STA 711 Homework 7

Due: Friday, March 28, 10:00pm on Canvas.

Instructions: Submit your work as a single PDF. You may choose to either hand-write your work and submit a PDF scan, or type your work using LaTeX and submit the resulting PDF. See the course website for a homework template file and instructions on getting started with LaTeX and Overleaf.

Tests for variances

1. Suppose that $X_1, ..., X_n \overset{iid}{\sim} N(0, \sigma^2)$. We wish to test the hypotheses $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_A : \sigma^2 = \sigma_1^2$, were $\sigma_0^2 < \sigma_1^2$.

    (a) Show that the most powerful test of these hypotheses rejects when $\sum\limits_{i=1}^n X_i^2 > c$, for some value c.

    (b) Find c such that the test in part (a) has size α.

2. Suppose that $X_1, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, with both μ and σ^2 unknown. Our hypotheses are $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_A : \sigma^2 \neq \sigma_0^2$. Propose a test statistic and rejection region for testing these hypotheses, such that the resulting test is size α.

Paired t-test

Many studies involve the analysis of *paired* data, in which two observations are taken on the same individual. For example, researchers studying whether a teaching intervention improves student learning may assess each student's knowledge before and after the intervention, and examine how much the scores changed.

Suppose that we observe pairs $(Y_{11}, Y_{12}), (Y_{21}, Y_{22}), ..., (Y_{n1}, Y_{n2})$. The pairs are independent, that is $(Y_{i1}, Y_{i2}) \perp\!\!\!\perp (Y_{j1}, Y_{j2})$ for $i \neq j$. Within each pair, we assume that

$$Y_{i2} = Y_{i1} + \varepsilon_i$$

where $\varepsilon_i \overset{iid}{\sim} N(\mu, \sigma^2)$, and both μ and σ^2 are unknown. We wish to test $H_0 : \mu = 0$ vs. $H_A : \mu \neq 0$.

3. Construct a test statistic for these hypotheses which follows a t_{n-1} distribution. Your answer should demonstrate that the statistic does indeed follow a t_{n-1} distribution.

Chi-squared goodness-of-fit test

A random variable X follows a *categorical* distribution with k categories if $X \in \{1, ..., k\}$ and the probability that X is in category j is $P(X = j) = p_j$, with each $p_j \in [0, 1]$ and $\sum\limits_{j=1}^k p_j = 1$. We write $X \sim Categorical(p_1, ..., p_k)$. (This is just a generalization of the Bernoulli to more than two categories).

Suppose that we observe $X_1, ..., X_n \overset{iid}{\sim} Categorical(p_1, ..., p_k)$. Let $n_j = \sum\limits_{i=1}^{n} \mathbb{1}\{X_i = j\}$ (the number of observations in category j), and note that $\sum_j n_j = n$. We are interested in testing the hypotheses

$$H_0 : (p_1, ..., p_k) = (p_{01}, ..., p_{0k}) \qquad H_A : (p_1, ..., p_k) \neq (p_{01}, ..., p_{0k})$$

(in other words, are the true probabilities for each category equal to hypothesized probabilities).

4. (a) Find the maximum likelihood estimators $\widehat{p}_j$ of each probability p_j. (*Hint:* You will need to add a constraint that $\sum_j \widehat{p}_j = 1$. Lagrange multipliers may be helpful.)

   (b) Let Λ denote the likelihood ratio test statistic for the hypotheses above. Show that $2\log(\Lambda)$ can be written in the form

   $$2\log(\Lambda) = 2\sum_{j=1}^{k} n_j \log\left(\frac{n_j}{e_j}\right),$$

   where you will need to define e_j.

   (c) Show that if each $|n_j - e_j|$ is small, then

   $$2\log(\Lambda) \approx \sum_{j=1}^{k} \frac{(n_j - e_j)^2}{e_j}.$$

   (*Hint:* Use a second-order Taylor approximation...)

Nonparametric estimation

So far, we have focused on estimated parameters in parametric distributions. But what if we want to estimate a distribution without assuming any parametric family? Let $X_1, ..., X_n$ be iid from some distribution with cdf F. The *empirical distribution function* F_n is a *nonparametric* estimate of F defined by

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{X_i \leq t\}.$$

5. Show that for each t, $F_n(t) \xrightarrow{p} F(t)$. (In other words, the empirical distribution function converges pointwise to the true cdf).

6. Let $t \in \mathbb{R}$ be given. Suppose for this specific t, we want to test the hypotheses

   $$H_0 : F(t) = p_0 \qquad H_A : F(t) \neq p_0.$$

   Derive a Wald test using the empirical distribution function F_n; you should state the test statistic, demonstrate that it has the desired asymptotic distribution, and specify when the test will reject the null hypothesis.