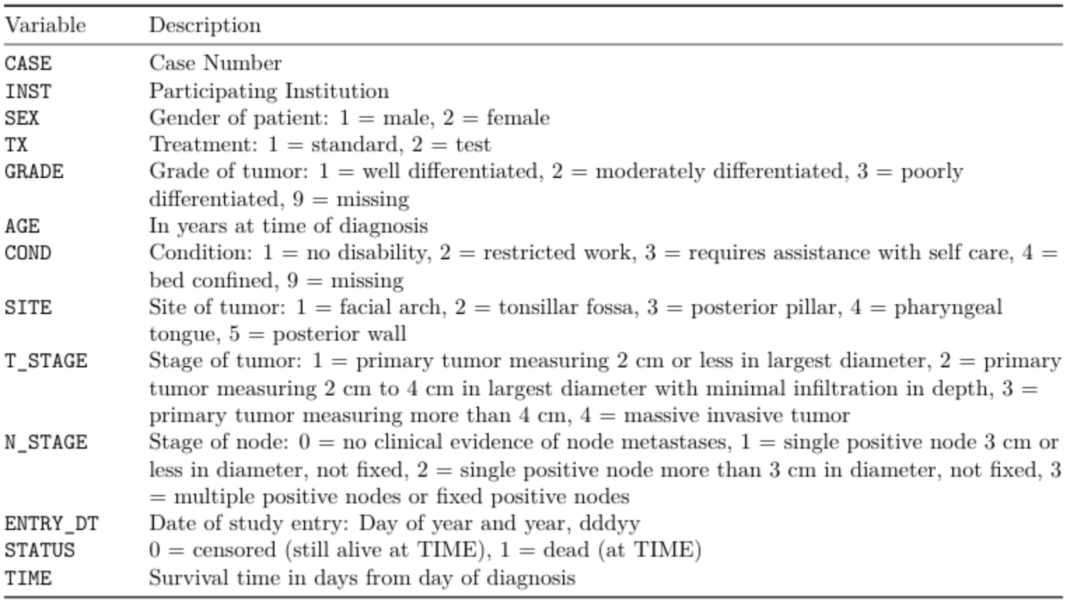
Lab 7 Homework: Inference for Categorical Data

Background

Oropharyngeal cancer affects the tissues of the oropharynx, which includes the back one-third of the tongue, the soft palate, the side and back walls of the throat, and the tonsils. The American Cancer Society estimated that in 2014 about 37,000 people would be diagnosed with oral cavity or oropharyngeal cancer, with approximately 7,300 deaths from these cancers. The five-year survival rate after diagnosis depends on the stage and location of the tumor. Treatment options typically involve a combination of surgery, chemotherapy, and radiation therapy.

The Data

The dataset `pharynx.csv` contains 14 variables from a large clinical trial conducted by the Radiation Therapy Oncology Group in the United States. Patients in this study had squamous carcinoma from three sites in the oropharynx. For this lab, we are interested in examining the relationship between survival (500 days post-diagnosis) and sex.



Variable	Description
CASE	Case Number
INST	Participating Institution
SEX	Gender of patient: 1 = male, 2 = female
TX	Treatment: 1 = standard, 2 = test
GRADE	Grade of tumor: 1 = well differentiated, 2 = moderately differentiated, 3 = poorly differentiated, 9 = missing
AGE	In years at time of diagnosis
COND	Condition: 1 = no disability, 2 = restricted work, 3 = requires assistance with self care, 4 = bed confined, 9 = missing
SITE	Site of tumor: 1 = facial arch, 2 = tonsillar fossa, 3 = posterior pillar, 4 = pharyngeal tongue, 5 = posterior wall
T_STAGE	Stage of tumor: 1 = primary tumor measuring 2 cm or less in largest diameter, 2 = primary tumor measuring 2 cm to 4 cm in largest diameter with minimal infiltration in depth, 3 = primary tumor measuring more than 4 cm, 4 = massive invasive tumor
N_STAGE	Stage of node: 0 = no clinical evidence of node metastases, 1 = single positive node 3 cm or less in diameter, not fixed, 2 = single positive node more than 3 cm in diameter, not fixed, 3 = multiple positive nodes or fixed positive nodes
ENTRY_DT	Date of study entry: Day of year and year, dddyy
STATUS	0 = censored (still alive at TIME), 1 = dead (at TIME)
TIME	Survival time in days from day of diagnosis

Practice

Make sure you use the option `correct=FALSE` when doing all one-proportion tests, two-proportion tests, and chi-square tests to ensure that R results exactly match corresponding hand calculations from the lab.

1. Investigate the association between survival (500 days post-diagnosis) and sex.

(a) Create a variable that indicates whether each patient survived at least 500 days.

(b) What are the two tests that could be used to answer this question? What are the null and alternative hypotheses for each test? When is each test appropriate?

(c) Perform the chi-squared test to investigate if there is an association between survival past 500 days and sex. Does one group (male or female) have a better survival outcome? Write a brief paragraph summarizing your results, including basic descriptive statistics.

(d) Explain why the analysis performed is not sufficient to make definitive conclusions regarding survival differences based on sex.

2. Check the assumptions for the chi-squared test.

(a) What assumption regarding expected cell counts must be satisfied for a valid chi-squared test?

(b) Is the assumption regarding expected cell counts satisfied for the previous test on survival past 500 days and sex? Why or why not? If assumptions are violated, explain why and suggest an appropriate alternative test. (You do not need to conduct and interpret the test.)

(c) Suppose you wanted to determine if there was an association between survival past 500 days and the stage of the tumor (`T_STAGE`). Is the assumption regarding expected cell counts satisfied? Why or why not? If assumptions are violated, explain why and suggest an appropriate alternative test. (You do not need to conduct and interpret the test.)

3. Calculate p-values for given test statistics.

Calculate p-values for the following chi-squared test statistic values and degrees of freedom. Indicate if each p-value would be significant at the $\alpha = 0.05$ level.

- $\chi^2 = 1,\ \text{df} = 1$ ________________
- $\chi^2 = 3,\ \text{df} = 1$ ________________
- $\chi^2 = 5,\ \text{df} = 1$ ________________
- $\chi^2 = 1,\ \text{df} = 2$ ________________
- $\chi^2 = 3,\ \text{df} = 2$ ________________
- $\chi^2 = 5,\ \text{df} = 2$ ________________

As the test statistic increases, the p-value (increases/decreases). Therefore, larger test statistics present (more/less) evidence against the null hypothesis. The same test statistic value with different degrees of freedom (can/cannot) result in a different conclusion for a specified level of significance (e.g., $\alpha = 0.05$).