# Data Science 1
# Analysis of Categorical Data
# Part 3

**Ann Maharaj**

# Analysis of Categorical Data: Part 3

1. Analysis of Categorical Data: Related Populations
2. McNemar Test for Equal Proportions: Two Populations
   - Example
3. Cochran Q for Equal Proportions: More than 2 Populations
   - Example

# Analysis of Categorical Data: Related Populations

- In the previous two topics we analysed categorical data for independent populations.
- In what follows, we analyse categorical data for related populations.

# McNemar Test for Equal Proportions: Two Populations

- We use the chi-square test of independence when categorical data are collected from two independent populations.
- The McNemar test is used for analysing categorical data from two related populations.
- This test is relevant for testing the null hypothesis that the proportion of items or individuals with the characteristic of interest is the same under two conditions or treatments.

# McNemar Test for Equal Proportions: Two Populations

- The data is presented in a 2 x 2 table but the cell frequencies are numbers of ‘pairs‘ not numbers of item or individuals.
- For example, in a before and after design each item or individual is measured twice for the presence or absence of the characteristic: before and after an intervention.
- The ‘pairs‘ are not two paired items or individuals but two measurements on the same item or individual.

# McNemar Test for Equal Proportions: Two Populations

**Assumptions**

- The outcome is binary: each item or individual is classified as having the characteristic present (Yes) or not having the characteristic present (No) under each condition.

- A random sample of items or individuals is selected but because they are subjected to two different conditions, they can be regard as being samples from two related populations.

# McNemar Test for Equal Proportions: Two Populations

| | | After Intervention | |
|---|---|---|---|
| Before Intervention | | **Yes** | **No** |
| | **Yes** | a | b |
| | **No** | c | d |

- a: number of subjects with characteristic present both before and after.
- b: number of subjects where characteristic is present before but not after.
- c: number of subjects where characteristic is present after but not before .
- d: number of subjects with the characteristic absent both before and after.
- b and c are referred to *discordants*

# McNemar Test for Equal Proportions: Two Populations

- **Hypotheses**

  $H_0$: Proportion of items or individuals in the population is identical before and after an intervention.

  $H_1$: Proportion of items or individuals in the population is not identical before and after an intervention.

- Test Statistic

$$\chi^2 = \frac{(b-c)^2}{b+c}$$

However if $b$ or $c$ are small and in particular, if $(b+c) < 25$, it is recommended that a binomial test be used to get an exact p-value.

## Example: McNemar Test for Equal Proportions - Two Populations

Suppose that 35 randomly selected employees from a market research company were given a skills test before and after a course of training with performance being graded as satisfactory or unsatisfactory. We want to know if the training program has improved skills. The data is given below.

| | | **After Training** | |
|---|---|---|---|
| | | Satisfactory | Unsatisfactory |
| **Before Training** | Satisfactory | 20 | 5 |
| | Unsatisfactory | 7 | 3 |

# Example: McNemar Test for Equal Proportions - Two Populations

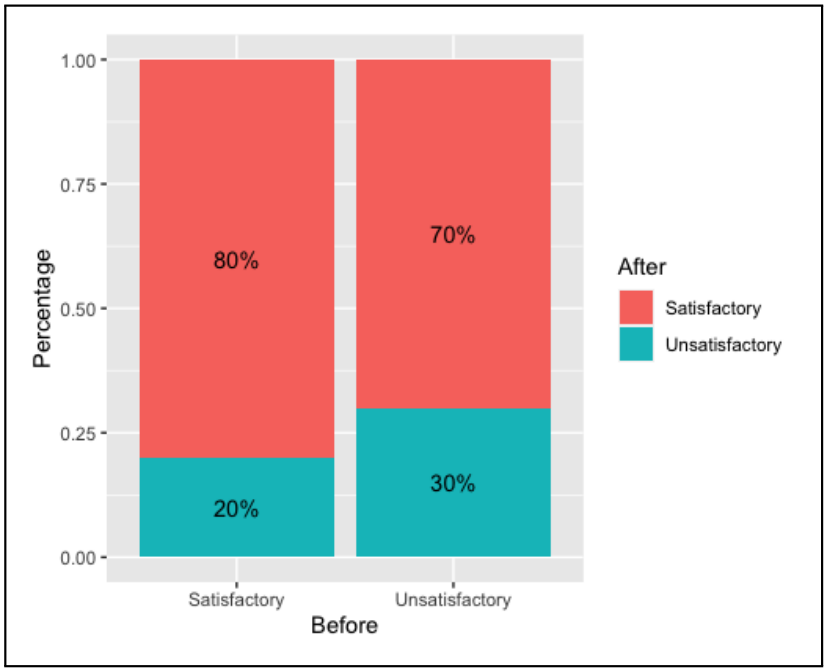
```
> #row percentages
> PercTable(ba_tab, row.vars = NULL, col.vars = NULL, justify = "right",
+           freq = TRUE, rfrq = "010", expected = FALSE, residuals = FALSE,
+           stdres = FALSE, margins = c(1,2), digits = 0)

                    Satisfactory   Unsatisfactory   Sum

Satisfactory        freq               20                5   25
                    p.row              80%              20%    .

Unsatisfactory      freq                7                3   10
                    p.row              70%              30%    .

Sum                 freq               27                8   35
                    p.row              77%              23%    .
```




We observe from the bar chart and the percentage of row totals that:

- The performances of 80% of employees that was satisfactory before training was also satisfactory after training.
- The performances of 20% of employees that was satisfactory before training was unsatisfactory after training.
- The performances of 70% of employees that was unsatisfactory before training was satisfactory after training.
- The performances of 30% of employees that was unsatisfactory before training was also unsatisfactory after training.

## Example: McNemar Test for Equal Proportions - Two Populations

$\pi_1$ and $\pi_2$: population proportions of employees whose performance in the skills test was unsatisfactory before and after the training, respectively.

1. Hypotheses:
   $H_0 : \pi_1 = \pi_2 \quad H_1 : \pi_1 \neq \pi_2$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $\chi^2 = 0.333$
4. P-value: $p - value = 0.564$
5. Decision Rule: Reject $H_0$ if p-value $< 0.05$
6. Conclusion: Since p-value $> 0.05$, do not reject $H_0$ at the 5% level and indeed any reasonably level of significance.
   There is not sufficient evidence to conclude that the training has improved performance in the skills test.

```
> #Mcnamar Test
> mcnemar.test(ba_tab,correct = FALSE)

        McNemar's Chi-squared test

data:  ba_tab
McNemar's chi-squared = 0.33333, df = 1, p-value = 0.5637
```

## Example: McNemar Test for Equal Proportions - Two Populations

Since $b + c = 5 + 7 = 12 < 25$, it is more appropriate to use the Exact Binomial test with $X = 5, n = 12, \pi_1 = \pi_2 = 0.5$.

1. Hypotheses:
   $H_0 : \pi_1 = \pi_2 \quad H_1 : \pi_1 \neq \pi_2$
2. Level of significance: $\alpha = 0.05$
3. Test statistic: $X = 5$.
4. P-value: $p - value = 0.774$
5. Decision Rule: Reject $H_0$ if p-value $< 0.05$
6. Conclusion: Since p-value $> 0.05$, do not reject $H_0$ at the 5% level and indeed any reasonably level of significance.
   There is not sufficient evidence to conclude that the training has improved performance in the skills test.

```
> #Since b+c < 25, we will use the exact test
> #To use the binomial test we use b as the number of successes and
> # n=b+c as the sample size
> binom.test(x=5, n=12 , p = 0.5)

        Exact binomial test

data:  5 and 12
number of successes = 5, number of trials = 12, p-value = 0.7744
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.1516522 0.7233303
sample estimates:
probability of success
             0.4166667
```

# Cochran Q for Equal Proportions: More than Two Populations

- In some investigations, the responses to an event (treatment) may take on only one of two values.
- We may arbitrarily designate these two possible outcomes, ‘success‘ or 1, and ‘failure‘ or 0.
- For example, in a medical experiment, we may assess the effectiveness of four pain-relieving drugs by giving each of several patients each of the drugs.
  - The patients are the blocks in the design.
  - If a given patient obtains relief from pain after receiving a drug, the response is assigned a score of 1.
  - If the patient does not obtain relief from pain, the response is assigned a score of 0.

# Cochran Q for Equal Proportions: More Than Two Populations

**Assumptions**

- The data for analysis consists of responses of $r$ blocks to $c$ independent conditions or treatments.
- The responses are 1 for 'success' and 0 for 'failure'.
- A random sample of items or individuals (blocks) is selected but because they are subjected to different conditions, they can be regard as being samples from $c$ related populations.

# Cochran Q for Equal Proportions: More Than Two Populations

- The Cochran Q test procedure is a generalisation of the McNemar technique to three or more populations but in this case are only two possible outcomes per condition or treatment.
- It is also an analogue to the Randomised Block Design (two-way ANOVA without replications) but in this case are only two possible outcomes per condition or treatment.

# Cochran Q for Equal Proportions: More Than Two Populations

- **Hypotheses**

  $H_0$: Proportion of items or individuals in the population is identical under different conditions

  $H_1$: Proportion of items or individuals in the population is not identical under different conditions

- **Test Statistic**

$$Q = \frac{c(c-1)\sum_{j=1}^{c} C_j^2 - (c-1)n^2}{cn - \sum_{i=1}^{r} R_i^2}$$

  $C_j$ = total of column $j$, $j = 1, 2, ..., c$

  $R_i$ = total of row $i$, $i = 1, 2, ..., r$

  $n$ = total sample size

- It has been shown as $r$ increases, $Q$ approaches an approximate chi-square distribution with $c$-1 degrees of freedom.

- Exact probabilities associated with $Q$ can also be determined to obtain the $p$-value of the test.

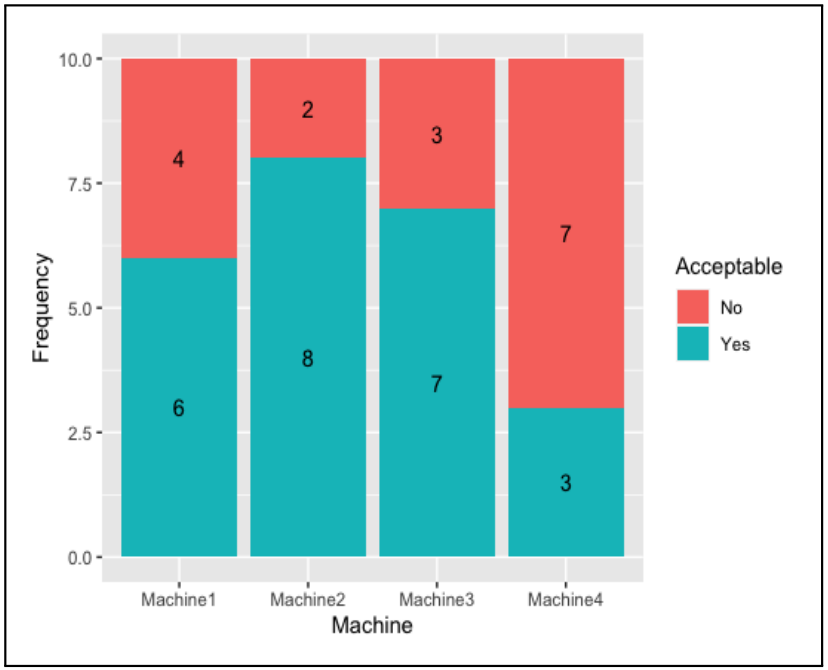
## Example: Cochran Q for Equal Proportions - More Than Two Populations

A manufacturer is considering the purchase of four machines for assembly of a key piece of equipment. An experiment is conducted to determine acceptably of the machines to employees. A random sample of 10 employees is selected, and each is assigned to operate each machine (in a random order) during a complete assembly cycle. Employees each gave a score of 1 if they find the machine acceptable and a score of 0 if the they do not. The results are shown below and in the file machines.csv. Do the data provide sufficient evidence to indicate that the four machines are not equally acceptable?

| Employee | Machine 1 | Machine 2 | Machine 3 | Machine 4 |
|---|---|---|---|---|
| 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 1 | 1 | 0 |
| 3 | 0 | 1 | 0 | 1 |
| 4 | 0 | 0 | 1 | 1 |
| 5 | 1 | 1 | 0 | 0 |
| 6 | 1 | 1 | 1 | 0 |
| 7 | 1 | 1 | 1 | 0 |
| 8 | 0 | 0 | 0 | 1 |
| 9 | 1 | 1 | 1 | 0 |
| 10 | 0 | 1 | 1 | 0 |

# Example: Cochran Q for Equal Proportions: More Than Two Populations

```
> # read in csv file
> machines<-read.csv("machines.csv", TRUE)
> mh <- as.matrix(machines[,2:5])
>
> #create tables
> Machine1 <-table(mh[,1])
> Machine2 <-table(mh[,2])
> Machine3 <-table(mh[,3])
> Machine4 <-table(mh[,4])
> m4 <- rbind(Machine1, Machine2, Machine3, Machine4)
> m4
         0 1
Machine1 4 6
Machine2 2 8
Machine3 3 7
Machine4 7 3
```




0: No
1: Yes

We observe from the table and from the bar chart that of the 10 employees, Machines 1, 2, 3, and 4 are

- acceptable to 6, 8, 7, and 3, respectively.
- not acceptable to 4, 2, 3 and 7, respectively.

## Example: Cochran Q for Equal Proportions: More Than Two Populations

1. Hypotheses:
   $H_0$: The four machines are equally acceptable to all employees.

   $H_1$: The four machines are not equally acceptable to all employees.
2. Level of significance: $\alpha = 0.05$
3. Test statistic: Cochran's Q $= 4.941$
4. P-value: $p - value = 0.176$
5. Decision Rule: Reject $H_0$ if p-value $< 0.05$
6. Conclusion: Since p-value $> 0.05$, do not reject $H_0$ at the 5% level and indeed any reasonably level of significance.

```
> #Cochrane Test
> CochranQTest(mh)

        Cochran's Q test

data:  y
Q = 4.9412, df = 3, p-value = 0.1762
```

There is not sufficient evidence that the four machines are not equally acceptable to all employees.