Stat 111 Spring 2025 Week 6: Hypothesis Tests (Rice Chapter 9)

1. **The Neyman-Pearson Paradigm**

   a) Describe the Neyman-Pearson paradigm for deciding between a null and alternative hypothesis. Define Type I and Type II errors and make a courtroom analogy (or any other analogy that seems relevant - e.g., replay challenges in sports) to point out the asymmetry in the two hypotheses.

   b) Imagine testing the defect rate θ of a manufactured component. The stated rate is $\theta_0 = 0.1$ and you are concerned that it might be higher than this (there's no problem if it is lower than 0.1). Suppose we take a sample of $n = 400$ and sound an alarm if 50 or more of the sampled components are defective. What are the hypotheses and significance level? Report the power of this test if $\theta = 0.11$, $\theta = 0.15$ or $\theta = 0.09$. How would things change if your threshold were 51 or more defectives? Discuss the connections between significance level and power.

   c) Define the P-value of a test and its connection to the significance level. With $n = 400$, suppose you observe 45 defects. What is the P-value?

   d) Describe the connection between power and sample size in the context of the defect rate example. Show how to solve for the smallest sample size to have power of at least 0.9 for the alternative in part c, when working at $\alpha \approx 0.05$. Use the Normal approximation to the Binomial to get approximate answers, and then use pbinom to make things precise.

2. **Likelihood Ratio Tests**

   a) State the Neyman-Pearson lemma and define the likelihood ratio test of two simple hypotheses. For the defect rate example in presentation 1, show that the likelihood ratio test rejects for large values of the sample proportion of defects.

   b) Give a Bayesian justification for the likelihood ratio as a decision tool. Show that the conditional (posterior) probability or odds of H_o decreases with the likelihood ratio, no matter what prior probabilities are assigned. Explain how the NP paradigm defines a decision rule without assigning prior probabilities (in order to remain 'objective'). Point out the limitations of this approach. For example, a test for Lyme disease is positive with probability 0.9 for a person who has Lyme (H_a), and with probability 0.05 for a person who does not have Lyme (H_o). Define a likelihood ratio test with significance level $\alpha = 0.05$.

   c) As an example where the LR test has higher power than an alternate test with the same significance level, consider a test of $H_o : \mu = 0$ vs $H_a : \mu = 3$ for the NFL home field advantage data with $n = 272$ and assuming the data are iid Normal with $\sigma = 14.0$ known. An alternative test to the usual z-test is based on the count Y of positive values, with $Y \sim \text{Binom}(n, 0.5)$ under H_o. Find an α close to 0.05 that you can achieve exactly with a Binomial test. Show that the test that rejects for large $\bar{X}$ has higher power than the test that rejects for large Y. Note, however, how the test based on Y (the "sign test") is valid even if the data are not Normal.

   d) Outline the formal proof of the Neyman Pearson lemma to show that no test of comparable significance level has higher power than the likelihood ratio test. See Rice 9.2.

3. **Uniformly Most Powerful tests and CI's**

   a) Explain how a most powerful test of simple hypotheses implies a uniformly most powerful (UMP) test for a 1-sided alternative hypotheses. Use the NFL home advantage data as an example. Identify the most powerful test of level $\alpha = 0.05$ and compute its power for a test of $H_o : \mu = 0$ against $H_a : \mu = 1.0$ and against $H_a : \mu = 3.0$. Explain why we cannot compute the power for a test of $H_o : \mu = 0$ vs. $H_a : \mu > 0$, but that we can identify a uniformly most powerful test for this 1-sided alternative.

   b) Show how to define a 2-sided test by taking the union of the rejection regions for two UMP 1-sided tests. Use the NFL example and carry out a test of $H_o : \mu = 0$ vs. $H_a : \mu \neq 0$. Find the 2-sided p-value for this test based on the observed value $\bar{x} = 1.95$ (and assuming $\sigma = 14.0$, with $n = 272$.

   c) Explain why there is no UMP 2-sided test. Define the directional power for a 2-sided test as the probability of rejecting $H_o : \theta = \theta_o$ in the correct direction when $\theta \neq \theta_o$. Graph the directional power for a test based on the NFL data as a function of μ, the true mean home advantage. Explain why this approaches $\alpha/2$ as μ approaches the null value $\mu_o = 0$ (whereas the power of a 1-sided test approaches α). Add lines for two 1-sided tests and show that one or the other always has higher power than the 2-sided test. Explain why we should still prefer two-sided tests in most situations.

   d) Explain how to define a 1-sided Confidence interval by inverting a 1-sided significance test, and a 2-sided CI by taking the intersection of two 1-sided CI's. Construct a 95% z-CI based on the NFL data. Explain the duality between the test and CI.

4. **Generalized Likelihood Ratio Tests.**

   a) Define the generalized likelihood ratio (GLR) test and say how it generalizes the likelihood ratio test for composite hypotheses.

   b) Consider the defect rate test from presentation 1. Explain why the GLR test of $H_o : \theta \leq \theta_o$ vs. $H_o : \theta > \theta_o$ is equivalent to the UMP test of $H_o : \theta = \theta_o$ vs. $H_a : \theta > \theta_o$.

   c) Let $X_1, \ldots, X_n$ represent the times until the first goal in $n = 19$ Swarthmore women's soccer matches and consider a test of $H_o : \mu = 30$ vs. $H_a : \mu \neq 30$. Assume the X_i's are iid Exponential$(1/\mu)$. Contrast the $\alpha = 0.05$ GLR test rejection region to the union of two $\alpha = 0.025$ 1-sided rejection regions, and argue that the latter is preferable (and easier to construct!). Find the implied 95% CI for the mean time until the first goal based on $\bar{x} = 37.5$ minutes with $n = 19$.

   d) For $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$, with both μ and σ^2 unknown, show that the GLR test of $H_o : \mu = \mu_o$ vs. $H_a : \mu \neq \mu_o$ rejects for large values of $|T|$, where $T = \frac{\bar{x} - \mu_o}{s/\sqrt{n}}$. Carry out the t-test with $\mu_o = 0$ for the NFL home advantage data and report a 95% t-confidence interval for μ, the mean home advantage based on $\bar{x} = 1.95$ and $s = 14.44$, with $n = 272$.

5. **Approximate GLR Tests**

   a) Rejecting for small values of the GLR test statistic Λ is equivalent to rejecting for large values of $-2\log(\Lambda)$. Theorem A on p. 341 states that for large samples from Exponential family distributions, the distribution of $-2\log\Lambda$ is approximately $\chi^2_{(\nu)}$, where ν is the difference in the number of parameters that need to be estimated overall, and under H_o. Show (as in Example A on p. 339) that this is exactly true for testing a Normal mean with known σ, and approximately true when σ is unknown (using the result of 4d).

b) Show how to use Lagrange multipliers to find the maximum likelihood estimates for multinomial cell probabilities.

c) Show how to test multinomial probabilities using the approximate GLR test. Show that the GLR test statistic is asymptotically equivalent to the Pearson Chi-square statistic $\sum \frac{(O-E)^2}{E}$, where O and E are observed and expected counts. As an example, suppose $n = 30$ rolls of a 6-sided die result in counts of (10,5,5,5,5,0) for the outcomes $1, 2, ..., 6$ (this is what I would expect for my biased die that has two 1's and no 6). Compute the P-value of a test for $H_o : p_1 = p_2 = \ldots = p_6 = 1/6$ against a general alternative using the Chi-square approximation. Also use simulation to compute an exact P-value (up to simulation error). Simulate many sets of 30 fair dice rolls and compute the GLR stat for each replicate data set. Estimate the P-value as the proportion of times you get a statistic as large or larger than the observed value.

d) (If time) Consider a test of $H_o : \theta = \theta_o$ vs. $H_a : \theta \neq \theta_o$, based on $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, \theta)$. Explain why this distribution is not in the Exponential family. Show the null distribution is exactly Chi-square, but with 2 df, not the 1 df prescribed by Theorem A.