STA261 (Summer 2022) - Assignment 4

These problems are meant to test your understanding of the concepts in Module 4. They are *not* to be handed in. Some of these have been modified (or in some cases taken directly) from questions in the *Additional Resources* listed in the course syllabus, and no claims of originality are made.

---

1. Suppose that $L(x)$ and $U(x)$ satisfy $\mathbb{P}_\theta\left(L(X) \leq \theta\right) = 1 - \alpha_1$ and $\mathbb{P}_\theta\left(U(X) \geq \theta\right) = 1 - \alpha_2$, where $L(x) \leq U(x)$ for all $x \in \mathcal{X}$. Show that $\mathbb{P}_\theta\left(L(X) \leq \theta \leq U(X)\right) = 1 - \alpha_1 - \alpha_2$.

2. Given a random sample $X_1, X_2, \ldots, X_n$ from each of the following pdfs, find a $1-\alpha$ confidence interval for θ:

   (a)
   $$f_\theta(x) = 1, \quad \theta - \frac{1}{2} < x < \theta + \frac{1}{2}, \quad \theta \in \mathbb{R}.$$

   (b)
   $$f_\theta(x) = \frac{2x}{\theta^2}, \quad 0 < x < \theta, \quad \theta > 0.$$

3. For the density in Example 4.11, show that $Q(\mathbf{X}, \theta) = X_{(1)} - \theta$ is a pivotal quantity, and use it to find a $1-\alpha$ confidence interval. Compare its length to the that of the LRT-based confidence interval in Example 4.11.

4. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ where $\mu \in \mathbb{R}$ and σ^2 is known. Find the minimum value of n to guarantee that a 0.95 confidence interval for μ will have length no more than $\sigma/4$.

5. Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \mathcal{N}\left(\mu, \sigma^2\right)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Find $k \in \mathbb{R}$ to make $(0, kS^2)$ a $1-\alpha$ confidence interval for σ^2.

6. (a) Show $\bar{X}_n - \mu$ is a pivotal quantity in a location family with pdf $f_\mu(x) = f(x - \mu)$.
   (b) Show that $\bar{X}_n/\sigma$ is a pivotal quantity in a scale family with pdf $f_\sigma(x) = \frac{1}{\sigma} f(x/\sigma)$.
   (c) Show that $(\bar{X}_n - \mu)/\sqrt{S^2}$ is a pivotal quantity in a location-scale family with pdf $f_{\mu,\sigma}(x) = \frac{1}{\sigma} f(\frac{x-\mu}{\sigma})$.

7. Let $X \sim \text{Beta}\,(\theta, 1)$, where $\theta > 0$. Find a pivotal quantity for θ and use it to construct a $1-\alpha$ confidence interval.

8. In the simple linear regression setup, find $1-\alpha$ confidence intervals for α and β using the test statistics from Assignment 3 Q19.

9. In Lecture 7, we argued that we can turn certain kinds of hypothesis tests into confidence regions,[1] and vice versa. Turn this into a rigorous statement and prove it.

10. We remarked in lecture that there isn't a very deep theory of optimal confidence intervals (at least compared to point estimation and hypothesis testing). There are some useful results, however. Here's one:

---

[1] A $(1-\alpha)$-confidence region is just like a $(1-\alpha)$-confidence interval, except it doesn't have to be an interval specifically – just a random set $C(\mathbf{X})$ such that $\mathbb{P}_\theta\left(\theta \in C(\mathbf{X})\right) \geq 1 - \alpha$ for all $\theta \in \Theta$. This relaxation makes the question a lot more straightforward.

Theorem 1. *Let f_θ be a unimodal pdf. If the interval $[a, b]$ satisfies*

i) $\int_a^b f_\theta(t)\,\mathrm{d}t = 1 - \alpha$

ii) $f_\theta(a) = f_\theta(b) > 0$, *and*

iii) $a \leq t^* \leq b$, *where t^* is the mode of f_θ,*

then $[a, b]$ is the shortest among all intervals that satisfy the first condition.

(a) Use the theorem to prove that if f_θ is a symmetric unimodal pdf, then of all the intervals $[a, b]$ that satisfy $\int_a^b f_\theta(t)\,\mathrm{d}t = 1 - \alpha$, the shortest is obtained by choosing a and b so that $\int_{-\infty}^a f_\theta(t)\,\mathrm{d}t = \int_b^\infty f_\theta(t)\,\mathrm{d}t = \alpha/2$.

(b) Show that the Z-interval and the t-interval are the shortest exact $(1-\alpha)$ confidence intervals for μ under their respective $\mathcal{N}(\mu, \sigma^2)$ models.

11. Prove that if we observe $\mathbf{X} = \mathbf{x}$, the observed ecdf $\hat{F}_n(t)$ satisfies the following properties:

(a) $\hat{F}_n(t)$ is an increasing function

(b) $\lim_{t\to\infty} \hat{F}_n(t) = 1$

(c) $\lim_{t\to-\infty} \hat{F}_n(t) = 0$

(d) (Optional) $\hat{F}_n(t)$ is right-continuous

12. Suppose the following observed sample is assumed to arise from a $\mathcal{N}(\mu, \sigma^2)$ distribution, with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$:

```
14.0   9.4   12.1   13.4   6.3   8.5   7.1   12.4   13.3   9.1
```

(a) Plot the standardized residuals

(b) Construct a Normal probability plot of the standardized residuals

(c) What conclusions can you draw?

13. Suppose a die is tossed 1000 times, and the following frequencies are observed for the number of pips up when the die comes to a rest:

x_1	x_2	x_3	x_4	x_5	x_6
163	178	142	150	183	184

Perform a chi-squared test to assess whether we have evidence that this is not a symmetrical die.

14. Let $S_1 \sim \text{Bin}(n_1, p_1)$ and $S_2 \sim \text{Bin}(n_2, p_2)$ be independent, where $p_1, p_2 \in (0, 1)$ and n_1, n_2 are known. We're interested in testing $H_0 : p_1 = p_2$ versus $H_A : p_1 > p_2$.

(a) Under H_0, let p be the common value of $p_1 = p_2$. Show that the joint pmf of (S_1, S_2) is given by

$$f_p(s_1, s_2) = \binom{n_1}{s_1}\binom{n_2}{s_2} p^{s_1+s_2}(1-p)^{n_1+n_2-(s_1+s_2)},$$

and show that $S := S_1 + S_2$ is sufficient for p.

(b) Given an observation $S = s$, explain why it's reasonable to use S_1 as a test statistic and reject H_0 when S_1 is large.

(c) Show that

$$\mathbb{P}(S_1 = s_1 \mid S = s) = \frac{\binom{n_1}{s_1}\binom{n_2}{s - s_1}}{\binom{n_1 + n_2}{s}}.$$

What is this distribution?

(d) Argue that given an observation $S = s$, a reasonable p-value for our test is given by

$$\sum_{j=s_1}^{\min\{n_1,s\}} \frac{\binom{n_1}{j}\binom{n_2}{s-j}}{\binom{n_1+n_2}{s}}.$$

The test characterized by this p-value is called *Fisher's exact test.* It's used to test for independence between the (categorical) variables in a contingency table.

(e) Suppose that $(A_1, B_1), (A_2, B_2), \ldots, (A_n, B_n)$ are iid pairs of categorical data taking values in $\{0, 1\} \times \{0, 1\}$. Define the following quantities:

$$\begin{aligned}
n_1 &= \sum_{i=1}^{n} \mathbb{1}_{B_i=0} \\
n_2 &= \sum_{i=1}^{n} \mathbb{1}_{B_i=1} \\
S_1 &= \sum_{i=1}^{n_1} (\mathbb{1}_{A_i=0} \mid B_i = 0) \\
S_2 &= \sum_{i=1}^{n_1} (\mathbb{1}_{A_i=0} \mid B_i = 1) \\
p_1 &= \mathbb{P}(A_i = 0 \mid B_i = 0) \\
p_2 &= \mathbb{P}(A_i = 0 \mid B_i = 1).
\end{aligned}$$

One can show that the test derived above is equivalent to testing the hypothesis that the A_i's are independent of the B_i's.[2] Suppose the following contingency table was obtained from classifying members of a sample of $n = 10$ from a student population according to the classification variables A and B, where $A = 0$ indicates male, $A = 1$ indicates female, $B = 0$ indicates conservative, and $B = 1$ indicates liberal:

	$B = 0$	$B = 1$
$A = 0$	2	1
$A = 1$	3	4

Use Fisher's exact test to check the model that says gender and political orientation are independent.

[2] Strictly speaking, n_1 and n_2 are random for the independence test, but that's not important here.