Math 4720: Statistical Methods

13^{th} and 14^{th} Week Summary (04/25/25)

- **Chi-square Tests** .....

- The **chi-square goodness-of-fit** test allows us to test whether a categorical variable follows a given probability distribution.

- A categorical variable has k possible outcomes, with probabilities $\pi_1, \pi_2, \pi_3, \ldots, \pi_k$. That is, π_i is the probability of the i^{th} outcome. We have n independent observations from this categorical variable.

- We use chi-square goodness-of-fit test for testing the null hypothesis that the probabilities have specified values:

  $H_0 : \pi_1 = \pi_{10}, \pi_2 = \pi_{20}, \pi_3 = \pi_{30}, \ldots, \pi_k = \pi_{k0}$

  The expected count of outcome i, $E_i = n\pi_{i0}$. Expected counts do not have to be round numbers.

- The chi-square statistic is a measure of how far observed counts are from expected counts under the null hypothesis. The formula for the statistic is:

  $$\chi^2 = \sum \frac{(\text{observed count - expected count})^2}{\text{expected count}} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

  Each of the k terms in the sum is called a chi-square component

- The chi-square goodness of fit test involving k outcomes refers to the chi-square distribution with $k-1$ degrees of freedom. The p-value is the area to the right of χ^2 under the density curve of this chi-square distribution.

- You can safely use the chi-square goodness-of-fit test with critical values from the chi-square distribution when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater. In particular, the chi-square goodness-of-fit test can be used when all the expected counts are 5 or greater.

- The **chi-square test of independence** is for checking the dependence of two categorical variable. The null hypothesis is given by :

  H_0 : Two categorical variables are independent of each other.

- The chi-square test for a two-way table with r rows and c columns uses critical values from the chi-square distribution with $(r-1)(c-1)$ degrees of freedom.

- You can safely use the chi-square test with critical values from the chi-square distribution when no more than 20% of the expected counts are less than 5 and all individual expected counts are 1 or greater. In the special case of a 2×2 table, all four expected counts should be 5 or greater.

- **Regression** .....

  A tool to describe how a response (dependent) variable y changes based on an explanatory (independent) variable x. Often, a straight line is used to model and predict y given x.

  $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

  where ϵ_i are normally distributed error terms with mean 0 and constant variance σ^2.

- Least Squares Estimation:

  The slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ are chosen to minimize the sum of squared residuals, $\sum(y_i - \hat{y}_i)^2$.

- Assumptions:

  Linear relationship between x and y.

  Independent, normally distributed error terms with constant variance.

  Watch for outliers or influential points that may distort the regression.

- ANOVA and $\boldsymbol{R^2}$:

  An ANOVA table splits total variation into that explained by the regression and that attributed to error. The coefficient of determination R^2 measures the proportion of the variation in y explained by the model.

- Inference:

  Hypothesis tests (e.g., $H_0 : \beta_1 = 0$) determine whether there is a statistically significant linear relationship. Confidence and prediction intervals provide estimates for the mean response and for individual new observations.

- Cautions:

  Regression lines only capture linear patterns; always check scatterplots.

  Extrapolating far beyond observed x-values can be unreliable.

  Residual plots are crucial for checking model assumptions.