Gov 2002: Problem Set 7

Problem Set Instructions

This problem set is due on **Apr. 5th, 11:59 pm** Eastern time. Please upload a PDF of your solutions to Gradescope. We will accept hand-written solutions but we strongly advise you to typeset your answers in Rmarkdown. Please list the names of other students you worked with on this problem set.

Question 1 (20 points)

A welfare policy is applied to a group of n people to counter job loss caused by trade shocks, resulting in $X_1, \ldots, X_n$ which are i.i.d. $\mathcal{N}(\theta, 1)$, where θ is the theoretical mean effect of the remedy (e.g., the logged wage difference between new and old jobs), defined so that $\theta > 0$ if the remedy is helpful on average, $\theta < 0$ if it is harmful on average, and $\theta = 0$ if it does nothing. Let our Type I error rate be $\alpha = 0.05$. Consider testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Let the rejection region be

$$\mathcal{R} = \{\mathrm{x} : |\hat{\theta}_{PI}| > c\},$$

where $\hat{\theta}_{PI}$ is the plug-in estimator of θ.

(a) Find c so that the test has Type I error rate (i.e., size) equal to α. (Note that c will depend on the sample size, n.)

(b) Find the power of the test, $\beta(\theta)$, for $\theta > 0$

(c) Prove that when $\theta \neq 0$, the power $\beta(\theta) \to 1$ as $n \to \infty$.

(d) Suppose now that $n = 10^4$ and we observe $\bar{x} = 0.02$. What is the p-value? Does the test say to reject or retain H_0?

Question 2 (20 points)

In recent years it has become common in statistics to want to perform many simultaneous hypothesis tests. Let $p_1 \ldots p_m$ be indepedent p-values, corresponding to m hypothesis tests. Each of the m hypothesis tests has a simple null. Suppose that m_0 of the m null hypotheses are true. We decide in advance to conduct these tests at level α (i.e., we reject the null for tests where the p-value is less than α). The *familywise error rate* is the probability of making at least one Type I error. For the purposes of this question, you may assume that the p-values follow a continuous uniform distribution (think about why).

(a) Find the familywise error rate. What happens to the familywise error rate as m_0 gets large?

(b) The *Bonferroni procedure* is described as follows: instead of rejecting the null hypotheses with $p_i < \alpha$, we reject the null hypotheses with $p_i < \frac{\alpha}{m}$. Show that under this procedure, the familywise error rate is at most α. (Hint: You might find Markov's Inequality helpful: $\mathrm{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$)

(c) In (b), why not instead reject the null hypotheses with $p_i < \frac{\alpha}{m_0}$, considering that $m_0 \leq m$ (and often in practice m_0 is much smaller than m), which would seem to result in rejecting more false nulls while still keeping the familywise error rate at most α?

(d) Another procedure is to reject all null hypotheses with $p_i < 1 - (1-\alpha)^{1/m}$ (This is known as the Sidak Procedure). Show that under this procedure, the familywise error rate is again at most α.

Question 3 (20 points)

In the case Hazelwood School District v. United States (1977), which went to the Supreme Court, the United States argued that the Hazelwood School District in Missouri was practicing racial discrimination in their hiring of teachers. Data were presented showing that only 15 out of 405(3.7%) of the teachers hired by the school district were black.

(a) In the original case, the District Court ruled in favor of the school district, saying "The number of black teachers employed by the Hazelwood district is undeniably meager. Nonetheless, it has kept pace with the small but steadily increasing black enrollment in the district. For the 1970-71 school year the six black teachers hired by the Hazelwood district comprised less than one percent of its total faculty. However, the number of black students enrolled during that period was likewise only one percent of the total district attendance." The United States appealed this decision, arguing that the relevant population for comparison was the teachers in St. Louis County and St. Louis City, not the students in the Hazelwood district. (St. Louis City borders but is not included in St. Louis County, since the city seceded from the county in 1876.) About 15.4% of the teachers in this population were black, which intuitively seems like a massive disparity compared with the 3.7% statistic in the Hazelwood district. The Appeals Court ruled in favor of the United States. The school district then appealed the Appeals Court decision to the Supreme Court, arguing that St. Louis City should be excluded from the population for comparison, due to the city having very different hiring guidelines than were present in the county. Discuss the principles and considerations you would use in deciding on the population to compare the Hazelwood district statistic with. (You do not need to resolve the question of whether teachers in St. Louis City should be included in the comparison population.)

(b) The Supreme Court ruled that the comparison should be with St. Louis County, excluding St. Louis City. In St. Louis County, 5.7% of teachers were black. Suppose that, before observing the data for Hazelwood district, we know that $n = 405$ teachers will be hired and that $X \sim \mathrm{Bin}(n, p)$ of these teachers are black. What would you use as your null and alternative hypotheses if you would like to show that Hazelwood district discriminates against black teachers in their hiring process? Conduct this test, and find the p-value. (We are just looking at disparities in hiring; of course in reality there are many complications such as who applies for which jobs, how people are recruited, what salaries are offered, etc.)

(c) Repeat (b), except with a Normal approximation to the Binomial

(d) The majority opinion stated that "the expected value of 23 would be less than two standard deviations from the observed total of 15″. Is this correct? The Supreme Court concluded that the null hypothesis of no discrimination could not be rejected (presumably setting $\alpha = 0.05$ and mentioning"two standard deviations" since $1.96 \approx 2$). In a dissenting opinion, Justice Stevens stated that "one of my law clerks advised me that, given the size of the two-year sample, there is only about a 5% likelihood that a disparity this large would be produced by a random selection from the labor pool." Was Justice Stevens's clerk correct? How can the discrepancy between the majority opinion failing to reject the null and Justice Stevens rejecting the null be reconciled?

Question 4 (20 points)

Let $X_1, \ldots, X_n$ be i.i.d. Exponential r.v.s with mean μ. We aim to test $H_0 : \mu \leq 1$ vs. $H_1 : \mu > 1$. Set your seed to 02138 for simulation exercises.

(a) Derive a test statistic based on the MLE for μ along with a rejection region that produces a test of size α. This rejection region should not be based on asymptotics. For $n = 100$ and $\alpha = 0.05$, calculate the critical value based on your analytical results (Hint: For $X_1, \ldots, X_n \sim^{iid} \text{Expo}(\lambda)$, $\sum_{i=1}^n X_i \sim \text{Gamma}(n, \frac{1}{\lambda})$, you may find R functions like `qgamma` and `pgamma`} helpful)

(b) Derive a second test statistic based on using large sample theory (asymptotics) for the MLE of μ (so now you should assume the sample size n is large). Derive the associated rejection region for a test that is approximately of size α.

(c) Simulate 1000 data sets each of size $n = 5$ i.i.d. Exponentials with mean $\mu = 1$. Conduct both tests (from (a) and (b)) on each data set for a Type I error rate of $\alpha = 0.025$, and calculate the fraction of times each test leads to a rejection of the null hypothesis across the 1000 data sets. Compare these fractions to the α. Discuss this result.

Question 5 (20 points)

Let $Y_1, \ldots, Y_n$ be i.i.d. random variables such that $E(Y_i) = 0, \operatorname{Var}(Y_i) = 1$, and the fourth moment $E(Y_i^4)$ exists.

(a) Define

$$S_n = \frac{1}{n} \sum_{i=1}^n Y_i^2, \qquad V_n = \frac{1}{n} \sum_{i=1}^n \left(Y_i^2 - 1\right)^2.$$

Prove the following asymptotic result:

$$\frac{\sqrt{n} \cdot (S_n - 1)}{\sqrt{V_n}} \xrightarrow{d} \mathcal{N}(0, 1).$$

Hint: Start by applying the CLT to S_n and the LLN to V_n.

(b) Find the constant c such that

$$c \cdot \frac{\sqrt{n}\left(\exp\left(S_n\right) - e\right)}{\sqrt{V_n}} \xrightarrow{D} \mathcal{N}(0,1)$$

Hint: what does $\sqrt{n}(\exp(S_n) - e)$ converge in distribution to?