Comparing More Two Population Proportions

Example – To compare the market share of iPhone between three Countries (A, B, C), a market research company collects a random sample of people in each of these countries (A, B and C) and records the **proportion of people using iPhone.**

The results are shown below.

- In the sample of people selected from Country A,
  there is a total of 300 people and 70% of them use iPhone.
- In the sample of people selected from Country B,
  there is a total of 400 people and 65% of them use iPhone.
- In the sample of people selected from Country C,
  there is a total of 500 people and 50% of them use iPhone.

From the results, the proportion of people using iPhone **varies** a lot among the three countries(A, B, C). However, it is **not** our concern.

In the **random samples of people** from the three countries (A,B,C)

A: 70% use iPhone

B: 65% use iPhone

C: 50% use iPhone

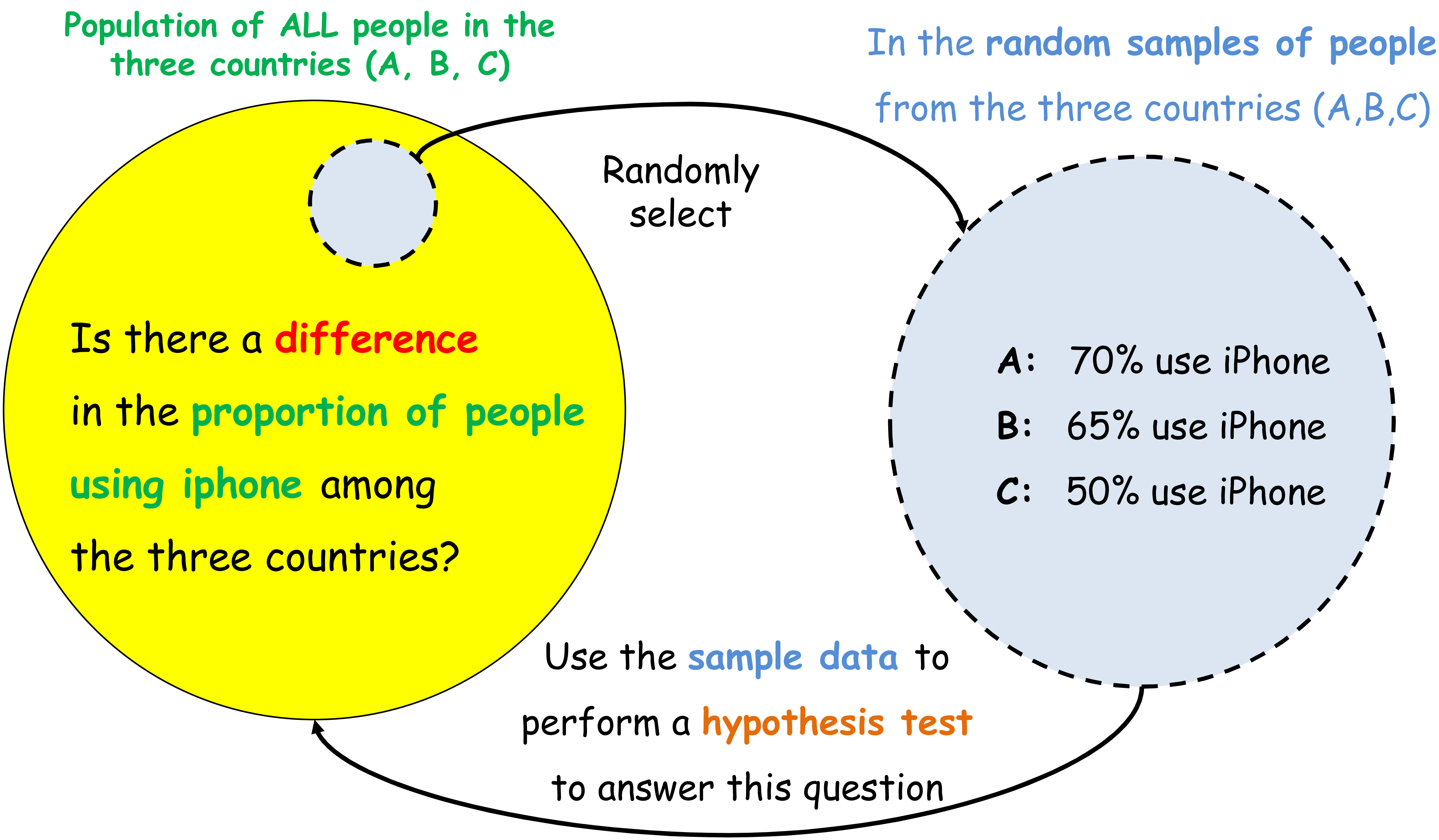
Question - In the **population of ALL people in the three countries (A, B, C)**, is there a **difference** in the proportion of people using iphone among the three countries?

Population of ALL people in the three countries (A, B, C),

Is there a **difference** in the **proportion of people using iphone** among the three countries?

We don't know the answer to this question because we do not have the **population data**.

Question - In the **population of ALL people in Country A, B nnd C**, is there a **difference** in the proportion of people using iphone among the three countries?



Question - In the **population of ALL people in the three countries (A,B,C)**, is there a **difference** in the proportion of people using iphone among the three countries?

Answer: We can perform multiple ***z***-tests to compare the proportion of people using iphone **between any two of countries such as**,

- A and B
- A and C
- B and C

But what is the problem?

- Whenever we perform a statistical hypothesis test,
- it is possible to make a **mistake**.
- Let's say that we perform each ***Z***-test at the 5% significance level.
- Since we reject Ho if the p-value is less than 5%,
- there is a **5% chance** of **making a mistake** that we **reject Ho in fact Ho is true**.
- When we perform ***Z***-tests **three** times ,
- the probability of **making a mistake** in at least one of the tests
- is about **3** * **5%** = 15%
- Therefore, a more efficient approach to test all proportions simultaneously is to perform a **Chi-Square Test.**

Step 1- State the Hypotheses

- The **hypotheses** are the statements about a **population**.
- There are two hypotheses we need to state at the beginning called
  - **Null Hypothesis** (denoted by **Ho**)
  - **Alternative Hypothesis** (denoted by **Ha**)

Null Hypothesis

The **null Hypothesis (Ho)** is the hypothesis that says there is **NO difference** between **population proportions**.

In other words, all **population proportions** are assumed to be **SAME** under the **null hypothesis**.

In this example,

Ho:	The proportion of ALL people using iPhone is the **SAME** in the three countries(A, B and C).

Alternative Hypothesis

- The **alternative Hypothesis (Ha)** is the hypothesis that says **at least one population proportion <u>differ</u>** from the others.

In our example, the alternative hypothesis is

Ha:	**At least one** country (either A, B or C) has **a different** proportion of all people using iPhone.

Note:

Under the **alternative hypothesis**, there are many possibilities;

For example.

1. The proportion of all people using iPhone is the **SAME** for Country A and B,

   but both **differ** from Country C.

2. The proportion of all people using iPhone in the three countries is **all different**

Before we move on to the next step, we need to make a **two-way table** that summarize the information from the samples.

	Country A	**Country B**	**Country C**
# of people **using iPhone**	70% of 300 = 210		
# of people **NOT using iPhone**			
Total			

Step 2 – Calculate the Expected Numbers assuming the Null Hypothesis is True

- When we perform the Chi-Square test, first,
- we assume that the null hypothesis is true that
- the **proportion of ALL people** using iPhone is the SAME in the three countries(A, B and C).
- Under this assumption, we combine the samples from the three countries
- Then use the combined samples to estimate
  - the proportion of all people using iPhone
  - the proportion of all people not using iPhone

	Country A	Country B	Country C	Total	Proportion
# of people using iPhone	210	260	250	720	
# of people NOT using iPhone	90	140	250	480	
Total	300	400	500	1200	

Notes: These percentages are only **correct** if the **null hypothesis is true** that the **proportion of ALL people** using iPhone is the *SAME* in the three countries

- Here is the important question we want to ask.
- Recall, we sample **300 people from Country A**

	Country A	Country B	Country C
Total	300	400	500

Question: If **300 people** are randomly selected from **Country A**,

- what are the expected numbers of people
  - using iPhone and
  - not using iPhone
- assuming the null hypothesis is true that the the **proportion of ALL people** using iPhone is the SAME in the three countries.

	Estimated Proportion	Expected number of people
Using iPhone	60%	$300 \times 60\% = 180$
Not using iPhone	40%	$300 \times 40\% = 120$

If **300 people** are randomly selected from **Country A**, we expect

- 180 of them use iPhone
- 120 of them do not use iPhone

assuming the null hypothesis is true that the the **proportion of ALL people** using iPhone is the SAME in the three countries

- Recall, we sample **400 people from Country B**

	Country A	**Country B**	**Country C**
Total	300	400	500

Question: If **400 people** are randomly selected from **Country B**,

- what are the **expected numbers** of people
  - **using iPhone** and
  - **not using iPhone**
- assuming the **null hypothesis is true** that the the **proportion of ALL people using iPhone** is the **SAME** in the three countries.

	Estimated Proportion	Expected number of people
Using iPhone	60%	$400 \times 60\% = 240$
Not using iPhone	40%	$400 \times 40\% = 160$

If **400 people** are randomly selected from **Country B**, we expect

- 240 of them use iPhone
- 160 of them do not use iPhone

assuming the null hypothesis is true that the the **proportion of ALL people** using iPhone is the SAME in the three countries

- Recall, we sample **500 people from Country C**

	Country A	**Country B**	**Country C**
Total	300	400	500

Question: If **500 people** are randomly selected from **Country C**,

- what are the expected numbers of people
  - using iPhone and
  - not using iPhone
- assuming the null hypothesis is true that the the **proportion of ALL people** using iPhone is the SAME in the three countries.

	Estimated Proportion	**Expected number** of people
Using iPhone	**60**%	
Not using iPhone	**40**%	

If 500 **people** are randomly selected from **Country C**, we **expect**

- ________ **of them use iPhone**
- ________ **of them do not use iPhone**

assuming the **null hypothesis is true** that the the **proportion of ALL people using iPhone** is the **SAME** in the three countries

Now we put all the expected numbers in the corresponding cells of the two-way table.

Expected Numbers	Country A	Country B	Country C
# of people using iPhone	180	240	300
# of people NOT using iPhone	120	160	200

Next, we need to compare

with the expected data we calculated based on the null hypothesis is true.

Actual Numbers	Country A	Country B	Country C
Using iPhone	210	260	250
NOT using iPhone	90	140	250

VS

Expected Numbers	Country A	Country B	Country C
Using iPhone	180	240	300
NOT using iPhone	120	160	200

To do that, we need the **Chi-Square Statistic**

Step 3 - Calculate the Chi-Square Statistic

The **Chi-Square Statistics** measures the **difference** between **Actual data** and **Expected data** assuming **the null hypothesis is true**.

Chi Square Statistic (χ^2) is defined as

$$\chi^2 = \sum_{all\ i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

χ^2	–	Notation of Chi-Square Statistic
O_{ij}	–	the **Actual numbers (or frequencies)** in the i^{th} row and j^{th} column of the two-way table
E_{ij}	–	the **Expected numbers (or frequencies)** in the i^{th} row and j^{th} column of the two-way table
$\sum$	–	sum over all i, j entries in the two-way table

The first step is to compare the actual number and the expected number in each cell of the two-way table. How?

1. We calculate **difference** between the actual number and the expected number
2. Then we **square** the **difference**
3. Finally, we **divide** the **squared-difference** by the expected number.

$$\frac{\left(\begin{matrix}\text{Actual} \\ \text{Number}\end{matrix} - \begin{matrix}\text{Expected} \\ \text{Number}\end{matrix}\right)^2}{\begin{matrix}\text{Expected} \\ \text{Number}\end{matrix}}$$

I usually call this term as **"Chi-Square Contribution"** that tells you how much contribution to the **Chi-Square statistic** from a particular entry.

Actual Numbers	Country A
Using iPhone	210

VS

Expected Numbers	Country A
Using iPhone	180

Let's look at the cell in the 1st row and 1st column

- the expected number of people using iPhone (if Ho is true) is 180
- the actual number of people using iPhone (in the sample) is 210

Chi-square contribution due to the cell in the 1st row and 1st column is:

$$\frac{\left(\begin{array}{c}Actual\\Number\end{array} - \begin{array}{c}Expected\\Number\end{array}\right)^2}{\begin{array}{c}Expected\\Number\end{array}} = \frac{(210 - 180)^2}{180} = 5$$

Actual Numbers	Country A	Country B	Country C
Using iPhone	210	260	250
NOT using iPhone	90	140	250

VS

Expected Numbers	Country A	Country B	Country C
Using iPhone	180	240	300
NOT using iPhone	120	160	200

Chi-Square Contribution	Country A	Country B	Country C
Using iPhone	$\frac{(210-180)^2}{180} = 5$	$\frac{(260-240)^2}{240} = 1.67$	$\frac{(250-300)^2}{300} = 8.33$
NOT using iPhone	$\frac{(90-120)^2}{120} = 7.5$	$\frac{(140-160)^2}{160} = 2.5$	$\frac{(250-200)^2}{200} = 12.5$

Chi-Square Contribution	Country A	Country B	Country C
Using iPhone	$\frac{(210-180)^2}{180} = 5$	$\frac{(260-240)^2}{240} = 1.67$	$\frac{(250-300)^2}{300} = 8.33$
NOT using iPhone	$\frac{(90-120)^2}{120} = 7.5$	$\frac{(140-160)^2}{160} = 2.5$	$\frac{(250-200)^2}{200} = 12.5$

Chi–Square Statistic (χ^2)

$$= 5 + 1.67 + 8.33 + 7.5 + 2.5 + 12.5$$

$$= 37.5$$

A Faster Way to Calculate the Chi-Square Statistic

- From the pervious calculation, we estimate the proportion of all people using iPhone is 0.6 under Ho.

	Country A	Country B	Country C	Total	Proportion
# of people using iPhone	210	260	250	720	$\frac{720}{1200} \times 100\% \rightarrow 60\%$
# of people NOT using iPhone	90	140	250	480	$\frac{480}{1200} \times 100\% \rightarrow 40\%$
Total	300	400	500	1200	

We can compare the proportion of people using iPhone observed in each country to the proportion under Ho by calculating the square of z-score.

$$z^2 = \left(\frac{\bar{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \right)^2 = \frac{n(\bar{p} - p_0)^2}{p_0(1-p_0)}$$

where

p_0 is the estimated / expected proportion under Ho.

$\bar{p}$ is the proportion we observed in the sample / group

n is the sample size in the sample / group

Let's process with the calculation

Country	Number of iPhone users	n	Observed Proportion	z^2
A	210	300	0.7	
B	260	400	0.65	
C	250	500	0.5	
Combine d	720	1200	0.6	

Estimated proportion under Ho

Chi-Square Statistic = 12.5 + 4.1667 + 20.8333 = 37.5

Properties of Chi-Square Statistic

What can we conclude if the Chi-Square Statistic is ZERO?

- The Chi-Square Statistic will be ZERO
- if the actual data are identical to the expected data.
- Recall, the expected data are calculated based the **null hypothesis is true** that
- the **proportion of ALL people** using iPhone is the SAME in the three countries (A, B and C).
- Therefore, we conclude that the **proportion of ALL people** using iPhone is the SAME in the three countrie.

Properties of Chi-Square Statistic

What can we conclude if the Chi-Square Statistic (χ^2) is a **"SMALL" number**?

- The Chi-Square Statistic will be a **"SMALL"** number
- if the **actual data slightly differ** the **expected data**
- although we **cannot** conclude that the **null hypothesis is true** that
- the **proportion of ALL people using iPhone** is the **SAME** in the three countries
- at least we **DO NOT reject** the **null hypothesis**.

Properties of Chi-Square Statistic

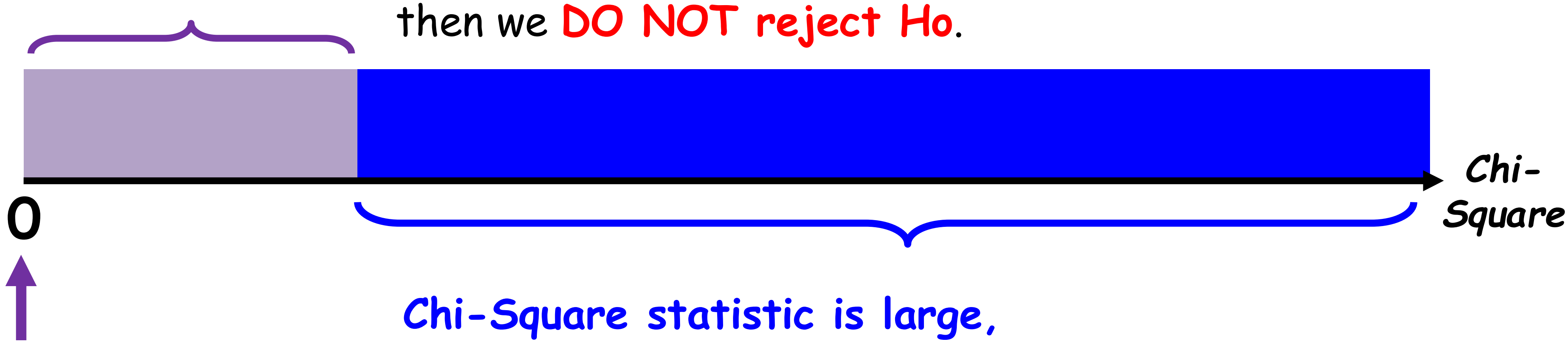
What can we conclude if the Chi-Square Statistic (χ^2) is a **"LARGE" number**?

- The Chi-Square Statistic will be a **"LARGE" number**
- if the actual data **differ a lot from** the expected data
- Recall, the expected data are calculated **assuming the null hypothesis is true**.
- When the actual data **differ a lot from** the expected data, it indicates that the null hypothesis may be NOT TRUE.
- Therefore, we should **reject null hypothesis** and
- support the **alternative hypothesis** that **at least one population proportion differ from the others**.

In summary

Chi-Square statistic is small (i.e. close to zero)

then we **DO NOT reject Ho**.



Chi-Square = 0

All population proportions are the same

Chi-Square statistic is large,

then we support the **alternative hypothesis (Ha)** that

at least one population proportion <u>differ</u> from the others.

Question: How can we decide whether

the **Chi-Square statistic** is **small** or **large**?

Answer: We can covert the **Chi-Square statistic** to its **p-value.**

Step 4 - Find the p-value and State the Conclusion

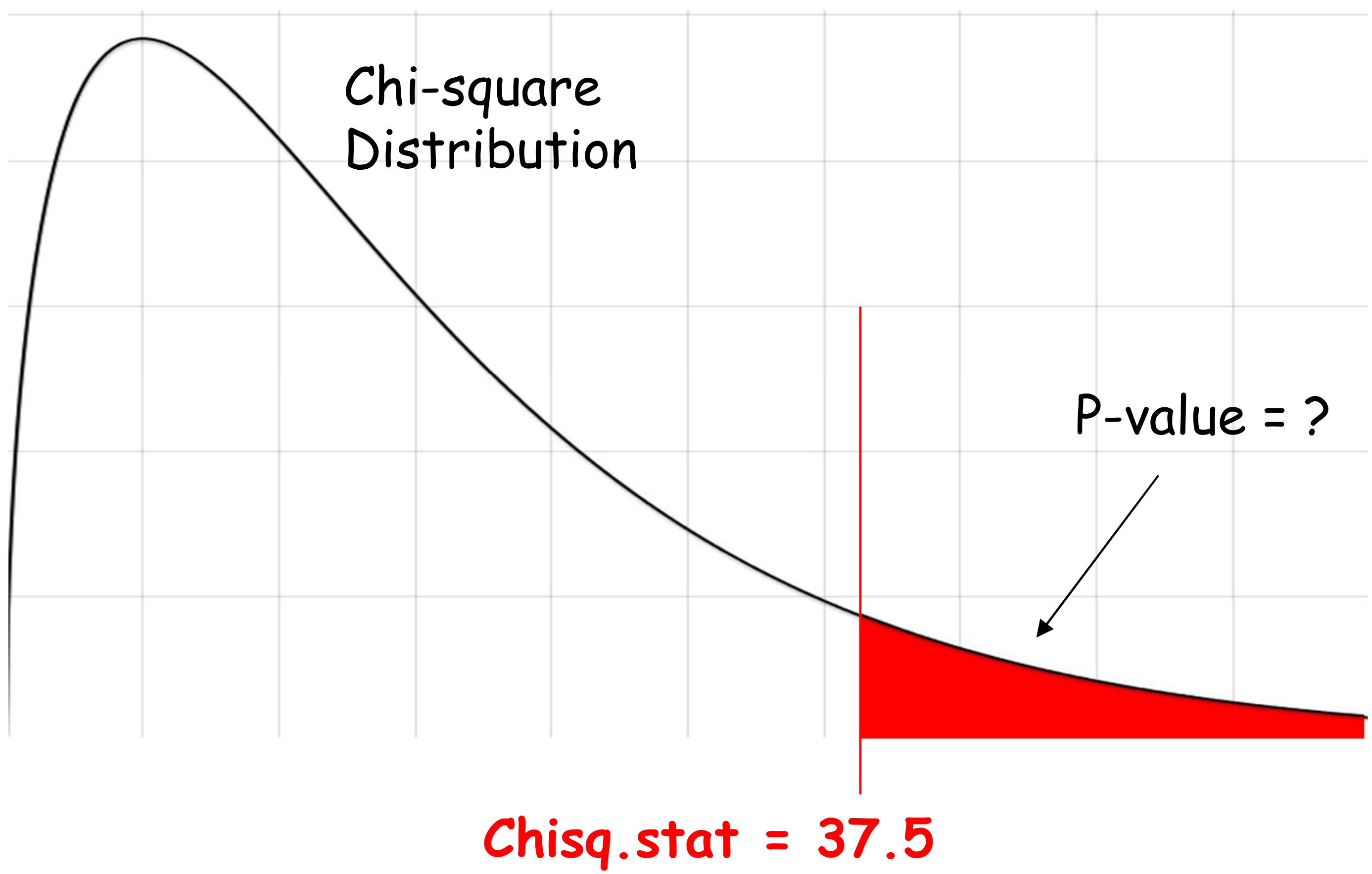
The **p-value** of the **Chi-Square statistic** is the **area** under the Chi-Square Distribution.

The center and spread of the Chi-Square distribution are controlled by a parameter called the **degree of freedom (DF)**

For comparing multiple proportions, **DF = # number of samples / groups – 1**

In the case, we compare 3 proportions -> **DF = 3 – 1 = 2**

The **p-value** is the **area of the right tail** bounded by **Chi-Square statistic (e.g. 37.5)** .



- To calculate the area under the chi-square distribution, we use statistical software such as R.
- We will use R to perform the chi-square test.
- Please see the attached R code.
- The following section presents the output.

```
> #Run the Chi-Square Test
> chisq.test( two.way.table )

        Pearson's Chi-squared test

data:  two.way.table
X-squared = 37.5, df = 2, p-value = 7.194e-09
```

Chi-square statistic = 37.5

P-value = $\underbrace{0.00000000}_{9\,zeros}7194$

Let's compare the p-value with the 5% significance level.

Since the p-value (0.000000007194) is **smaller than** the significance level (0.05), we **reject Ho**.

At the 5% significance level, the sample data **provide sufficient evidence** to conclude that **at least one** country (either A, B or C) has **a different** proportion of all people using iPhone.

Conditions Required for a Valid Chi-Square Test

- Similarly, **NOT** all datasets can be used to perform the Chi-square test **unless it satisfies certain conditions**.
- If the dataset **does not** satisfy the required **conditions,** the conclusion drawn from the Chi-Square will **NOT be correct**.
- What is the **condition** required for the Chi-Square Test?

Mainly, it requires the expected numbers (or frequencies) in all cells of the two-table are at least 5.

Let's look at the expected data in our example.

Expected Numbers	Country A	Country B	Country C
Using iPhone	180 ≥ 5	240 ≥ 5	300 ≥ 5
NOT using iPhone	120 ≥ 5	160 ≥ 5	200 ≥ 5

Since the expected numbers in all cell is at least 5, the conclusion drawn from the Chi-Square test should be correct and can be trusted.

Warning

- In the Chi-Square test, we can only determine whether
- **at least one proportion is different from the one**.
- But we are **NOT able** to tell:

1. which proportion is **different** from the other proportions
2. which proportion is **higher** or **lower** than the other proportions

Follow-up the results from Chi-Square Test

- From the Chi-Square test, we found that **at least one country** gives a **different proportion of people using iPhone**
- Which country has the **largest** **proportion of people using iPhone?**
- To answer the question, we can use a **confidence interval**
- to compare the **proportion of people using iPhone**
  - between Country A and Country B
  - between Country A and Country C
  - between Country B and Country C
- Question: What **confidence level** should we use?

- Let's say that we use a **95% confidence interval** to **compare two proportions**.
- There is a **5%** probability of getting ***incorrect results*** that
- the interval **does not contain** the **true value of a parameter**
- When we construct **three** 95% confidence intervals to compare **three pairs of proportions** (A vs B, A vs C, B vs C)
- The **overall error rate**, that is the probability of getting **incorrect results** in at least one of the **95% confidence intervals** is about **3** * **5%** = 15%
- How can we reduce the **overall error rate?**

- Recall, the confidence interval for a difference between two population proportions is given by the following:

Proportion for i^{th} sample

Proportion for j^{th} sample

$$\underbrace{(\bar{p}_i - \bar{p}_j)}_{\text{Difference between two sample proportions}} \pm \underbrace{\underbrace{z_c}_{\text{Z-Critical Value}} \times \sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n_i} + \frac{\bar{p}_j(1-\bar{p}_j)}{n_j}}}_{\text{Margin of Error}}$$

Sample Size for i^{th} / j^{th} sample

- To reduce the **error rate**, we can increase the margin of error, thereby making each confidence interval wider.
- How? We replace the **z-critical value** (z_c) with a larger critical value,
- **the square root of the chi-square critical value.**
- This procedure is known as **Marascuilo Procedure**

- Here is the formula to calculate the confidence interval for the difference between two proportions using the **Marascuilo Procedure**

Proportion for i^{th} sample

Proportion for j^{th} sample

$$\underbrace{(\bar{p}_i - \bar{p}_j)}_{\text{Difference between two sample proportions}} \pm \underbrace{\sqrt{\underbrace{\chi_c^2}_{\text{Chi-Square Critical Value}}}\sqrt{\frac{\bar{p}_i(1-\bar{p}_i)}{n_i} + \frac{\bar{p}_j(1-\bar{p}_j)}{n_j}}}_{\text{Margin of Error}}$$

Sample Size for i^{th} / j^{th} sample

- To find the Chi-Square critical value, first we need to set the overall error rate.
- This represents the probability of getting incorrect results in at least one of the intervals
- Let's the overall error rate to be 5%

  (it is should as the significance level used in the Chi-Square test).
- Then, we need to read the Chi-Square Table.
- To read the Chi-Square table, we need to know the degree of freedom (DF)
- Degrees of freedom (DF) = Number of proportions being compared − 1

  = 3 - 1 = 2

Significance level = 0.05 →

$\alpha = 0.05$	
DF	Critical Value
1	3.841
2	5.991
3	7.815
4	9.488

It is your exercise to complete the calculation of the confidence intervals.

	A	B	C
Proportion of using iPhone	0.7	0.65	0.5
Sample size (n)	300	400	500
χ_c^2	5.991		

CI for the difference in the **proportion of people using iphone between Countries A and B**

$$(0.7-0.65)\pm\sqrt{5.991}\sqrt{\frac{0.7(1-0.7)}{300}+\frac{0.65(1-0.65)}{400}}=(-0.0372, 0.1372)$$

CI for the difference in the **proportion of people using iphone between Countries A and C**

CI for the difference in the **proportion of people using iphone between Countries B and C**

- From the intervals, we can conclude that
  - The **proportion of all people using iPhone** in Country A/B is **higher** Country C.
  - But we **cannot** conclude that which **country A or B** has a **higher** **proportion of all people using iPhone**.


THANK
YOU