Bivariate statistics chi-square

Andri Setiyawan Benedikt Meyer Niri Gala
Yosep Dwi Kristanto

```r
library(tidyverse)
library(haven)
library(janitor)
library(knitr)
library(kableExtra)
library(DescTools)
library(ggtext)
library(scales)
```

Problem

A researcher was interested in whether animals could be trained to line-dance. He took 200 cats and tried to train them to line-dance by giving them either food or affection as a reward for dance-like behaviour. At the end of the week he counted how many animals could line-dance and how many could not. There are two categorical variables here: training (the animal was trained using either food or affection, not both) and dance (the animal either learnt to line-dance or it did not). By combining categories, we end up with four different categories.
Open the SPSS file “Cats.sav” to work on the following questions.

1. Count the frequencies how many cats fall into each category and create a contingency table!

2. Which question could be investigated? Write down a null hypothesis and alternative hypothesis.

3. What test can we use to investigate whether there’s a relationship between these categorical variables (i.e., does the number of cats that line-dance relate to the type of training used?)? Check the assumptions and then conduct an appropriate test.

Question

Is there a relationship between the type of training (using food as a reward or affection as a reward) and the cats' ability to dance (yes or no)?

Hypotheses

H_0: There is no relationship between the type of training (using food as a reward or affection as a reward) and the cats' ability to dance (yes or no).

H_1: There is a relationship between the type of training (using food as a reward or affection as a reward) and the cats' ability to dance (yes or no).

Contingency table

Based on the dataset from Cats.sav, we created a contingency table, as shown in Table 1.

```
# Import the data
cats_data <- read_sav("Cats.sav") |>
  mutate(
    training = as_factor(Training),
    dance = as_factor(Dance)
  ) |>
  select(training, dance)

# Make a contingency table
contingency_table <- cats_data |>
  count(training, dance) |>
  pivot_wider(
    names_from = dance,
    values_from = n
  ) |>
  adorn_totals(
    where = c("row", "col")
  )
contingency_table |>
  kbl(
    linesep = "",
    booktabs = TRUE,
    col.names = c("Training", "(Dance?) No", "(Dance?) Yes", "Total")
  ) |>
  kable_styling(
```

Table 1: The observed frequencies for the cats experiment

Training	(Dance?) No	(Dance?) Yes	Total
Food as Reward	10	28	38
Affection as Reward	114	48	162
Total	124	76	200

```
    full_width = TRUE,
    bootstrap_options = c("striped", "condensed"),
     latex_options = c("striped", "hold_position")
  )
```

The percentages for each cell in Table 1 are presented in the Figure 1.

```
cats_data |>
  ggplot(aes(y = training, fill = dance)) +
  geom_bar(position = "fill") +
  scale_x_continuous(labels = label_percent()) +
  scale_fill_viridis_d(
    name = "Dance"
  ) +
  theme_minimal() +
  theme(
    axis.title.x = element_blank()
  ) +
  labs(
    y = "Training"
  )
```

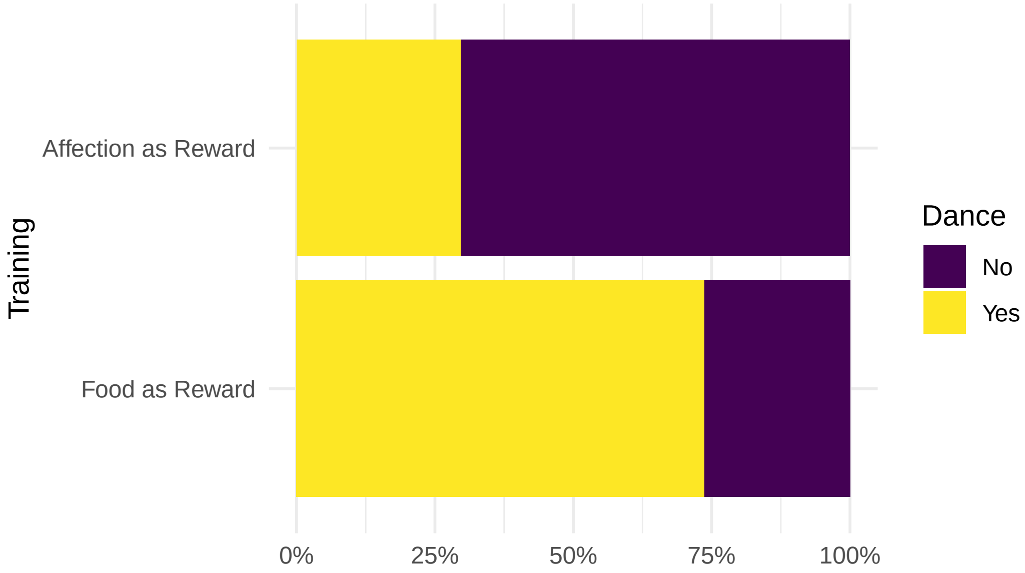

Figure 1: The observed percentages for the cats experiment

Expected frequencies table

Table 2 presents the expected frequencies based on the observed frequencies from Table 1.

```
# Make a table for expected frequencies
chisq_cats <- chisq.test(
  cats_data$training,
  cats_data$dance
)
exp_freq_table <- chisq_cats$expected |>
  as.table() |>
  as_tibble() |>
  rename(
    training = "cats_data$training",
    dance = "cats_data$dance"
  ) |>
  pivot_wider(
    names_from = dance,
    values_from = n
  ) |>
  adorn_totals(where = c("row", "col"))
exp_freq_table |> kbl(
    linesep = "",
```

Table 2: The expected frequencies for the cats experiment

Training	(Dance?) No	(Dance?) Yes	Total
Food as Reward	23.56	14.44	38
Affection as Reward	100.44	61.56	162
Total	124.00	76.00	200

```
    booktabs = TRUE,
    col.names = c("Training", "(Dance?) No", "(Dance?) Yes", "Total")
  ) |>
  kable_styling(
    full_width = TRUE,
    bootstrap_options = c("striped", "condensed"),
     latex_options = c("striped", "hold_position")
  )
```

Assumption checking

- The data has two categorical variables, i.e. `training` (Food as Reward or Affection as Reward) and `dance` (Yes or No).
- All the categories are mutually exclusive.
- As shown in Table 2, all expected frequencies are greater than or equal to 5, meeting the assumptions that (a) all expected frequencies must be greater than 1, and (b) no more than 20% of the cells in the contingency table should have an expected frequency less than 5.

Calculating the χ^2 statistics

```
# Calculate chi-squared statistics
chisq_cats$statistic
```

```
X-squared
 23.52028
```

Testing the significance of χ^2

```r
# Show the degree of freedom
chisq_cats$parameter
```

```
df 
 1 
```

```r
# Calculate the p-value
chisq_cats$p.value
```

```
[1] 1.236041e-06
```

We obtained a p-value of 1.2360413×10^{-6}. This means that, assuming the null hypothesis is true (i.e., there is no relationship between the type of training and the cats' ability to dance), the probability of obtaining the data in Cats.sav is 1.2360413×10^{-6}.

The visualization of the p-value is shown in Figure 2.

```r
# Make chi-squared distribution
ggplot(data.frame(x = c(0, 30)), aes(x = x)) +
  stat_function(
    fun = dchisq,
    args = list(df = 1),
    linewidth = 1
  ) +
  stat_function(
    fun = dchisq,
    args = list(df = 1),
    color = "blue",
    xlim = c(23.52, 30),
    linewidth = 2
  ) +
  geom_vline(
    xintercept = as.numeric(chisq_cats$statistic),
    linetype = "dashed",
    color = "blue"
  ) +
  geom_textbox(
    x = 26.5,
    y = .15,
    label = "Tail area (1 / 1 million) is too small to see",
    maxwidth = .25
  ) +
  scale_x_continuous(
    breaks = c(0, 10, 20, 23.52, 30)
  ) +
  theme_minimal() +
  theme(
    axis.title.y = element_blank(),
    axis.text.y = element_blank()
  ) +
  labs(
    x = "Chi-squared"
  )
```

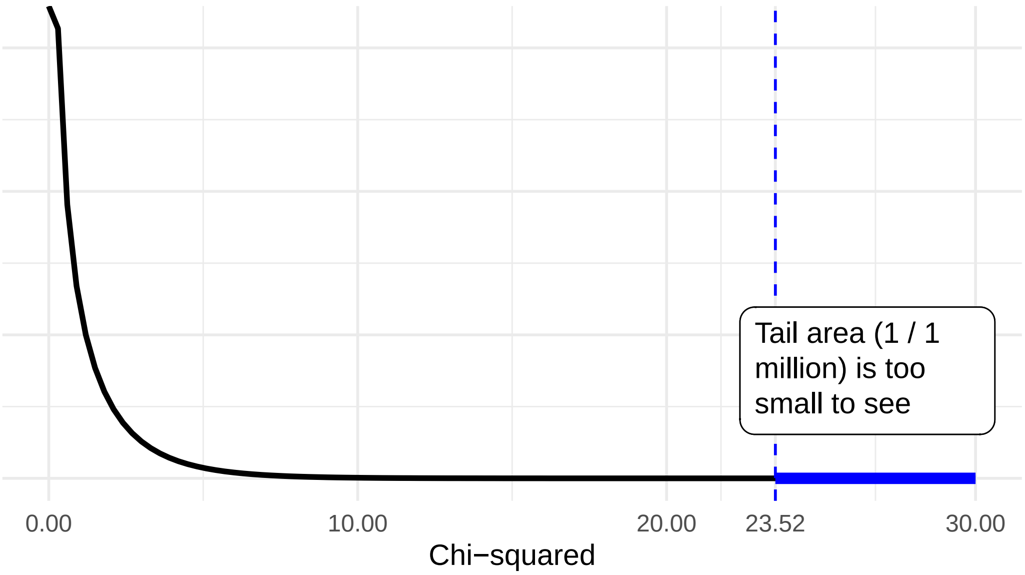

Figure 2: Visualization of the p-value for $\chi^2 = 23.52$ when $df = 1$

Interpreting χ^2

From the chi-square test, we get χ^2 23.5202781 and p-value 1.2360413 × 10^{-6}. Since the p-value is less than .05, we reject the null hypothesis.

Effect size

```r
# Calculate Phi and Cramer's V
Phi(cats_data$training, cats_data$dance)
```

```
[1] 0.3560596
```

```r
CramerV(cats_data$training, cats_data$dance)
```

```
[1] 0.3560596
```

```r
# Calculate odds ratio manually
odds_ratio <- (28 / 10) / (48 / 114)
print(odds_ratio)
```

```
[1] 6.65
```

Reporting the findings

A chi-square test of independence showed that there was a significant association between the type of training (using food as a reward or affection as a reward) and the cats' ability to dance, $\chi^2\,(1, N = 200) = 23.5202781$, $\mathrm{p} = 1.2360413 \times 10^{-6}$. The effect size is moderate, i.e. 0.3560596. The odds ratio indicates that cats were seven times more likely to be successfully trained to dance when rewarded with food rather than affection.