中独立或能自由变化的数据的个数，称为该统计量的自由度。

维基百科

- 指的是计算某一统计量时，取值不受限制的变量个数。

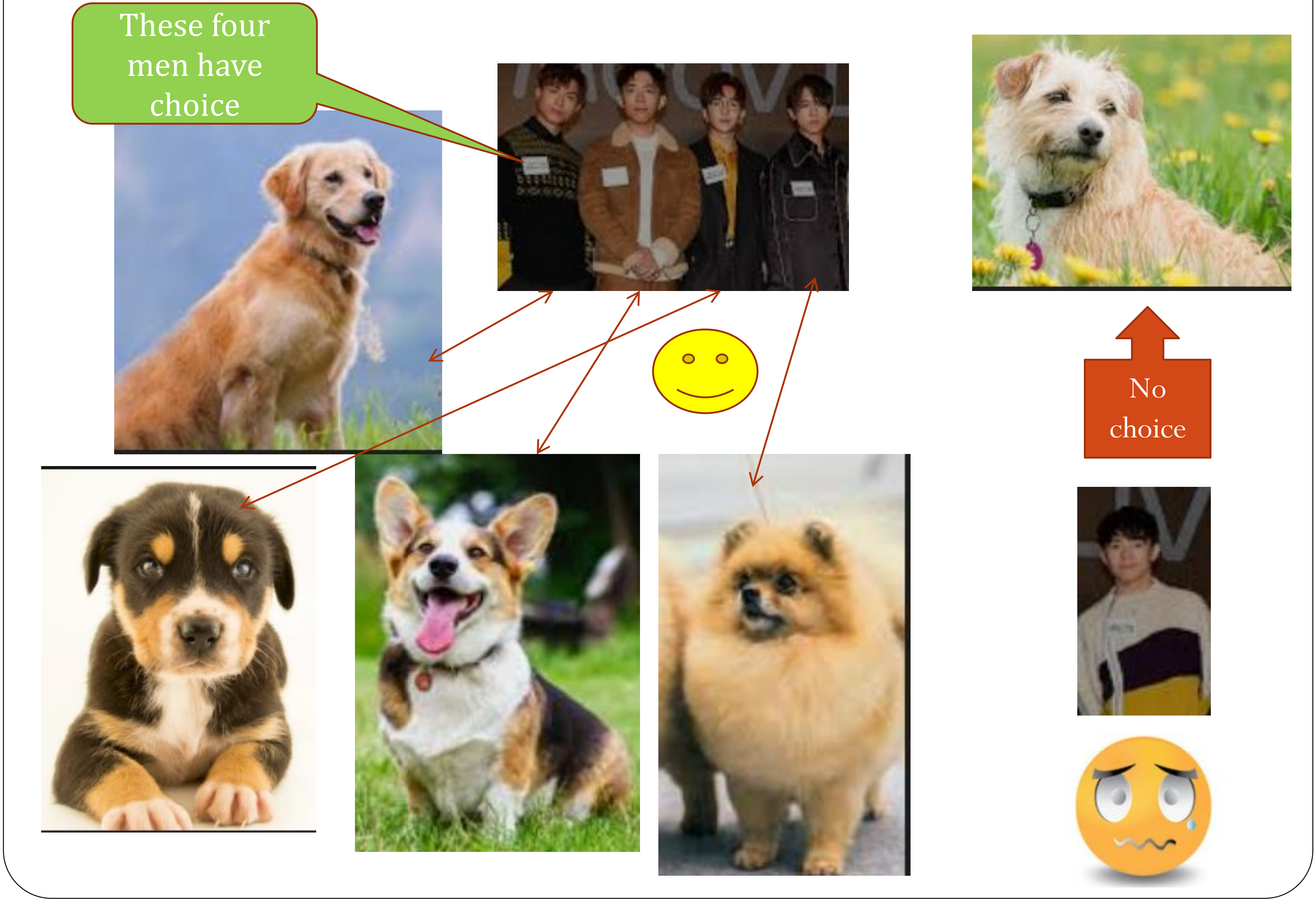
  通常 df=n-k。

其中n为样本数量，k为被限制的条件数或变量个数，或计算某一统计量时用到其它独立统计量的个数

OR

df = (number of rows-1)(number of column-1)

Degree of freedom自由度(df)



These four men have choice

No choice

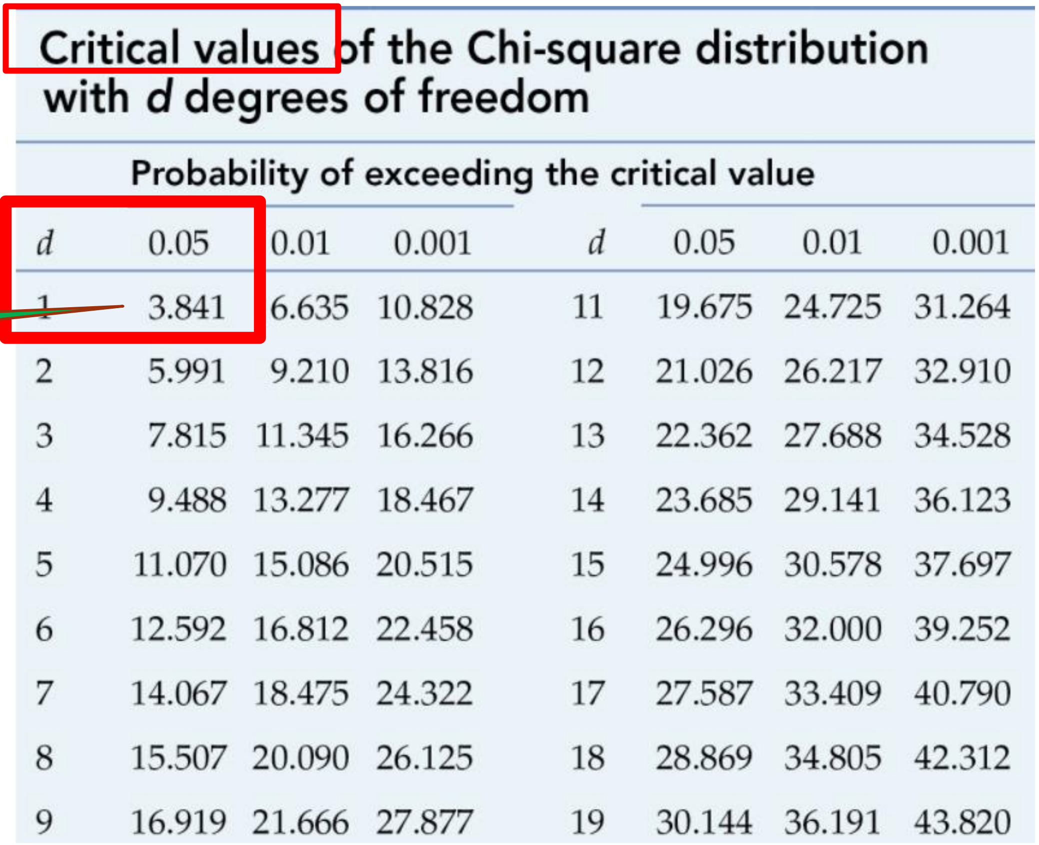
Calculate the degree of freedom 自由度

df=(number of rows -1)(number of columns-1)

df=(2-1)(2-1) = 1

	Dead	Alive
Drug A	a 30	b 40
Drug B	C 60	d 70

$X^2 = 0.198$



Critical values of the Chi-square distribution with *d* degrees of freedom

	Probability of exceeding the critical value						
d	0.05	0.01	0.001	*d*	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820

X^2 < critical values

0.198 < 3.841

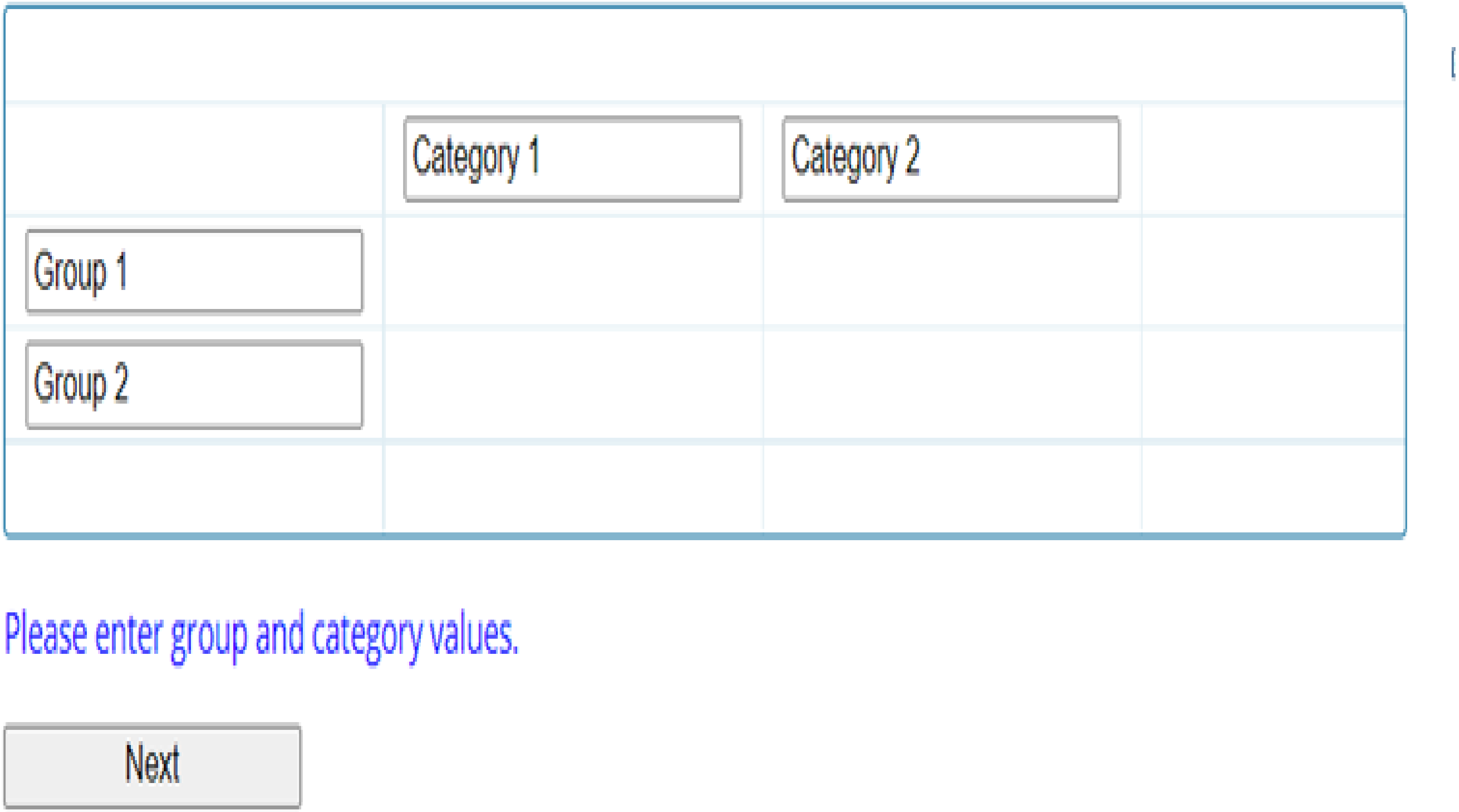
Do **NOT** reject

Null Hypothesis

不要拒绝零假设

https://www.socscistatistics.com/tests/chisquare/default2.aspx

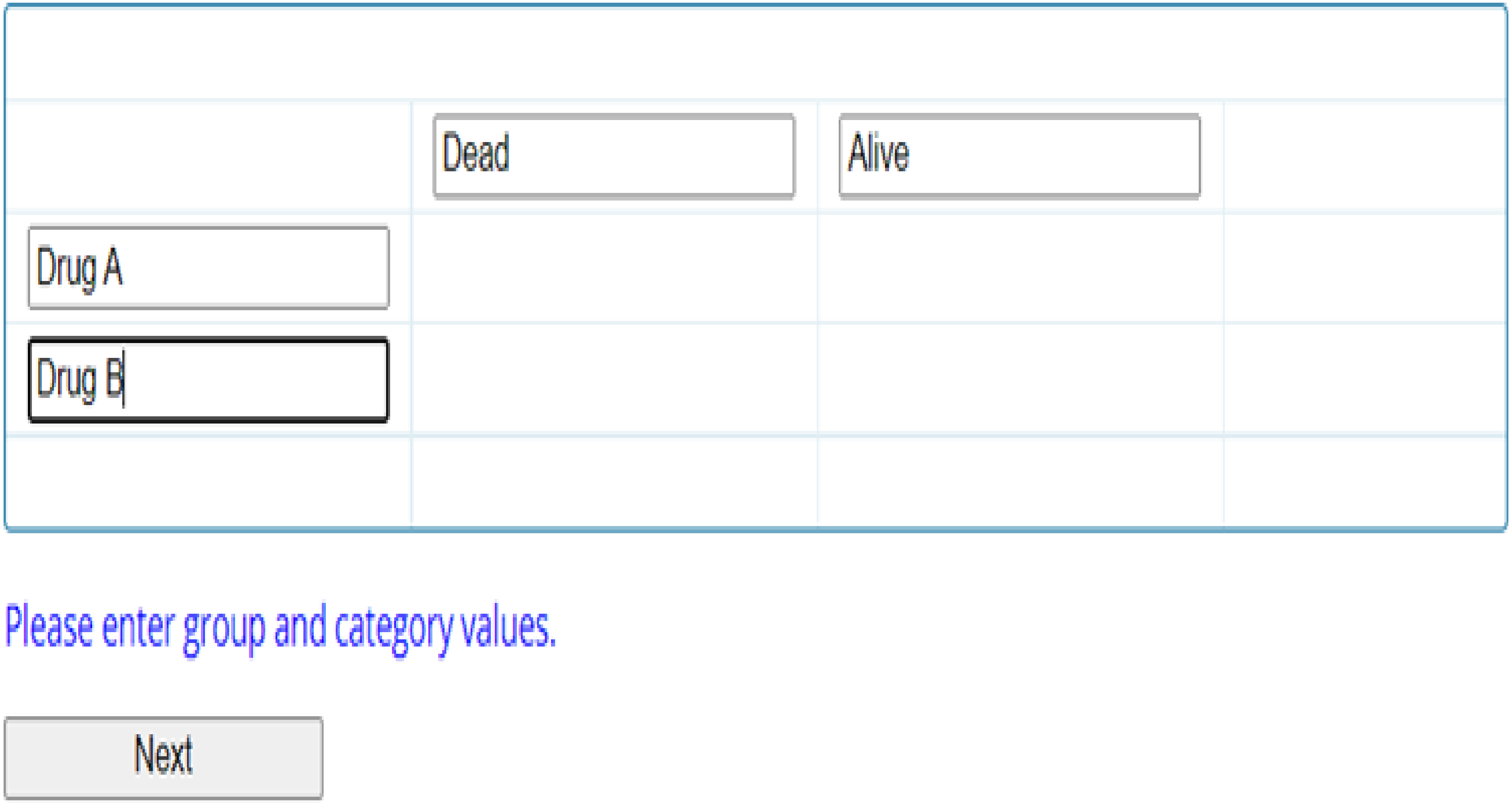
Use Chi Square Calculator



	Category 1	Category 2	
Group 1			
Group 2			

Please enter group and category values.

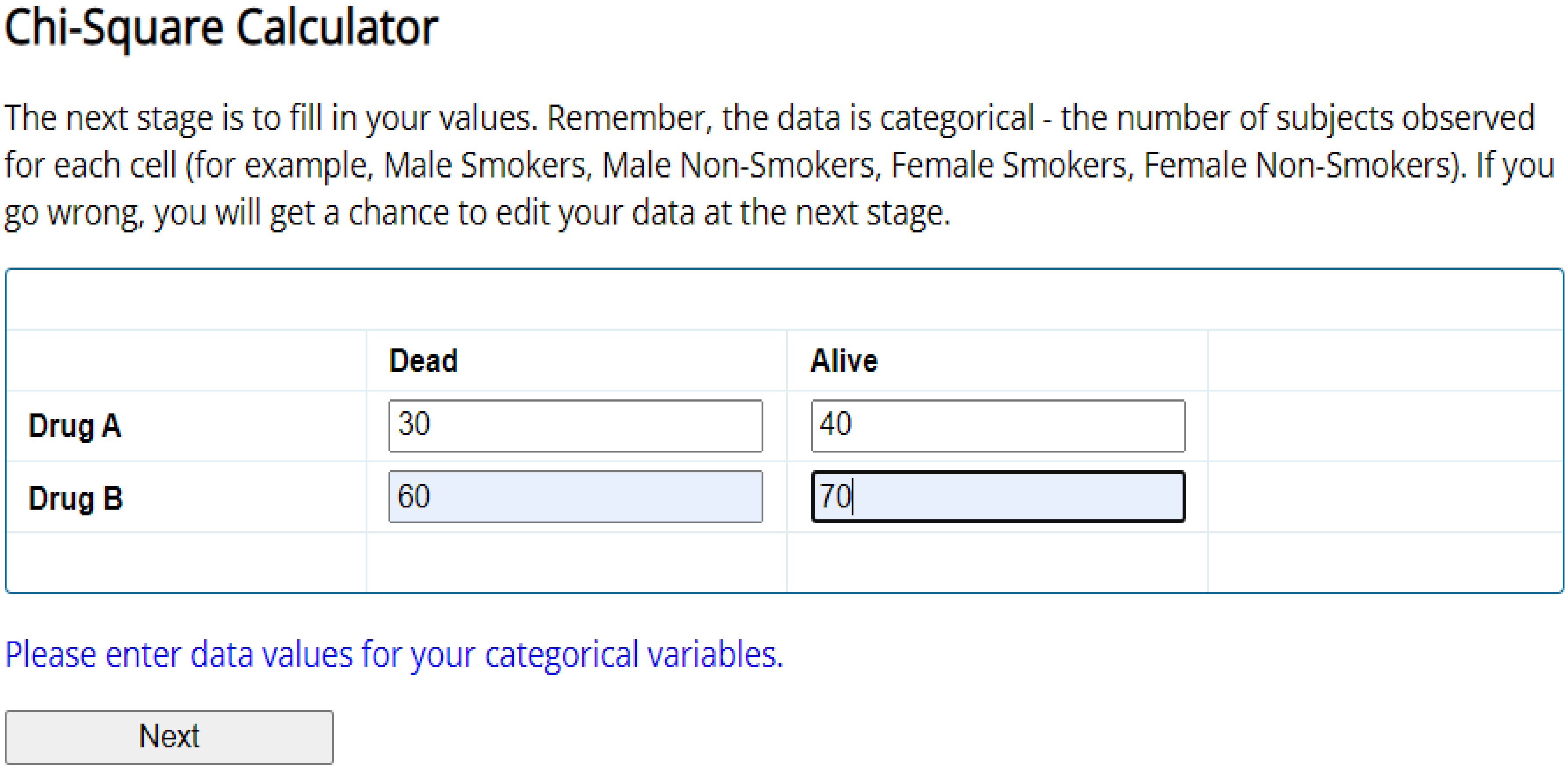
Step 1



	Dead	Alive	
Drug A			
Drug B			

Please enter group and category values.

Step 2



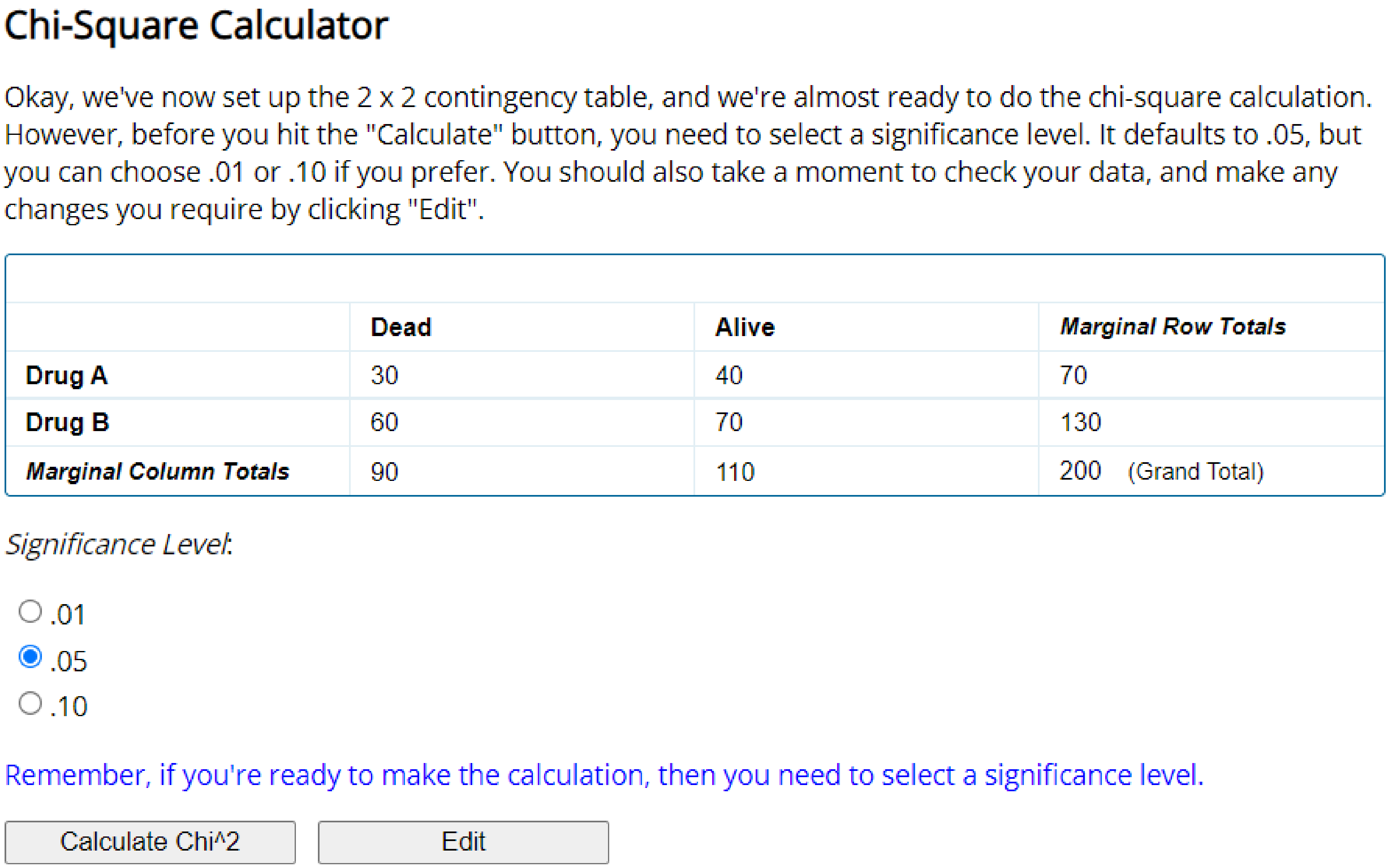
Chi-Square Calculator

The next stage is to fill in your values. Remember, the data is categorical - the number of subjects observed for each cell (for example, Male Smokers, Male Non-Smokers, Female Smokers, Female Non-Smokers). If you go wrong, you will get a chance to edit your data at the next stage.

	Dead	Alive	
Drug A	30	40	
Drug B	60	70	

Please enter data values for your categorical variables.

Step 3



Chi-Square Calculator

Okay, we've now set up the 2 x 2 contingency table, and we're almost ready to do the chi-square calculation. However, before you hit the "Calculate" button, you need to select a significance level. It defaults to .05, but you can choose .01 or .10 if you prefer. You should also take a moment to check your data, and make any changes you require by clicking "Edit".

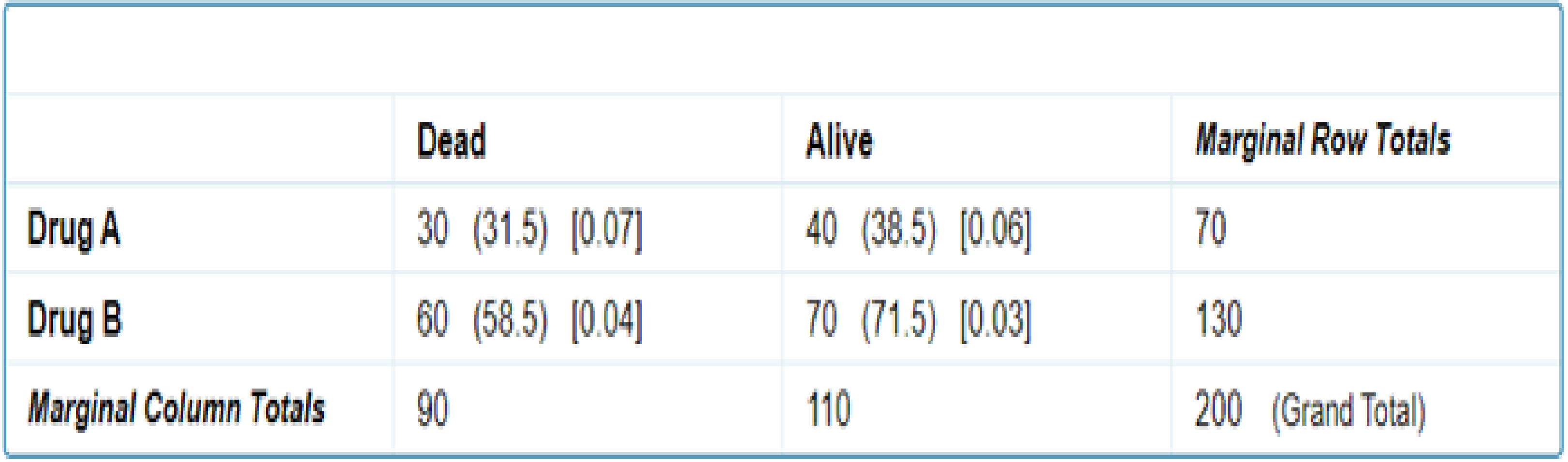
	Dead	**Alive**	***Marginal Row Totals***
Drug A	30	40	70
Drug B	60	70	130
Marginal Column Totals	90	110	200 (Grand Total)

Significance Level:

○ .01
● .05
○ .10

Remember, if you're ready to make the calculation, then you need to select a significance level.

Step 4



	Dead	Alive	*Marginal Row Totals*
Drug A	30 (31.5) [0.07]	40 (38.5) [0.06]	70
Drug B	60 (58.5) [0.04]	70 (71.5) [0.03]	130
Marginal Column Totals	90	110	200 (Grand Total)

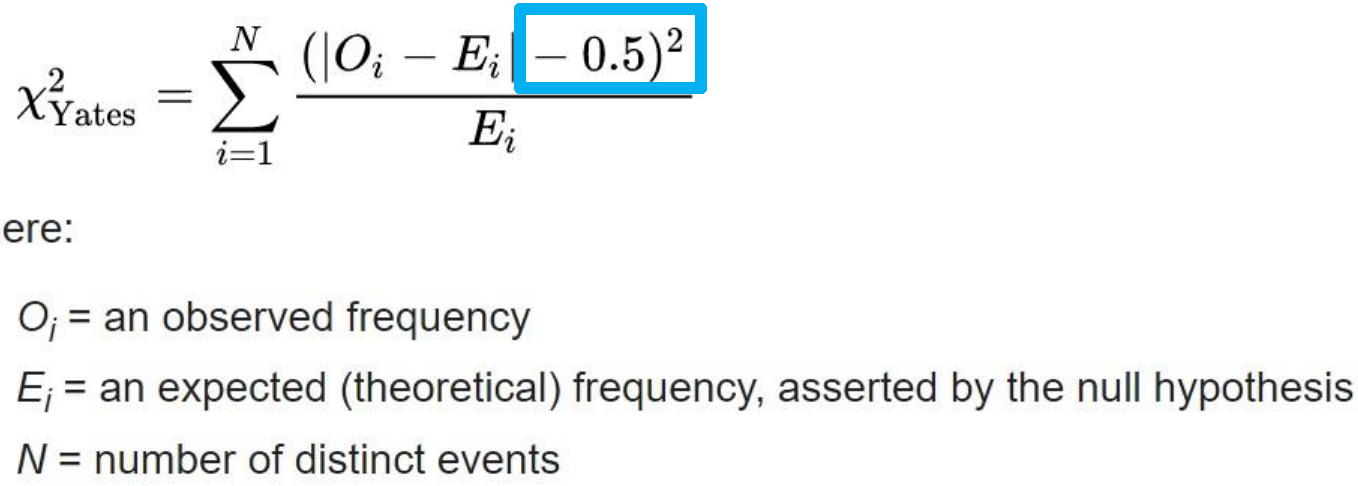
The chi-square statistic is 0.1998. The *p*-value is .654882. *Not* significant at $p < .05$.

The chi-square statistic with Yates correction is 0.0888. The *p*-value is .765708. *Not* significant at $p < .05$.

Yates correction 耶茨 修正

- Aims at correcting the error introduced by assuming that the discrete probabilities of observed binomial frequencies in the contingency table can be approximated by a continuous chi-squared distribution
- Also called the continuity correction for the chi-square test
- 卡方检验的连续性校正
- To adjust the observed frequency in each cell of a 2x2 table, Frank Yates suggested a correction by the following formula by subtracting 0.5 from the difference between each observed value and its expected value
- The correction is used only when there is one degree of freedom
- 修正仅在有一个自由度时使用

检定统计量：



$$\chi^2_{\text{Yates}} = \sum_{i=1}^{N} \frac{(|O_i - E_i| \boxed{- 0.5)^2}}{E_i}$$

where:

O_i = an observed frequency

E_i = an expected (theoretical) frequency, asserted by the null hypothesis

N = number of distinct events

X2 with Yates correction

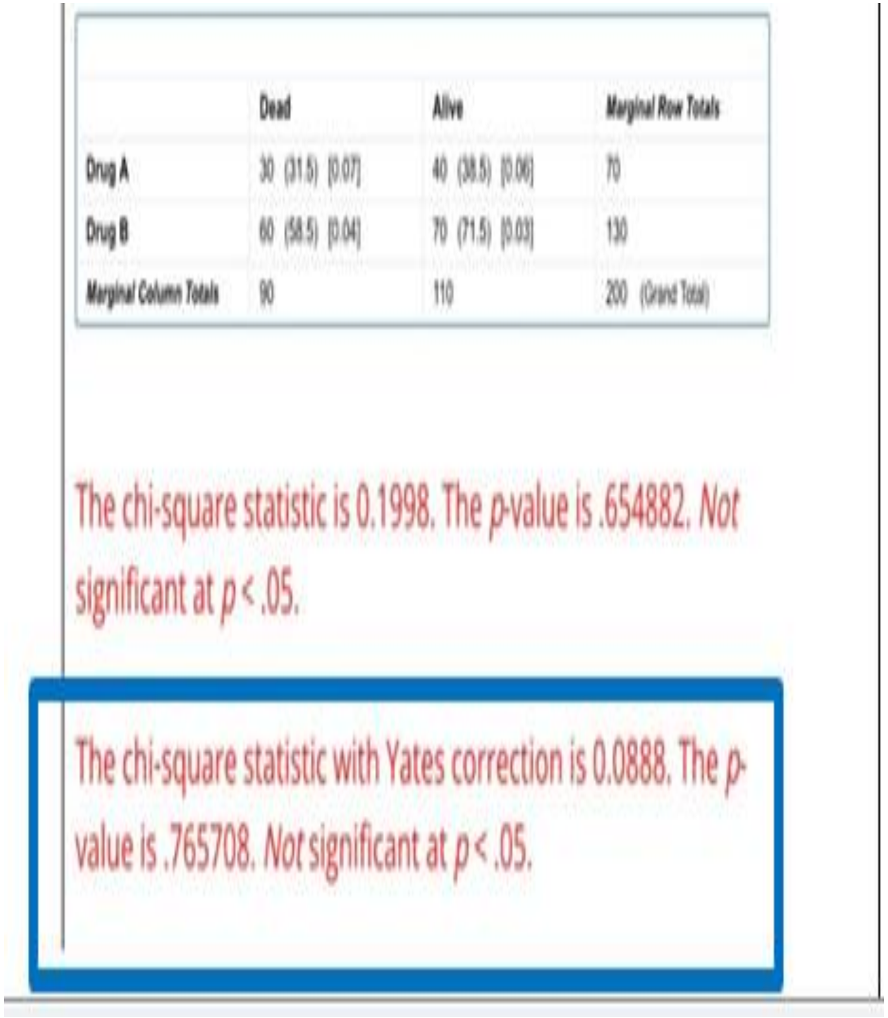
(use chi square calculator with Yates correction)

	Dead	Alive	
Drug A	a 30	b 40	70 (Row Total) a+b
Drug B	C 60	d 70	130 (Row Total) c+d
	90 (Column Total) a+c	110 (Column Total) b+d	200 Overall Total (a+b+c+d)

Use the same example

$X^2 = 0.198$

X^2 with Yates correction = 0.0888



	Dead	Alive	Marginal Row Totals
Drug A	30 (31.5) [0.07]	40 (38.5) [0.06]	70
Drug B	60 (58.5) [0.04]	70 (71.5) [0.03]	130
Marginal Column Totals	90	110	200 (Grand Total)

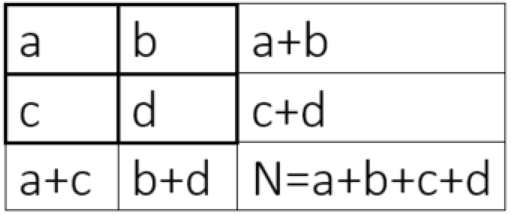
The chi-square statistic is 0.1998. The *p*-value is .654882. *Not* significant at *p* < .05.

The chi-square statistic with Yates correction is 0.0888. The *p*-value is .765708. *Not* significant at *p* < .05.

Fisher's Exact Test 确切概率法

- Another statistical significance test used in the analysis of contingency tables
- Significance of the deviation from a null hypothesis (e.g. *p*-value) can be calculated
- Employed when sample sizes/expected frequencies are small

对列联表进行关联性检定时，其方格内(如下表之 a、b、c、d)样本大小 n < 5，费氏精确检定法比较精准

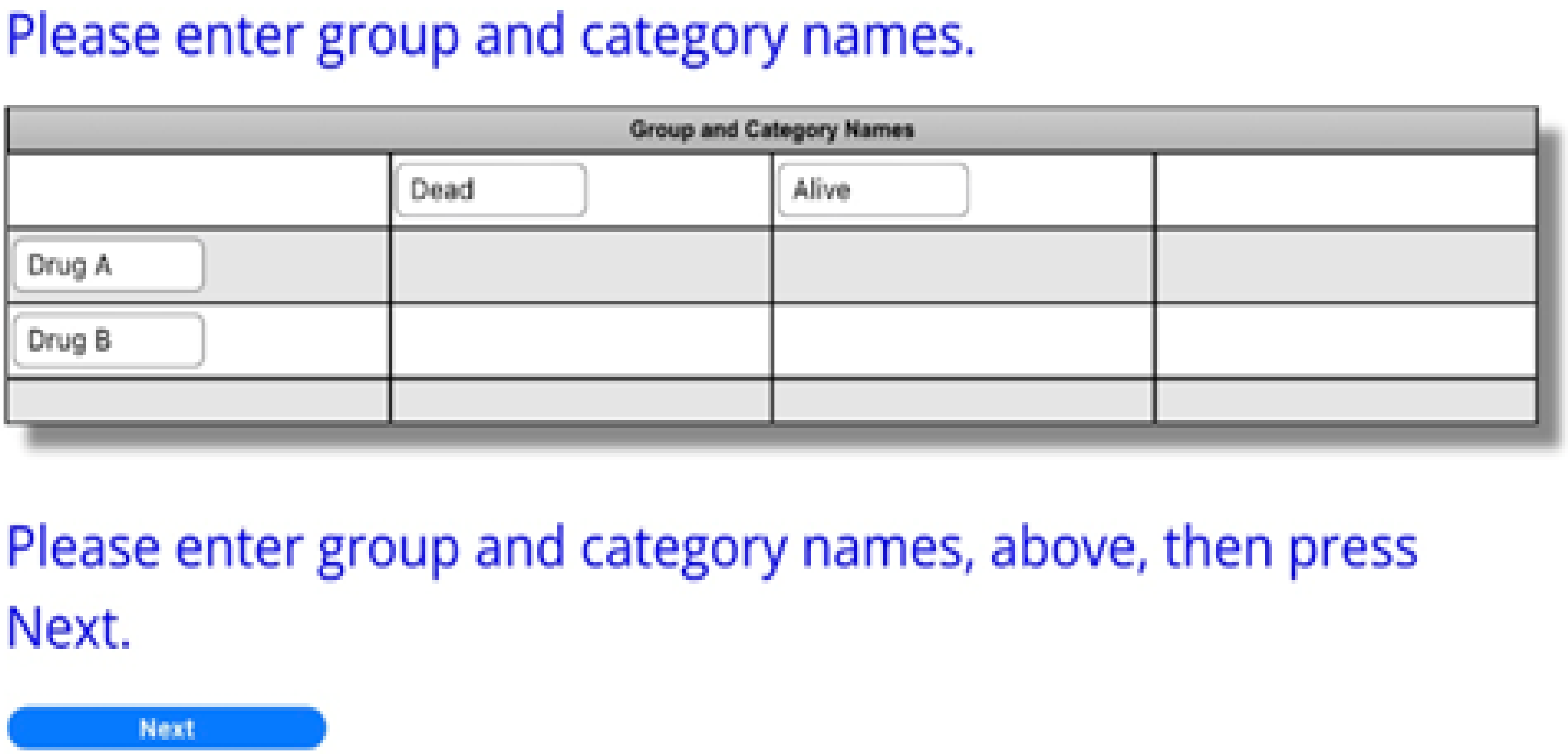


a	b	a+b
c	d	c+d
a+c	b+d	N=a+b+c+d

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{N}{a+c}}$$

Fisher’s Exact Test 确切概率法

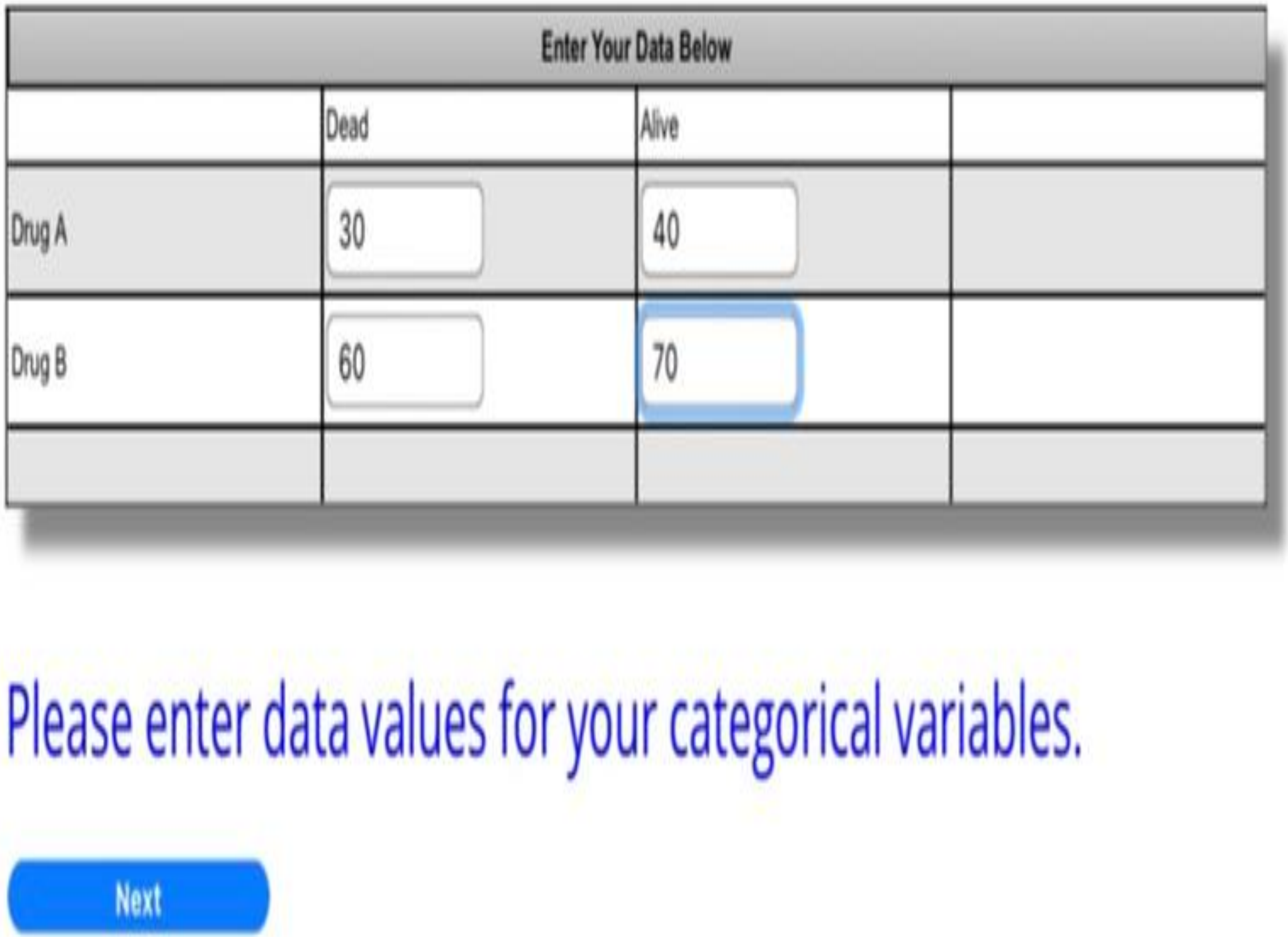
(use Fisher’s Exact Test calculator)



Please enter group and category names.

Group and Category Names			
	Dead	Alive	
Drug A			
Drug B			

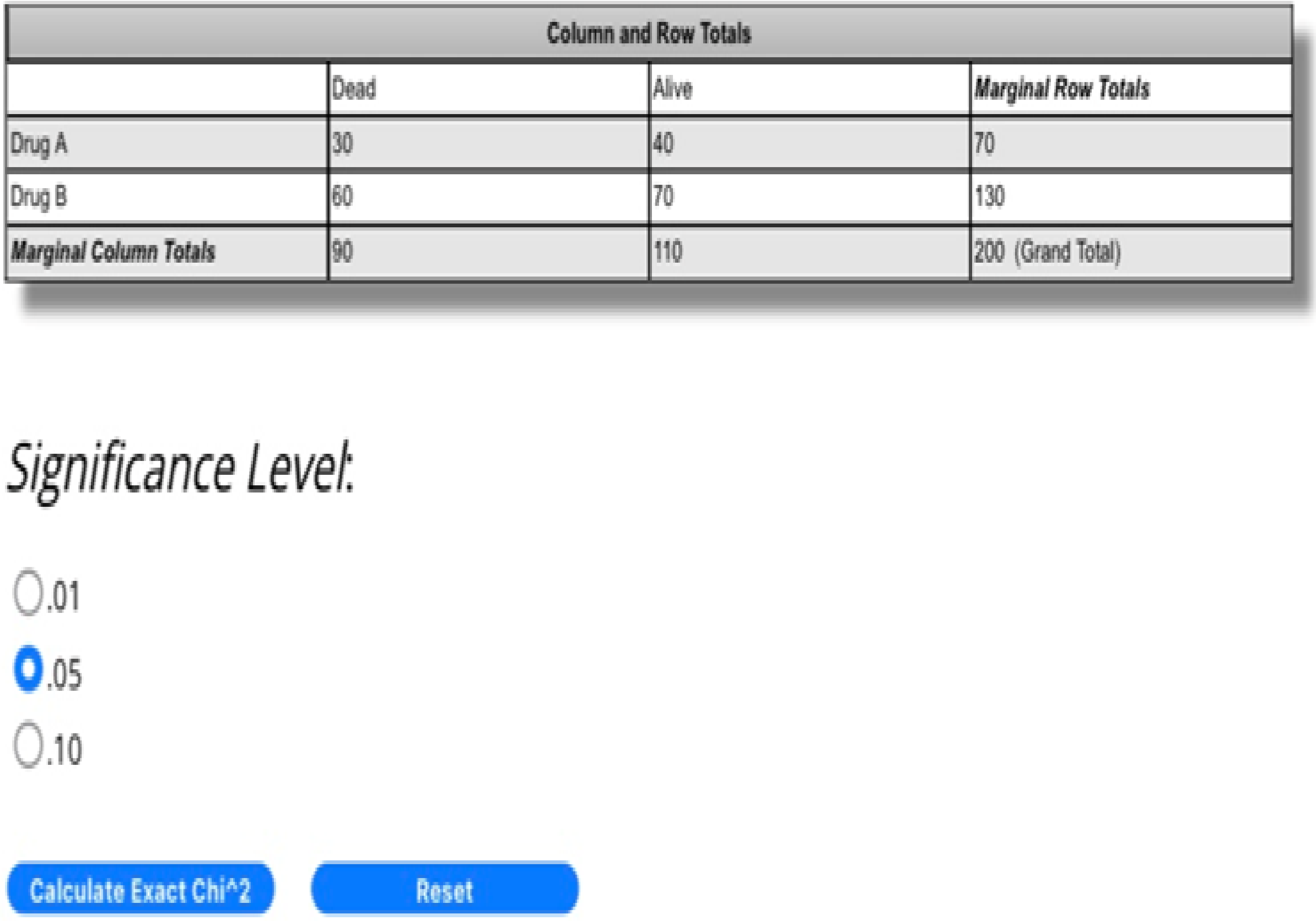
Please enter group and category names, above, then press Next.



Enter Your Data Below

	Dead	Alive	
Drug A	30	40	
Drug B	60	70	

Please enter data values for your categorical variables.



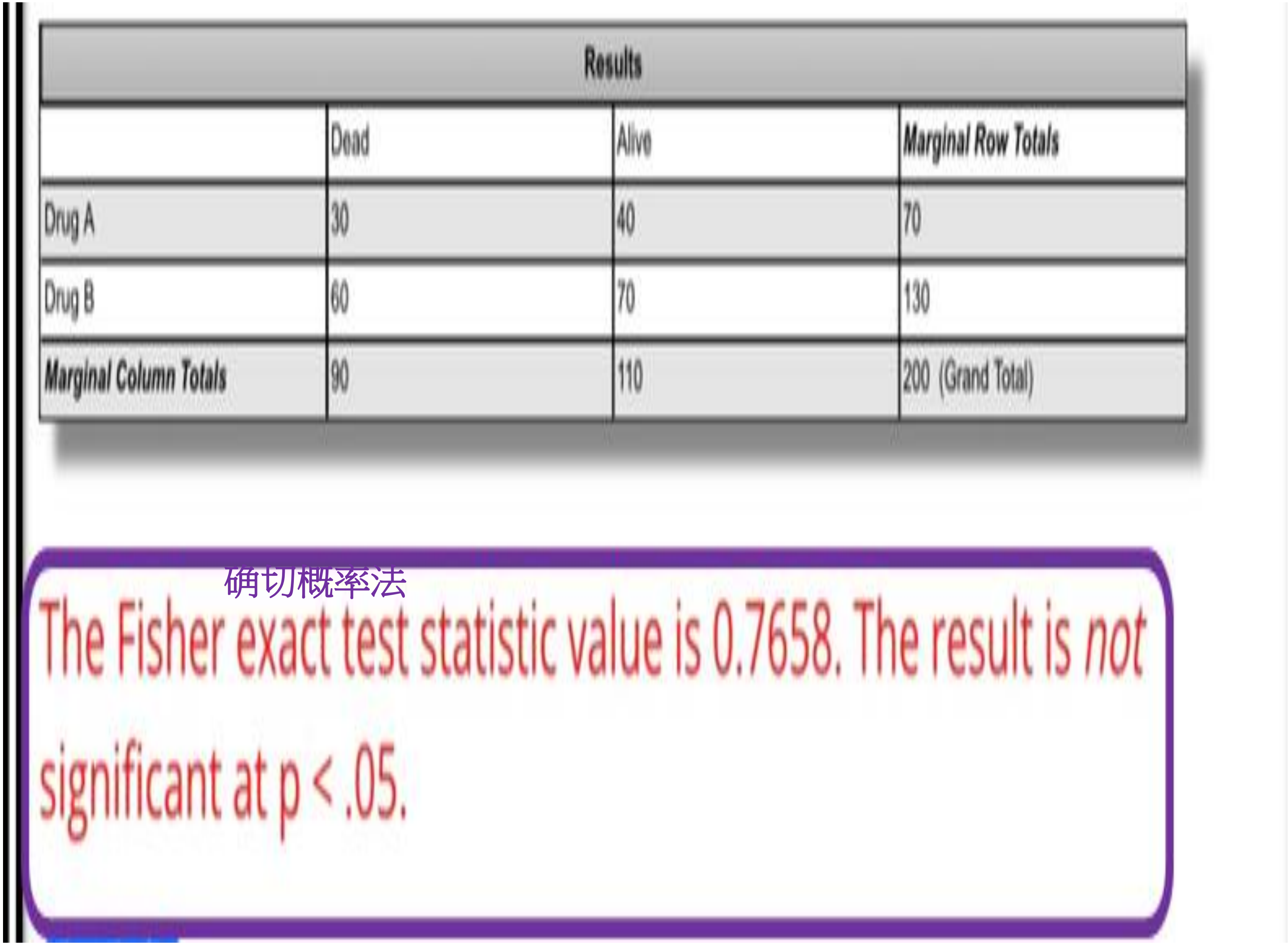
Column and Row Totals			
	Dead	Alive	***Marginal Row Totals***
Drug A	30	40	70
Drug B	60	70	130
Marginal Column Totals	90	110	200 (Grand Total)

Significance Level:

○ .01

● .05

○ .10

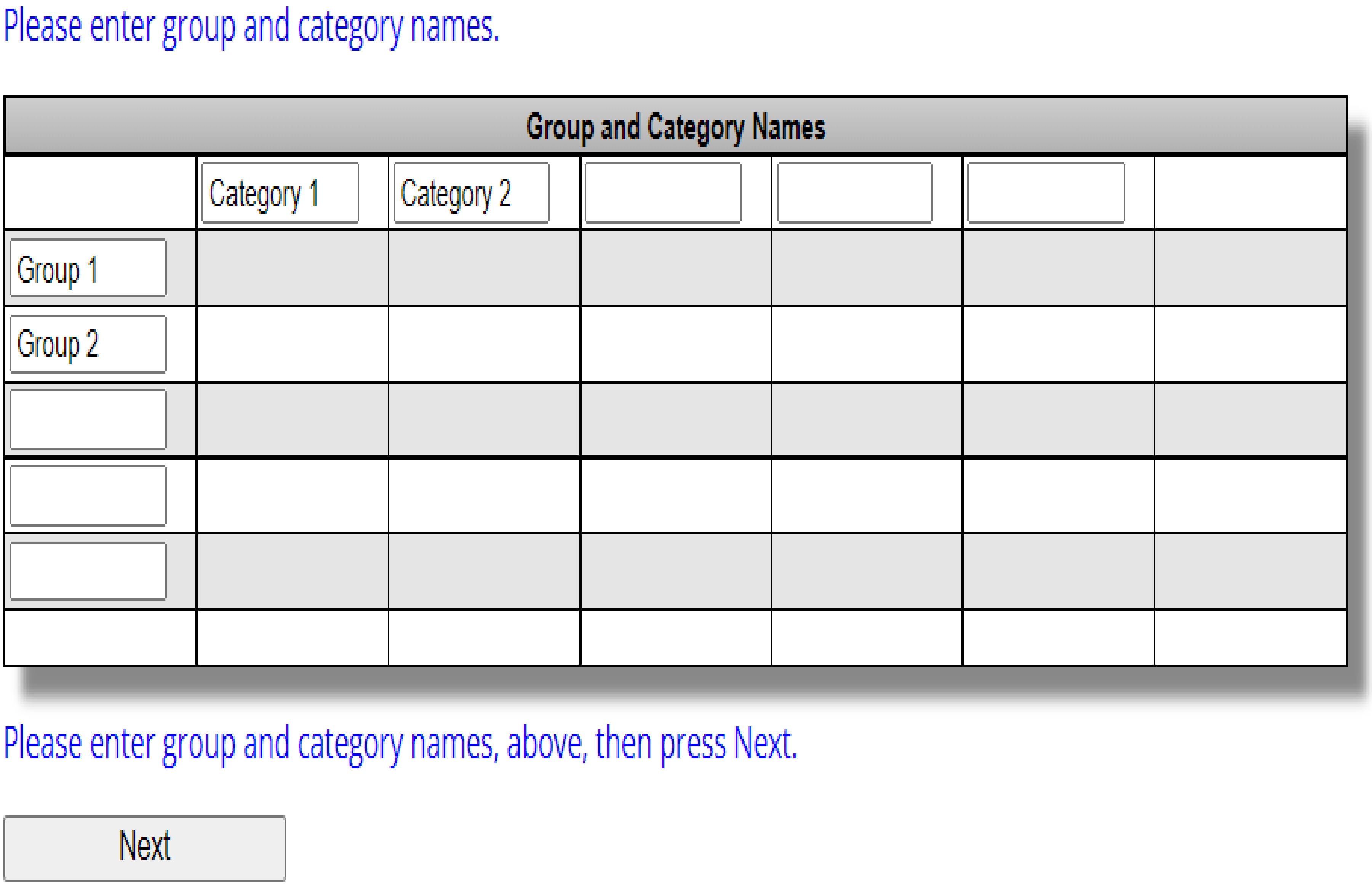


Results			
	Dead	Alive	*Marginal Row Totals*
Drug A	30	40	70
Drug B	60	70	130
Marginal Column Totals	90	110	200 (Grand Total)

确切概率法

The Fisher exact test statistic value is 0.7658. The result is *not* significant at $p < .05$.

r by c chi-square test 卡方分割（Calculator）



Please enter group and category names.

Group and Category Names						
	Category 1	Category 2				
Group 1						
Group 2						

Please enter group and category names, above, then press Next.

Thank You!!!