Practical 2.5: Categorical data

ADS2

Semester 2, 2023/24

Learning Objectives

- Describing and visualising categorical data
- Chi-square test of categorical data

1. Simulation of the probability of goodness-of-fit test

In the lecture you learned how to use the chi-square value to measure the discrepancy between observed data and expected data. Based on Season preferences data, use a simulation to get the distribution of χ^2 and calculate the p-value (the probability that the discrepancy is larger than the observed data).

```
Poll_seasons <- data.frame(Spring = 40, Summer = 30, Autumn = 18, Winter = 28)
```

Hint: If there is no preference for a particular season, the frequencies should follow a 0.25 in each category. Generate a (large) population with the expected proportions, sample 116 values (same as the values in the poll) and calculate the χ^2, replicate this a large number of times to get an approximate χ^2 distribution.

```
equal_preferences <- sum(Poll_seasons) * 0.25
```

Use *plot(density(. . .))* to visualise the distribution of simulated χ^2 values. What is the curve like? Do you get the similar probability as in the lecture material?

How the curve changes when considering a higher population size? How about a larger sample size?

Compare the probability with the result from *chisq.test()*.

2. Chi-square distribution and degree of freedom

Generate random chi-square values with different degrees of freedom. Use it as your simulation tool to get the curves as in the lecture. Hint: use *rchisq()* to directly obtain χ^2 values for each degree of freedom.

3. Chi-square test of homogeneity

Input the data from the two categories (season preference and reported allergy) into a data frame. Visualise the data as bar, balloons and mosaics. Hint: Try *mosaicplot()*

Perform chi-square test on the data.

4. Chi-square test and Fisher's exact test

Input the data from the survival after geneX KO into a matrix.

Perform a chi-square test on the data. Turn off the Yates's continuity correct assigning the *correct* argument to *FALSE*. What warning message do you get? If you turn on the correction, what changes?

Perform a Fisher's exact test on the data. Hint: use *fisher.test()*

Previous version by Hugo Samano.

Last update by DJ MacGregor in 2024