A tutorial for Categorical Data Analysis in R

Version 1.0.0

Konstantinos I. Bougioukas

18/11/2021

Contents

Objectives

- Applying hypothesis testing
- Investigate the possible association between two categorical variables
- Interpret the results

We will need to load the following packages

- `{rstatix}`,
- `{ggsci}`,
- `{finalfit}`,
- `{janitor}`,
- `{exact2x2}`,
- `{epitools}`,
- `{modelsummary}`,
- `{scales}`,
- `{magrittr}`,
- `{here}`,
- `{patchwork}`,
- `{tidyverse}`

1 Introduction

If we want to look at the association between two categorical variables then we can't use the mean or any similar statistic because we don't have any variables that have been measured continuously. Therefore, when we have measured categorical variables, we analyze frequencies. That is, we analyze the number of subjects that fall into each combination of categories. We can tabulate these frequencies and this is known as a cross-tabulation table or a two-way table or a contingency table.

2 Pearson's Chi-squared test of independence

If we want to see whether there's an association between two categorical variables we can use the Pearson's chi-squared test, often called chi-squared test of independence. This is an extremely elegant statistic based on the simple idea of comparing the frequencies we observe in certain categories to the frequencies we might expect to get in those categories by chance.

We will use the "Survival from Malignant Melanoma" dataset named "meldata". The data consist of measurements made on patients with malignant melanoma, a type of skin cancer. Each patient had their tumor removed by surgery at the Department of Plastic Surgery, University Hospital of Odense, Denmark, between 1962 and 1977.

For the purposes of this lesson, we are interested in the association between tumor ulceration and death from melanoma.

Ho and H1 Hypotheses:

- H_0: There is no association between the two categorical variables (they are independent)
- H_1: There is association between the two categorical variables (they are dependent)

Note In practice, the null hypothesis of independence, for our particular question, is no difference in the proportion of patients with ulcerated tumors who die compared with non-ulcerated tumors.

2.1 Preraring the data

We import the data:

```
library(readxl)
meldata <- read_excel(here("data", "meldata.xlsx"))
glimpse(meldata)
```

```
## Rows: 205
## Columns: 7
## $ time      <dbl> 10, 30, 35, 99, 185, 204, 210, 232, 232, 279, 295, 355,~
## $ status    <chr> "Alive", "Alive", "Alive", "Alive", "Died", "Died", "Di~
## $ sex       <chr> "Male", "Male", "Male", "Female", "Male", "Male", "Male~
## $ age       <dbl> 76, 56, 41, 71, 52, 28, 77, 60, 49, 68, 53, 64, 68, 63,~
## $ year      <dbl> 1972, 1968, 1977, 1968, 1965, 1971, 1972, 1974, 1968, 1~
## $ thickness <dbl> 6.76, 0.65, 1.34, 2.90, 12.08, 4.84, 5.16, 3.22, 12.88,~
## $ ulcer     <chr> "Present", "Absent", "Absent", "Absent", "Present", "Pr~
```

Table 1: Meldata (first and last 5 rows)

time	status	sex	age	year	thickness	ulcer
10	Alive	Male	76	1972	6.76	Present
30	Alive	Male	56	1968	0.65	Absent
35	Alive	Male	41	1977	1.34	Absent
99	Alive	Female	71	1968	2.9	Absent
185	Died	Male	52	1965	12.08	Present
...	NA	NA	...	...	...	NA
4492	Alive	Male	29	1965	7.06	Present
4668	Alive	Female	40	1965	6.12	Absent
4688	Alive	Female	42	1965	0.48	Absent
4926	Alive	Female	50	1964	2.26	Absent
5565	Alive	Female	41	1962	2.9	Absent

Next, we will convert the categorical variables of interest to factors.

```
meldata <- meldata %>%
  convert_as_factor(status, ulcer)

glimpse(meldata)
```

```
## Rows: 205
## Columns: 7
## $ time      <dbl> 10, 30, 35, 99, 185, 204, 210, 232, 232, 279, 295, 355,~
## $ status    <fct> Alive, Alive, Alive, Alive, Died, Died, Died, Alive, Di~
## $ sex       <chr> "Male", "Male", "Male", "Female", "Male", "Male", "Male~
## $ age       <dbl> 76, 56, 41, 71, 52, 28, 77, 60, 49, 68, 53, 64, 68, 63,~
## $ year      <dbl> 1972, 1968, 1977, 1968, 1965, 1971, 1972, 1974, 1968, 1~
## $ thickness <dbl> 6.76, 0.65, 1.34, 2.90, 12.08, 4.84, 5.16, 3.22, 12.88,~
## $ ulcer     <fct> Present, Absent, Absent, Absent, Present, Present, Pres~
```

What about to categorize the age variable into four age categories? The `cut()` function can be used for this, but first let's find the minimum and maximum value for `age` variable:

```
min(meldata$age)
```

```
## [1] 4
```

```
max(meldata$age)
```

```
## [1] 95
```

```
meldata <- meldata %>%
  mutate(age_cut = cut(age, breaks = c(4, 25, 40, 60, 95), include.lowest = TRUE,
                       labels  = c("<=25", "26 to 40", "41 to 60", ">60")))
```

Note The requirement for 'include.lowest = TRUE' when we specify breaks and the lowest cut-point is also the lowest data value.

Should we convert a continuous variable to a categorical variable?

This is a common question and something which is frequently done. Take for instance the variable age. Is it better to leave it as a continuous variable, or to chop it into categories, e.g., 26 to 40, 41 to 60 etc.?

The clear disadvantage in doing this is that information is being thrown away. Which feels like a bad thing to be doing. This is particularly important if the categories being created are large. For instance, if age was dichotomised to "young" and "old" at say 42 years (the current median age in Europe), then it is likely that relevant information to a number of analyses has been discarded.

Second, it is unforgivable practice to repeatedly try different cuts of a continuous variable to obtain a statistically significant result. This is most commonly done in tests of diagnostic accuracy, where a threshold for considering a continuous test result positive is chosen post hoc to maximise sensitivity/specificity, but not then validated in an independent cohort.

But there are also advantages to converting a continuous variable to categorical. Say the association between age and an outcome is not linear, but rather u-shaped, then fitting a regression line is more difficult. If age is cut into 10-year bands and entered into a regression as a factor, then this non-linearity is already accounted for.

Third, when communicating the results of an analysis to a lay audience, it may be easier to use a categorical representation. For instance, an odds of death 1.8 times greater in 70-year-olds compared with 40-year-olds may be easier to grasp than a 1.02 times increase per year.

So what is the answer? **Do not do it unless you have to**. Plot and understand the continuous variable first. If you do it, try not to throw away too much information. Repeat your analyses both with the continuous data and categorical data to ensure there is no difference in the conclusion (often called a sensitivity analysis).

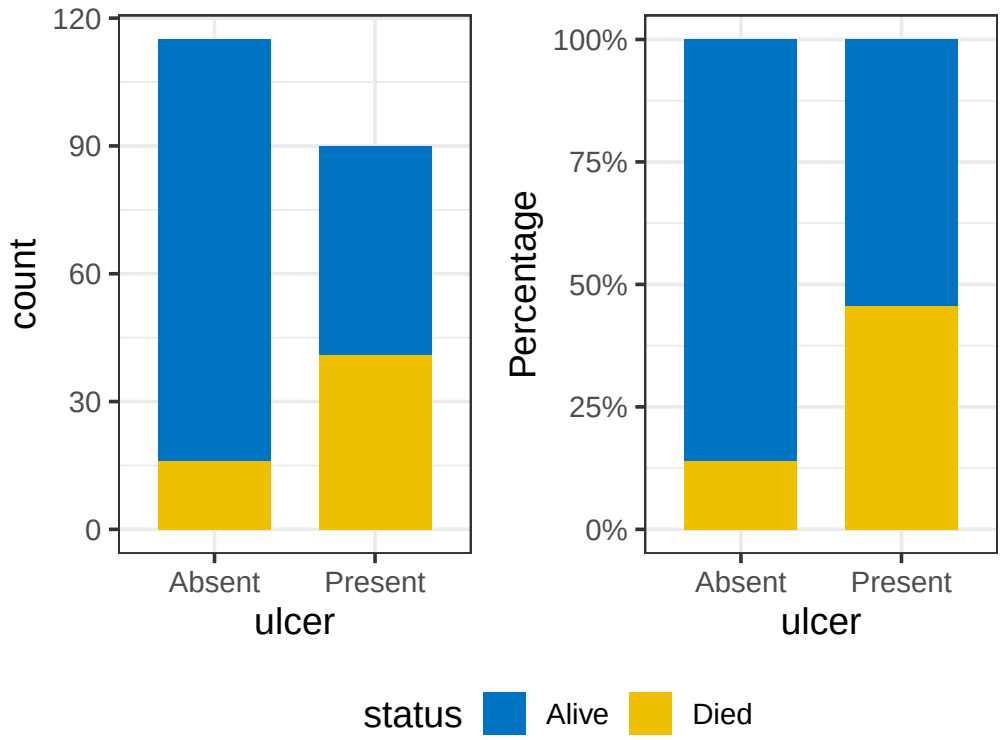
2.2 Plot the data

We are interested in the association between tumor ulceration and death from melanoma. It is useful to plot the data as counts but also as percentages. It is percentages we are comparing, but you really want to know the absolute numbers as well.

```
p1 <- meldata %>%
  ggplot(aes(x = ulcer, fill = status)) +
  geom_bar(width = 0.7) +
  scale_fill_jco() +
  theme_bw(base_size = 14) +
  theme(legend.position = "bottom")


p2 <- meldata %>%
  ggplot(aes(x = ulcer, fill = status)) +
  geom_bar(position = "fill", width = 0.7) +
  scale_y_continuous(labels=scales::percent) +
  scale_fill_jco() +
  ylab("Percentage") +
  theme_bw(base_size = 14) +
  theme(legend.position = "bottom")

p1 + p2 +
  plot_layout(guides = "collect") & theme(legend.position = 'bottom')
```



Just from the plot, death from melanoma in the ulcerated tumor group is around 40% and in the non-ulcerated group around 13%. The number of patients included in the study is not huge, however, this still looks like a real difference given its effect size.

2.3 Contigency table and Expected frequencies

First, we will create a contingency 2x2 table (this means two categorical variables with exactly two levels each) with the frequencies using the Base R.

```
tb1 <- table(meldata$ulcer, meldata$status)
tb1
```

```
##
##            Alive Died
##    Absent     99   16
##    Present    49   41
```

Next, we will also create a more informative table using the function `summary_factorlist()` which is included in `{finalfit}` package for obtaining marginal totals and row percentages.

```
row_tb1 <- meldata %>%
  summary_factorlist(dependent = "status", add_dependent_label = T,
                     explanatory = "ulcer", add_col_totals = T,
                     include_col_totals_percent = F,
                     column = FALSE, total_col = TRUE)
```

Dependent: status		Alive	Died	Total
Total N		148	57	205
ulcer	Absent	99 (86.1)	16 (13.9)	115 (100)
	Present	49 (54.4)	41 (45.6)	90 (100)

The same table can be produced using the `datasummary_crosstab()` from the {modelsummary} package:

```
datasummary_crosstab(ulcer ~ status, data = meldata)
```

From the raw frequencies, there seems to be a large difference, as we noted in the plot we made above. The proportion of patients with ulcerated tumors who die equals to 45.6% compared with non-ulcerated tumors 13.9%.

A commonly stated assumption of the chi-squared test is the requirement to have an expected count of at least 5 in each cell of the `2x2` table. For larger tables, all expected counts should be > 1 and no more than 20% of all cells should have expected counts < 5.

We can calculate the **expected frequencies** for each cell using the `expected()` function from {epitools} package:

```
epitools::expected(tb1)
```

```
##
##               Alive     Died
##   Absent  83.02439 31.97561
##   Present 64.97561 25.02439
```

Here, as we observe the assumption is fulfilled.

2.4 Run Pearson's chi-squared test

Finally we run the `chisq_test()` function:

```
chisq <- chisq_test(tb1)
chisq
```

n	statistic	p	df	method	p.signif
205	23.631	<0.001	1	Chi-square test	****

There is evidence for an association between the ulcer and status (reject H0). The proportion of patients with ulcerated tumors who died (45.6%) is significant larger compared with non-ulcerated tumors (13.9%) (p-value <0.001).

Similarly, using the Base R function `chisq.test()` :

```
chisq.test(tb1)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  tb1
## X-squared = 23.631, df = 1, p-value = 1.167e-06
```

2.5 Risk Ratio and Odds ratio

From the data in the following table

```
table.margins(tb1)
```

```
##
##           Alive Died Total
##   Absent     99   16   115
##   Present    49   41    90
##   Total     148   57   205
```

we can calculate the risk ratio by hand:

$$Risk\ Ratio = \frac{\frac{41}{90}}{\frac{16}{115}} = \frac{0.4556}{0.1391} = 3.27$$

The risk ratio with the 95% CI using R:

```
riskratio(tb1)$measure
```

```
##          risk ratio with 95% C.I.
##           estimate    lower    upper
##   Absent  1.000000       NA       NA
##   Present 3.274306 1.970852 5.439819
```

The risk of dying is 3.27 (95% CI: 1.97, 5.4) times higher for patients with ulcerated tumors compared to non-ulcerated tumors.

We can also calculate the odds ratio by hand:

$$Odds\ Ratio = \frac{\frac{41}{49}}{\frac{16}{99}} = \frac{0.837}{0.162} = 5.17$$

The odds ratio with the 95% CI using R:

```
oddsratio(tb1, method = "wald")$measure
```

```
##          odds ratio with 95% C.I.
##           estimate    lower   upper
##   Absent  1.000000       NA      NA
##   Present 5.177296 2.645152 10.1334
```

The odds of dying is 5.17 (95% CI: 2.65, 10.13) times higher for patients with ulcerated tumors compared to non-ulcerated tumors patients.

Finnaly, we can also reverse the odds ratio:

$$\frac{1}{OR} = \frac{1}{5.17} = 0.193$$

```
oddsratio(tb1, method = "wald", rev = "rows")$measure
```

```
##          odds ratio with 95% C.I.
##           estimate      lower   upper
##   Present 1.000000         NA      NA
##   Absent  0.193151 0.09868354 0.37805
```

The non-ulcerated tumors patients has 0.193 (95% CI: 0.098, 0.378) times the odds (of dying) of the ulcerated tumors. This means that the non-ulcerated tumors patients has (0.193 - 1= -0.807) 80.7% lower odds of dying than ulcerated tumors.

3 Fisher's exact test

If this assumption for the expected frequencies is not fulfilled, an alternative test can be used.

Fisher came up with a method for computing the exact probability of the chi-square statistic that is accurate when sample sizes are small. This method is called Fisher's exact test even though it's not so much a test as a way of computing the exact probability of the chi-square statistic. This procedure is normally used on 2×2 contingency tables and with small samples. However, it can be used on larger contingency tables and with large samples, but in this case it becomes computationally intensive and we might find R taking a long time to give us an answer.

For example, in a survey there are two treatment regimens studied for controlling bleeding in 25 patients with hemophilia undergoing surgery. We want to investigate if there is an association between the treatment regimen (treatment A or B) and the bleeding complications (no or yes). The null hypothesis (H0) is that the bleeding complications are independent from the treatment regimen, while the alternative (Ha) is that are dependent.

3.1 Preraring the data

We import the data:

```
library(readxl)
hemophilia <- read_excel(here("data", "hemophilia.xlsx"), col_names=TRUE)
glimpse(hemophilia)
```

```
## Rows: 28
## Columns: 2
## $ treatment <chr> "A", "A", "A", "B", "A", "B", "B", "A", "A", "A", "B", ~
## $ bleeding  <chr> "no", "no", "no", "yes", "no", "no", "no", "no", "yes",~
```

Table 2: Hemophilia Data (first and last 5 rows)

treatment	bleeding
A	no
A	no
A	no
B	yes
A	no
NA	NA
B	yes
B	no
A	yes
B	no
A	no

Next, we will convert the categorical variables to factors.

```
hemophilia <- hemophilia %>%
  convert_as_factor(treatment, bleeding)

glimpse(hemophilia)
```

```
## Rows: 28
## Columns: 2
## $ treatment <fct> A, A, A, B, A, B, B, A, A, A, B, A, B, B, A, A, B, B, B~
## $ bleeding  <fct> no, no, no, yes, no, no, no, no, yes, no, no, no, yes, ~
```

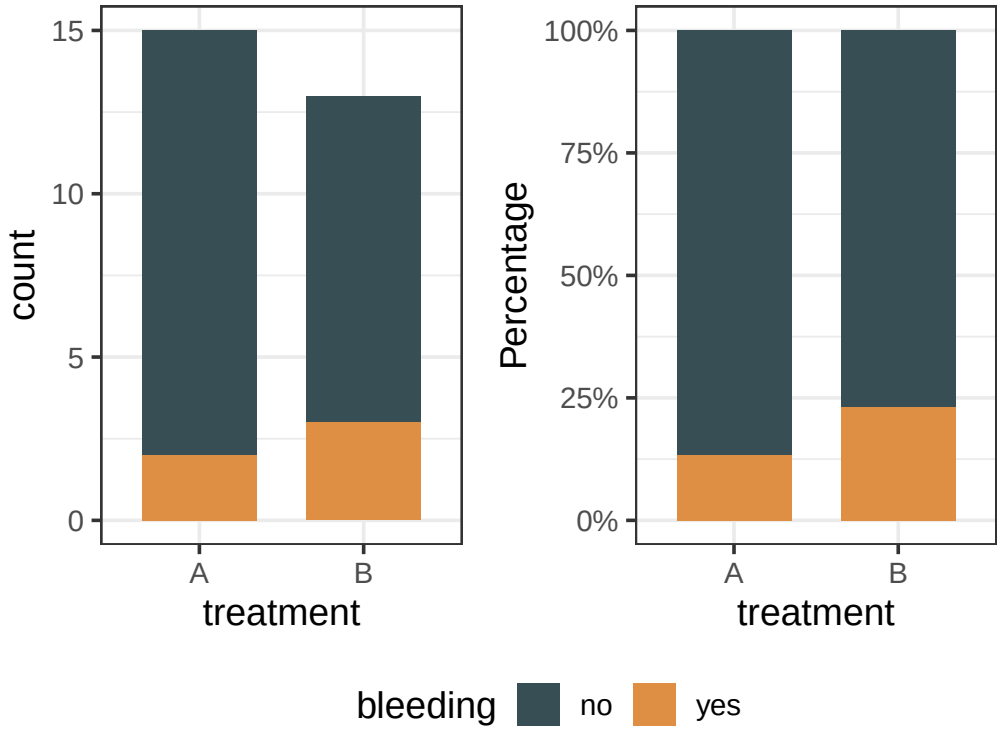
3.2 Plot the data

We count the number of patients with bleeding in the two regimens. It is useful to plot this as counts but also as percentages and compare them.

```r
p3 <- hemophilia %>%
  ggplot(aes(x = treatment, fill = bleeding)) +
  geom_bar(width = 0.7) +
  scale_fill_jama() +
  theme_bw(base_size = 14) +
  theme(legend.position = "bottom")


p4 <- hemophilia %>%
  ggplot(aes(x = treatment, fill = bleeding)) +
  geom_bar(position = "fill", width = 0.7) +
  scale_y_continuous(labels=scales::percent) +
  scale_fill_jama() +
  ylab("Percentage") +
  theme_bw(base_size = 14) +
  theme(legend.position = "bottom")

p3 + p4 +
  plot_layout(guides = "collect") & theme(legend.position = 'bottom')
```



The above bar plots with counts show graphically that the number of patients who had bleeding complications was similar in the two regimens. Note that the number of patients included in the study is small (n=28).

3.3 Contigency table and Expected frequencies

First, we will create a contingency 2x2 table with the frequencies.

```
tb2 <- table(hemophilia$treatment, hemophilia$bleeding)

tb2
```

```
##
##     no yes
##   A 13   2
##   B 10   3
```

Next, we will also create a more informative table using the function `summary_factorlist()` which is included in `{finalfit}` package for obtaining marginal totals and row percentages.

```
row_tb2 <- hemophilia %>%
  summary_factorlist(dependent = "bleeding", add_dependent_label = T,
                     explanatory = "treatment", add_col_totals = T,
                     include_col_totals_percent = F,
                     column = FALSE, total_col = TRUE)
```

Dependent: bleeding		no	yes	Total
Total N		23	5	28
treatment	A	13 (86.7)	2 (13.3)	15 (100)
	B	10 (76.9)	3 (23.1)	13 (100)

The same table can be produced using the `datasummary_crosstab()` from the {modelsummary} package:

```
datasummary_crosstab(treatment ~ bleeding, data = hemophilia)
```

From the row frequencies, there is not actually difference, as we noted in the plot we made above.

Now, we will calculate the **expected frequencies** for each cell using the `expected()` function from {epitools} package:

```
epitools::expected(tb2)
```

```
##
##           no      yes
##   A 12.32143 2.678571
##   B 10.67857 2.321429
```

In the above table there are 2 cells (50%) with expected counts less than 5 (specifically 2.67 and 2.32), so the Chi-square test is not the appropriate one. In this case the Fisher's exact test should be used instead.

3.4 Run Fisher's exact test

Finally we run the `fisher_test()` function:

```
fisher_exact <- fisher_test(tb2)
fisher_exact
```

n	p	p.signif
28	0.639	ns

The p-value = 0.64 is higher than 0.05. There is absence of evidence for an association between the treatment regimens and bleeding complications (failed to reject H0).

Similarly, using the Base R function `fisher.test()`:

```
fisher.test(tb2)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tb2
## p-value = 0.6389
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##   0.1807204 26.9478788
## sample estimates:
## odds ratio
##    1.90363
```

3.5 Having only the counts

When we read an article which reports a chi-square or a fisher exact analysis we will see only the counts in a table without having the raw data of the categorical variables. In this instance, we can create the table using the `matrix()` function and run the tests. For our example of hemophilia we have the following table:

```
dat <- c(13, 10, 2, 3)
mx <- matrix(dat, nrow = 2, dimnames = list(c("A", "B"), c("no", "yes")))
mx
```

```
##   no yes
## A 13   2
## B 10   3
```

4 McNemar test

The `asthma.xlsx` file contains data from a survey 86 children with asthma who attended a camp to learn how to self-manage their asthmatic episodes. The children were asked whether they knew (yes or not) how to manage their asthmatic episodes appropriately at both the start and completion of the camp.

4.1 Research question

Was a significant change in children's knowledge of asthma management between the beginning and completion of the health camp?

4.2 H0 and H1 Hypotheses

- H_0: there was no change in children's knowledge of asthma management between the beginning and completion of the health camp.
- H_1: there was change in children's knowledge of asthma management between the beginning and completion of the health camp.

4.3 Preraring the data

We import the data:

```
library(readxl)
asthma <- read_excel(here("data", "asthma.xlsx"), col_names=TRUE)
glimpse(asthma)
```

```
## Rows: 86
## Columns: 2
## $ know_begin <chr> "yes", "no", "yes", "no", "no", "no", "yes", "no", "no~
## $ know_end   <chr> "yes", "no", "no", "no", "no", "no", "yes", "yes", "ye~
```

Table 3: Asthma Data (first and last 5 rows)

know_begin	know_end
yes	yes
no	no
yes	no
no	no
no	no
NA	NA
yes	yes
no	yes
yes	yes
no	no
no	no

Next, we will convert the categorical variables to factors.

```
asthma <- asthma %>%
  convert_as_factor(know_begin, know_end)

glimpse(asthma)
```

```
## Rows: 86
## Columns: 2
## $ know_begin <fct> yes, no, yes, no, no, no, yes, no, no, yes, no, no, ye~
## $ know_end   <fct> yes, no, no, no, no, no, yes, yes, yes, yes, yes, no, ~
```

4.4 Contigency table

We can obtain the cross-tabulation table of the two categorical variables:

```r
tb3 <- table(know_begin = asthma$know_begin, know_end = asthma$know_end)
tb3
```

```
##           know_end
## know_begin no yes
##        no  27  29
##        yes  6  24
```

Caution! There is a basic difference between this table and the more common two-way table. In this case, the count represents the number of pairs, not the number of individuals.

We want to compare the proportion of children's knowledge of asthma management at the beginning with the proportion of children's knowledge of asthma management at the end. We can create a more informative table using the functions from `{janitor}` package for obtaining total percentages and marginal totals.

```r
asthma %>%
  tabyl(know_begin, know_end) %>%
  adorn_totals(c("row", "col")) %>%
  adorn_percentages("all") %>%
  adorn_pct_formatting(digits = 1) %>%
  adorn_ns %>%
  adorn_title
```

```
##               know_end
##  know_begin         no        yes       Total
##          no 31.4% (27) 33.7% (29)  65.1% (56)
##         yes  7.0%  (6) 27.9% (24)  34.9% (30)
##       Total 38.4% (33) 61.6% (53) 100.0% (86)
```

The proportion of children who knew to manage asthma at the beginning is (6+24)/86 = 0.349 or 34.9%. The proportion of children who knew to mange asthma at the end is (29+24)/86 = 0.616 or 61.6%.

The same table can be produced using the datasummary_crosstab() from the {modelsummary} package:

```
datasummary_crosstab(know_begin ~ know_end,
                     statistic = 1 ~ 1 + N + Percent(),
                     data = asthma)
```

Caution! The basic assumption of the test is that the sum of the **discordant** cells should be larger than 25 (that is fulfilled in our example).

4.5 Run McNemar test

Finally, we run the `mcnemar_test()` function which is applied for 2x2 tables:

```
mc_test <- mcnemar_test(tb3)
mc_test
```

n	statistic	df	p	p.signif	method
86	13.829	1	0	***	McNemar test

The proportion of children who knew to manage asthma at the end (61.6%) is significant larger compared with the proportion of children who knew to manage asthma at the beginning (34.9%) (p-value <0.001).

Similarly, using the Base R function `mcnemar.test()` :

```
mcnemar.test(tb3)
```

```
##
##  McNemar's Chi-squared test with continuity correction
##
## data:  tb3
## McNemar's chi-squared = 13.829, df = 1, p-value = 0.0002003
```

4.6 Exact binomial test

Exact binomial test for 2x2 table when the **discordant** cells are less than 25:

```
mcnemar.exact(tb3)
```