chi- Square Test

Chi Square tests are multiple sample tests used when dealing with count or categorical data

eg- Is there a difference in the number of people responding Yes/No to a direct mail offer, based on income categories?

→ The dependent variable or the target variable is a frequency count

→ There are two common applications of chi Square Tests

(i) Tests of association

(ii) Goodness of fit

e.g. Using Categorical or Tabular Data

As a retailer you look at brand ROI (Return of Investment) to assess shelf space effectiveness. Looking at particular category, carbonated beverages, you know across all your stores the shares of wallet for top Brands A, B and all other C is as listed in the first table

Brand	Transaction share
A	52%
B	35%
C (All other)	13%

You take a random sample of data from a particular store - 300 purchases of carbonated beverages

Brand	# of Transations	%
A	177	59
B	78	26
C	45	15

Before you start on any analysis, you first need to check if this difference implies this store is not like the population

The idea is to check the difference between what you expected in your sample, and then assess the chances of seeing that difference purely by chance

If there was no difference between this store and all other stores, what would we expect to see as the # of transactions for Brands A, B and all other (c)?

Column 1	Brand A	Brand B	Brand C
observed	177	78	45
Expected	156	105	39

→ A chi square test uses these observed and expected frequencies, to generate a conclusion about the statistical significance of the observed differences

Mathematically, the quantity

$$\sum \frac{(f_o - f_e)^2}{f_e}$$

follows a chi square distribution, with $k-1$ degrees of freedom

$f_o \Rightarrow$ observed frequency
$f_e \Rightarrow$ expected frequency

$k \Rightarrow$ No of samples / No of cells.

→ A chi square distribution is an asymmetric distribution that depends only on sample size

→ It is generated as the square of std scores (z) from a normal distribution

→ As sample size increases, chi square tends to normal

The Chi-Square statistic is built as:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

For our example, chi-square Test statistics:

$$\frac{(177-156)^2}{156} + \frac{(78-105)^2}{105} + \frac{(45-39)^2}{45} = 10.69$$

To use a Table : we need df

Df = No of cells - 1 = 3 - 1 = 2

We will use chi-Square table to find Critical stat value.

In excel,

= CHITEST (Actual_range, expected_range)

Chi Square (Association Tests)

Q8.

You look at preferences for beverages by age to understand if there is an association between age and brand preference, in order to decide if you need differentiated marketing strategies by age

You do a survey on a random sample, and get the following results:

Observed · Preference

Brand	M15-25	M26-40	M41-55	Total
Coke	49	50	69	168
Pepsi	24	36	38	98
Sprite	19	22	28	69
Total	92	108	135	335

Expected Value –

Calculate the expected values under the assumptions that the Null Hypothesis is TRUE

NULL HYPOTHESIS : It is there No association between Brand preference and Age

$$\text{Expected Value} = \frac{(\text{Row Total} \times \text{Column Total})}{n}$$

Expected Brand	Preference M 15-25	M 26-40	M 41-55	Total
Coke	46.14	54.16	67.70	168
Pepsi	26.91	31.59	39.49	98
Sprite	18.95	22.24	27.81	69
Total	92	108	135	335

Chi Square (Goodness-of-Fit Tests)

Very popular use of Goodness-of-fit tests if the data follows a particular distribution or not (expected)

e.g.

A gambler is playing a new game in a casino, which involves rolling three dice at a time. Winnings are directly proportional to the number of 6's rolled

This is what we observed in 100 rolls of the dice

Number of 6's	Rolls
0	48
1	35
2	15
3	2

Would you have cause to believe that the gambler is maybe "too" lucky and is playing with loaded dice?

What distribution would you expect the outcome of seeing a 6 on rolled dice to follow? → Binomial

What should be the expected probabilities of number of 6's in three rolled dice?
- We will calculate using Binomial distribution formula

Number of 6's	Expected Prob	Expected Prob in 100 throws
0	0.5787	57.8704
1	0.34722	34.7222
2	0.06944	6.94444
3	0.00463	0.46296

Number of 6's	Observed	Expeted
0	48	57.8704
1	35	34.7222
2	15	6.9444
3	2	0.4629

Testing for a Normal Distribution or any type of distribution

→ Calculate expected distribution using the probability distribution formula

→ If checking for normality, for example, follow below steps:

1) Calculate mean and std deviation of your data

2) Bin the data into sub intervals

3) Calculate expected probability of those sub-intervals (using normal probability fn)

4) Compare that frequency observed in data

5) Construct chi square and Test

→ The chi Square reviewed so far have been non-parametric tests.

→ To apply these tests, we do not need the underlying population to follow any specific distribution

→ There are many kinds of non-parametric tests, an equivalent one for every parametric test.

→ Non parametric tests are "better" than parametric tests because you are not bound to have a data distribution of a particular type

→ Why use parametric tests then?

Non-parametric tests are less powerful than parametric tests in the sense that they use more information and are sometimes less flexible in terms of testing different kinds of hypothesis.

→ Also, as sample size increases, it turns out that non-parametric test distributions approximate normal distributions.

chi Square Parametric Test

→ This is a test of variance of sample tested against a population variance

→ The Central Limit Theorem states that the distribution of sample means will follow a normal distribution

→ What about variance of the samples?

⇒ only the means will follow Normal distribution

⇒ Variance of the samples will follow chi square distribution

→ The chi Square distribution can also be used for testing Variance and standard Deviation (So for all hypothesis testing we have reviewed were concerned with testing means)

→ Test Statistics

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

$n \Rightarrow$ No sample size

$s \Rightarrow$ sample std dev.

$\sigma \Rightarrow$ Population std dev

Q. A call center is experimenting with different approaches to improve customer experience, with the aim of consistant call resolution time.

- Currently average resolution time is 6.5 minutes, with a ~~variance~~ std dev of 4-5 minutes.
- A new approach has been tested resulting in an average resolution time of 6 minutes and a ~~variance~~ std dev of 3 minutes across 30 calls.

Is the new approach sufficiently different from the standard to justify investment in it?

- If our aim is consistency, we check if there is a significant reduction in variance of resolution time

H0 : Variance = 4.5 minutes
H1 : Variance < 4.5 minutes

$$\text{Chi Square Statistics} = \frac{(n-1)s^2}{\sigma^2}$$

$$= \frac{(30-1) \times 3^2}{4.5^2}$$

$$= 12.88$$

We could use table to compare calculated Test stat against a Critical Value.

OR

We can directly calculate p-value in Excel

=chisq.dist (~~12.88,29~~ (x, deg_freedom, cumulative)

=chisq.dist (12.88, 29, True)

= 0.00~~2~~4259

pvalue < alpha value.

Therefore, we will Reject the null and conclude Variance of calls has reduced.

* R Code :

```
## ANOVA

# One Way ANOVA

# Create a data set with 3 Groups

mpg = c(34, 35, 34.3, 35.5, 35.8, 35.3, 36.5,
        36.4, 37, 37.6, 33.3, 34, 34.7, 33, 34.9)

brand = factor(c("A", "A", "A", "A", "A", "B",
        "B", "B", "B", "B", "C", "C", "C", "C", "C"))

mileage = data.frame(mpg = mpg, brand = brand)

mileage
```

```
ANOVA <- aov (mpg ~ brand, data=
                              mileage)

summary (ANOVA)


# Two way ANOVA

  tw <- read.csv ( )

Res <- aov (satisfaction ~ Format + subject
            + format : subject, data = tw)

summary (Res)
```

Format	Subject	satisfaction
Online	stats	10
Online	stats	9
Online	stats	8
Online	Eng	7
Online	Eng	6
Online	Eng	5
Online	Science	4
Online	Science	3
Online	Science	2
Hybrid	stats	9
Hybrid	stats	8
Hybrid	stats	7
Hybrid	Eng	6
Hybrid	Eng	5
Hybrid	Eng	4
Hybrid	Science	3
Hybrid	Science	2
Hybrid	Science	1
offline	↓	8 to 0.
↓		↓

chi Square Test of Goodness-of-fit

Create a dataset with observations and expectations

```
obs = c(195, 165, 47, 15, 30, 35, 8, 5)

pct = c(0.374, 0.357, 0.085, 0.034, 0.066,
        0.063, 0.015, 0.006)

dat < data.frame(Blood_type = c("O+", "A+",
    "B+", "AB+", "O-", "A-", "B-", "AB-"),
    pct, obs, Exp = sum(obs) * pct)

chisq.test(x = dat$obs, p = dat$pct)
```

chi Square test of factor independence

```
Monthly <- c(91, 150, 109)

Occassionally <- c(90, 200, 198)

Never <- c(51, 155, 172)

dat <- data.frame(Monthly, Occassionally,
                  Never)

row.names(dat) <- c("under 45", "45-49" and
                    "over")

dat

chisq.test(dat)
```

Case study

* Chi Square Tests are used when dealing with Count data

Q. In the bank telemarketing dataset, the bank is trying to understand customer profiles. It wants to target some offers to customers on the basis of their education and job level, and therefore wants to understand if there is an association between level of education and job types

It takes a random sample of 100 customers and looks at their education levels and associated job types

Observed

Job Type	Primary	Secondary	Tertiary	
Blue collar	11	40	9	60
Management	3	5	17	25
Services	3	12	0	15
	17	57	26	100

Expected

	10.2	34.2	15.6
	4.25	14.25	6.5
	2.55	8.55	3.9

* ANOVA is used when we want to check impacts of discrete factors on a continuous outcome variable

Q. As a step of preliminary understanding in the bank telemarketing, the bank wants to understand if customers that already have loans have lower bank balances on average (and therefore are more unlikely to sign up for a long term deposit)

It takes a sample of customers from its database, and looks at average balance in the account for customers that have: housing loans (Yes/No) and personal loans (Yes/No)

1: Yes 2: No

Housing Loan	Personal Loan	Balance
1	1	39
1	1	282
1	1	137
1	1	152
1	1	101
1	1	723
1	2	54
1	2	90
1	2	91
1	2	81
1	2	144
1	2	351
1	2	839
2	1	582
2	1	61
2	1	173
2	1	283
2	1	1640
2	1	0
2	2	91
2	2	164
2	2	50
2	2	383
2	2	79
2	2	