Testing independence

2024-09-02

```
stopifnot(
  require(patchwork),
  require(glue),
  require(here),
  require(tidyverse),
  require(vcd),
  require(vcdExtra),
  require(ggmosaic),
  require(skimr),
  require(plotly),
  require(DT),
  require(GGally),
  require(ggforce),
  require(ggfortify)
)

tidymodels::tidymodels_prefer(quiet = TRUE)

old_theme <-theme_set(theme_minimal(base_size=9, base_family = "Helvetica"))
```

- **M1 MIDS/MFA**
- **Université Paris Cité**
- Année 2024-2025
- Course Homepage

- Moodle

> **Objectives**

Confidence Intervals

We start with Confidence Intervals in a simple Gaussian setting. We have $X_1, \dots, X_n \sim_{i.i.d.} \mathcal{N}(\mu, \sigma^2)$ where μ and σ are unknown (to be estimated and/or tested).

The maximum likelihood estimator for (μ, σ^2) is $(\overline{X}_n, \hat{\sigma}^2)$ where

$$\overline{X}_n = \sum_{i=1}^{n} \frac{1}{n} X_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2$$

By Student's Theorem $\overline{X}_n$ and $\hat{\sigma}^2$ are stochastically independent $\overline{X}_n \sim \mathcal{N}(\mu, \hat{\sigma}^2/n)$ and $n\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-1}$.

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\hat{\sigma}} \sim t_{n-1}$$

where t_{n-1} denotes the Student's t distribution with $n-1$ degrees of freedom.

We have the following confidence interval for μ at confidence level $1-\alpha$:

$$\left[\overline{X}_n - \frac{\hat{\sigma} t_{n-1,\alpha/2}}{\sqrt{n}}, \overline{X}_n + \frac{\hat{\sigma} t_{n-1,\alpha/2}}{\sqrt{n}}\right]$$

Question

Simulate $N = 1000$ Gaussian samples of size $n = 100$.
Compute the empirical coverage of confidence intervals for $\alpha = 5\%$ and $\alpha = 10\%$.
Plot a histogram for replicates of $\frac{\overline{X}_n - \mu}{\hat{\sigma}\sqrt{n}}$. Overlay the density of t_{n-1}.

Testing independence

In data gathered from the 2000 General Social Survey (GSS), one cross classifies *gender* and *political party identification*. Respondents indicated whether they identified more strongly with the Democratic or Republican party or as Independents. This is summarized in the next contingency table (taken from Agresti *Introduction to Categorical Data Analysis*).

```
# GSS <- vcdExtra::GSS

T <- tribble(~ Democrat, ~ Independent, ~ Republican,
             762, 327, 468,
             484, 239, 477)
rownames(T) <- c('Females', 'Males')
```

```
Warning: Setting row names on a tibble is deprecated.
```

```
T <- as.matrix(T)
T <- as.table(T)
names(dimnames(T)) <- c("Gender", "Party identification")
```

```
prop.table(T)
```

```
         Party identification
Gender      Democrat Independent Republican
```

```
  Females 0.27638738  0.11860718 0.16974973
  Males   0.17555314  0.08668843 0.17301415
```

```
margin.table(T, 1)
```

```
Gender
Females   Males
   1557    1200
```

```
margin.table(T, 2)
```

```
Party identification
   Democrat Independent  Republican
       1246         566         945
```

> **Question**
>
> - Draw mosaicplot for the cross classification table
> - Compute the Pearson chi-square statistic for testing independence
> - Comment

Visualizing multiway categorical data

Consider the celebrated `UCBAdmissions` dataset

According to `R` documentation, this dataset is made of

> Aggregate data on applicants to graduate school at Berkeley for the six largest departments in 1973 classified by admission and sex.

This is a compilation of 4526 application files.

For each application, three variables have been reported: the `department` , the `gender` of the applicant, and whether the applicant has been `admitted`.

The dataset is a trivariate sample, which is summarized by a 3-way contingency table.

```
data("UCBAdmissions")
```

> **Question**
>
> Turn the 3-way contingency table into a dataframe/tibble with columns `Gender`, `Dept`, `Admit`, `n`, where the first columns are categorical, and the last column counts the number of co-occurrences of the values in the first three columns amongst the UCB applicants.

> **Question**
>
> Make it a bivariate sample by focusing on `Gender` and `Admit`: compute the *margin table* Draw the corresponding mosaicplot and compute the chi-square independence statistic. Comment.

Question

Visualize the three-way contingency table using double-decker plots from `vcd`

Question

Question

Viewing the `UCBAdmissions` dataset, which variable would you call a *response* variable? Which variable would you call *covariates*?
Test independence between `Gender` and `Dept`.

Question

For each department of application (`Dept`), extract the partial two-way table for `Gender` and `Admit`. Test each two-way table for independence. How many departments pass the test at significance level 1%, 5%?

Note that the two-way cross-sectional slices of the three-way table are called partial tables.

What we observed has a name.

Simpson's paradox

The result that a marginal association can have different direction from the conditional associations is called Simpson's paradox. This result applies to quantitative as well as categorical variables.