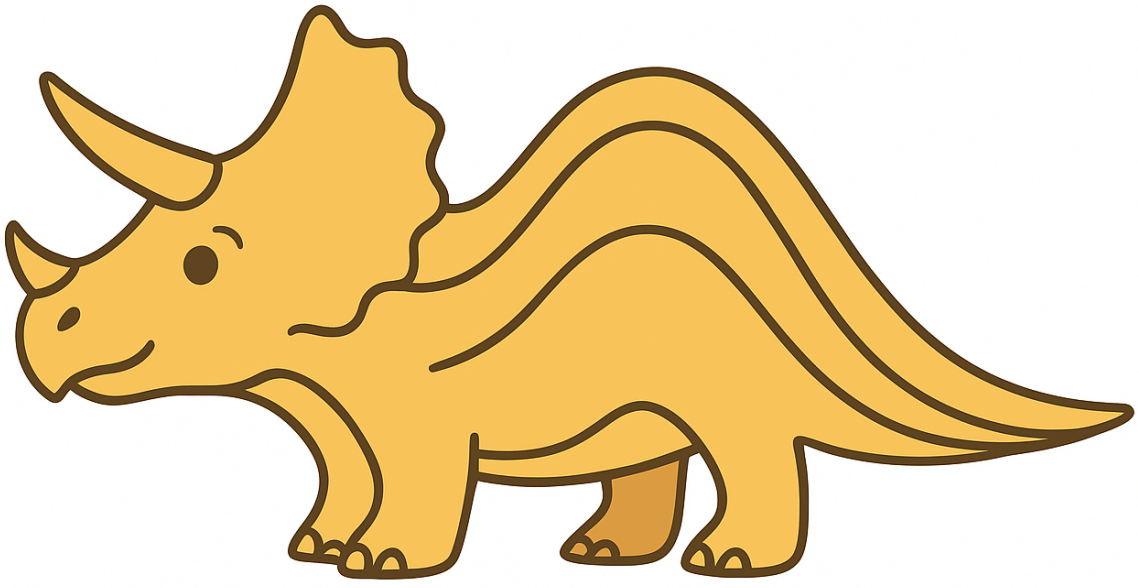
Day 24 - Theory-based inference for multiple proportions

Learning outcomes

In this section, we explore how to assess the relationship between two categorical variables, regardless of how many levels each variable has using theory-based methods.



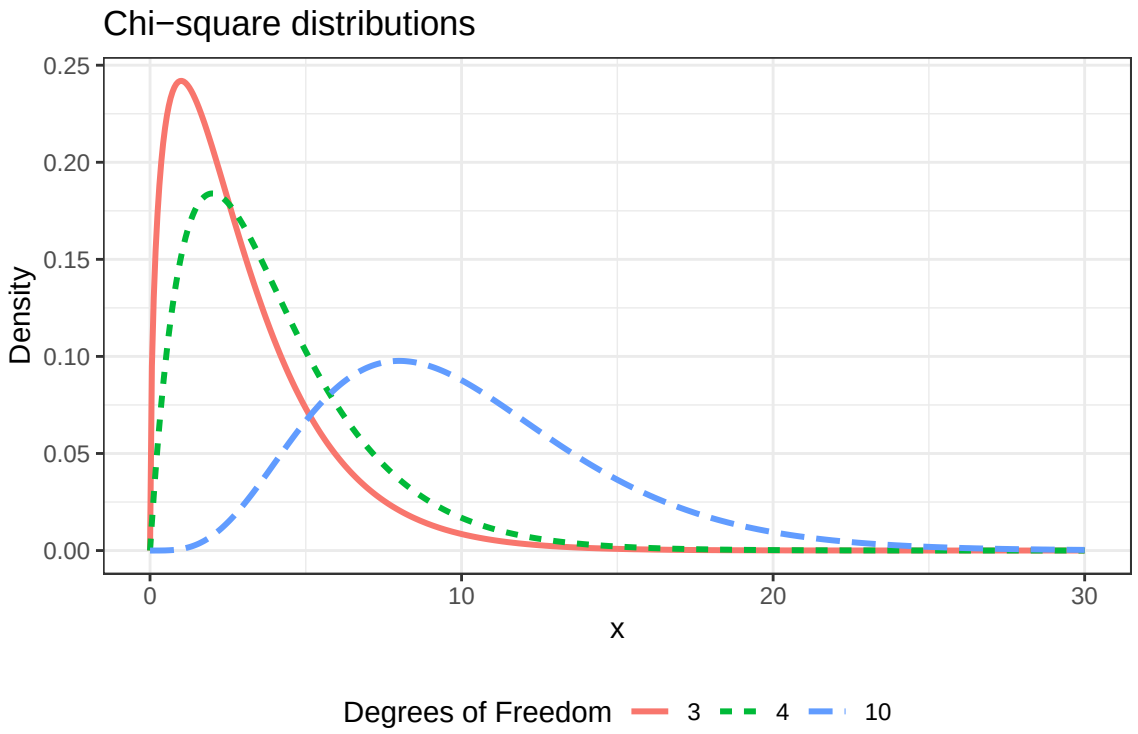
Chi-squared distribution

Chi-squared distribution

The Chi-squared distribution is described by its _____.

$$\chi^2_{df}$$

where df represents the degrees of freedom. Under the null hypothesis, $df = (r-1) \times (c-1)$ where r is the number of rows and c is the number of columns.



Conditions for inference

- **Independence**: We assume that the observations are _____.
- **Large enough sample**: The *expected* cell count should be at least _____, for all cells in the table.

  To calculate the expected count for the i^{th} cell in the table, compute:

$$E_i = \frac{\text{(row total)} \times \text{(column total)}}{\text{table total}}$$

Titanic survival

Is there a statistically significant association between passenger class and survival for passengers of the Titanic? The data set below contains a sample of 891 passengers from the Titanic and a collection of passenger characteristics.

```
head(titanic)
```

```
# A tibble: 6 x 12
  PassengerId Survived Pclass Name    Sex     Age SibSp Parch Ticket  Fare Cabin
        <int>    <int>  <int> <chr>   <chr> <dbl> <int> <int> <chr>  <dbl> <chr>
1           1        0      3 Braund~ male     22     1     0 A/5 2~  7.25 ""
2           2        1      1 Cuming~ fema~    38     1     0 PC 17~ 71.3  "C85"
3           3        1      3 Heikki~ fema~    26     0     0 STON/~  7.92 ""
4           4        1      1 Futrel~ fema~    35     1     0 113803 53.1  "C12~
5           5        0      3 Allen,~ male     35     0     0 373450  8.05 ""
6           6        0      3 Moran,~ male     NA     0     0 330877  8.46 ""
# i 1 more variable: Embarked <chr>
```

1. Create a stacked bar chart to visualize these data.
2. Is the independence assumption reasonable for these data? Why or why not?

3. Is the large enough sample assumption met for these data? Why or why not?

Recall: the X^2 statistic

Given some contingency table, the X^2 statistic may be computed in the following way

$$X^2 = \sum_i \left(\frac{O_i - E_i}{\sqrt{E_i}} \right)^2$$

where i indexes the elements in the table, O_i is the observed count, and E_i is the expected count.

4. Use R to calculate the test statistic.
5. What distribution does the test statistic follow under the null hypothesis?

Tip

Make sure to provide the degrees of freedom!

6. Use R to plot the distribution of the X^2 statistic under the null hypothesis. Plot the test statistic as a line on the distribution. *Hint:* see `?dchisq`

7. Look back at the simulation-based null distribution that we created last week. How does it compare to this distribution?

8. Use `R` to find the p-value. *Hint:* see `?pchisq`

9. Interpret the p-value as a conditional probability in the context of the problem.

10. Interpret the p-value as strength of evidence in the context of the problem.

11. Provide a mosaic plot of the standardized residuals, and discuss what you see.