Chi-Square Test

Chi square (χ^2) test are used with nominal data.

T-test and ANOVA can be used with continuous data only.

With nominal data, we usually calculate frequency of each group.
Chi-square evaluates whether the frequencies of categories actually observed(observed frequencies, f_0) match with the frequency you expect if randomly selected from null-hypothesis population (expected frequencies, f_e).
Chi-square test is also called a goodness-of-fit test.

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

Chi-square is the sum of the squared differences between observed and expected frequencies, each divided by the expected frequency.

For example: If we conduct an experiment in which we randomly sample 150 beer drinkers and let them taste the three leading brands. We want to know whether there is a difference in their preference for the three brands of light beer?
Preference for brands:

Brand A	Brand B	Brand C	Total
45	40	65	150

In this case, we will use Chi-square test as we are calculating frequency for each group.

Step1: Null and Alternate hypothesis
Null hypothesis(H0): There is no difference among beer drinkers in their preference for different brands of light beer.
Alternate hypothesis(H1): There is a difference among beer drinkers in their preference for different brands of light beer.

Step2: Expected frequencies under the H0
Assuming H0 is true, the expected frequencies (f_e) is the total number of participants divided by the number of categories.

Expected frequencies under H0

Brand A	Brand B	Brand C	Total	
45 (50)	40 (50)	65 (50)	150 (150)	Observed Expected

df = 3-1 = 2

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

$$= \frac{(45-50)^2}{50} + \frac{(40-50)^2}{50} + \frac{(65-50)^2}{50}$$

$$= 0.50 + 2 + 4.50$$

$$= 7.00$$

Critical value of χ^2 for degree of freedom(df) 2 is 5.991
Therefore, we can reject null hypothesis as here value of χ^2 is 7 which is greater than critical value.

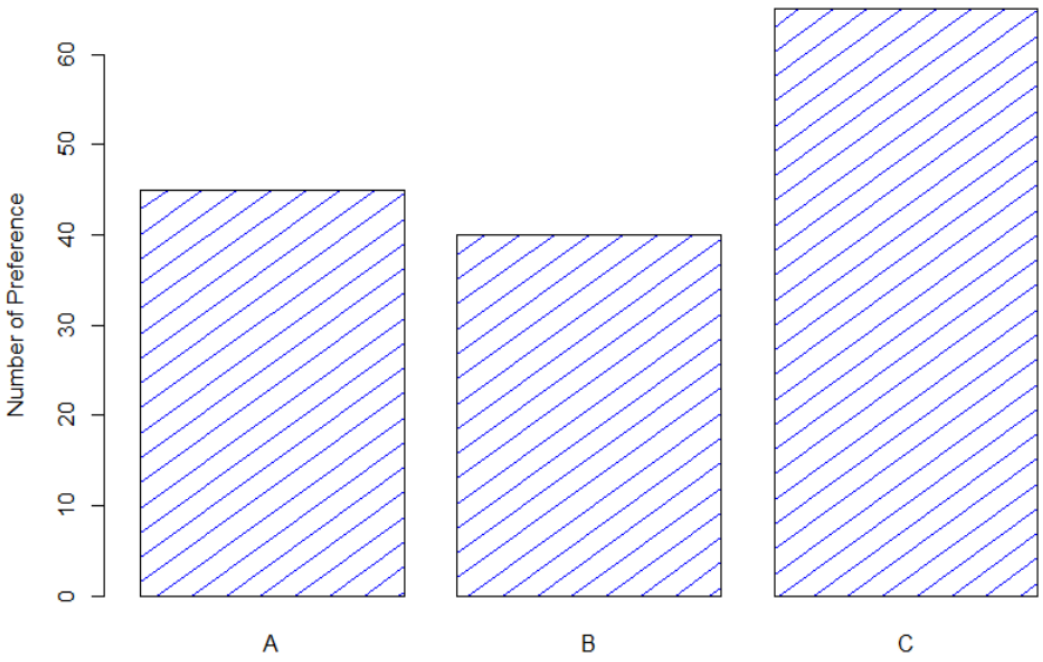
R code:

```r
observed <- c(45, 40, 65)
names(observed) <- c("A", "B", "C")
barplot(observed, density = 10, col = "blue", ylab =
          "Number of Preference")
result1 <- chisq.test(observed, correct = FALSE)
result1



# Here X-squared value is 7 and p value = 0.0302
# As p value is less than 0.05, we can reject null hypothesis
# and can say that there is a difference among beer drinkers in
# their preference for different brands of light beer.
```

```
> observed <- c(45, 40, 65)
> names(observed) <- c("A", "B", "C")
> barplot(observed, density = 10, col = "blue", ylab =
+           "Number of Preference")
> result1 <- chisq.test(observed, correct = FALSE)
> result1

        Chi-squared test for given probabilities

data:  observed
X-squared = 7, df = 2, p-value = 0.0302

>
```

Files | Plots | Packages | Help | Viewer

Frequency plot:



Chi- Square with two variables:

If we have two nominal variables, Chi-square test can be used to determine

whether they are related or it can also be used to test independence between two categorical variables.

For example:
Consider we randomly sample of 200 Republicans and 200 Democrats and ask them whether they are in favor of a particular bill or against the bill or undecided. Contingency table is:

Attitude →

	For	Undecided	Against	Row Marginal
Republican	68	22	110	200
Democrat	92	18	90	200
Column Marginal	160	40	200	400

Step1: Null and Alternate hypothesis
Null hypothesis (H0) - The attitude towards bill is not related to political party.
Alternate hypothesis (H1) - The attitude toward the bill is related to political party.

Step2: Expected frequencies under the NULL hypothesis

Here, f_e = (corresponding row marginal freq.* corresponding column marginal freq.)/ total elements.

For example, for cell 1: f_e = (160*200)/400

And degree of freedom, df = (r-1)(c-1)
where, r is number of categories in variable 1
c is number of categories in variable 2

And, $$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

R code:

```
contable1= matrix(c(68, 22, 110, 92, 18, 90), byrow = TRUE,
                  ncol = 3)
result2 <- chisq.test(contable1, correct = FALSE)
result2
```

```
>
> contable1= matrix(c(68, 22, 110, 92, 18, 90), byrow = TRUE,
+                   ncol = 3)
> result2 <- chisq.test(contable1, correct = FALSE)
> result2

        Pearson's Chi-squared test

data:  contable1
X-squared = 6, df = 2, p-value = 0.04979

>
```

Here, we can see p value is less than 0.05. Therefore, we can reject NULL value. Also, we can see from the contingency table, democrats are more likely to be in favor of the bill than republicans.

Conditions for Chi-square test to be valid:

1) Sample size needs to be large enough. Expected frequency in each cell is at least 5 for tables where r or c is 3 or greater. If table is 1X2 or 2X2 then each expected frequency should be at least 10.
2) Observations should be independent.

Problem with Chi-square test:

When the data contain few observations, the obtained chi-square does not follow a chi-square distribution. Therefore, using Chi-square distribution as reference distribution will lead to misleading inferences.
To solve this issue, we normally use Yate's correction or Fisher's exact test. It enumerates all possible outcomes given the marginal totals, and asks what percentage of them is more extreme than the result we obtained(this serves as an estimate of the p value).
It does not rely on any approximation to some theoretical distribution.
With large sample sizes, the Yates' correction makes little difference, and the chi-square test works very well. With small sample sizes, chi-square is not accurate,

with or without Yates' correction. Fisher's exact test, as its name implies, always gives an exact P value and works fine with small sample sizes. Normally, Fisher's test is preferred over Yate's test.

R code:

Yate's correction:
result <- chisq.test(contable1, correct = TRUE)
print(result)

Fisher's correction:
Fisher_test <- fisher.test(contable1)
print(Fisher_test)

On applying Fisher's test instead of chi-squared test in previous example.

```
contable1= matrix(c(68, 22, 110, 92, 18, 90), byrow = TRUE,
                  ncol = 3)
result2 <- chisq.test(contable1, correct = FALSE)
result2


# On using Fisher's exact test instead of Chi-square test

Fisher_test <- fisher.test(contable1)
print(Fisher_test)
```

```
>
>
> Fisher_test <- fisher.test(contable1)
> print(Fisher_test)

        Fisher's Exact Test for Count Data

data:  contable1
p-value = 0.04959
alternative hypothesis: two.sided

>
```

For R code:
Visit [Github]