**541B Midterm 2 Solutions**[1]

## 1. Question 1

Let $X_1, X_2, \ldots : \Omega \to \mathbf{R}$ be i.i.d random variables. Let $Y_1, Y_2, \ldots$ be a sequence of estimators so that for any $n \geq 1$, $Y_n = t_n(X_1, \ldots, X_n)$ for some $t_n \colon \mathbf{R}^n \to \mathbf{R}$. For any $n \geq 1$, define the **jackknife estimator** of $Y_n$ to be

$$Z_n := nY_n - \frac{n-1}{n} \sum_{i=1}^{n} t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n). \qquad (*)$$

- Assume that there exists $\theta, a, b \in \mathbf{R}$ such that

  $$\mathbb{E}Y_n = \theta + a/n, \quad \forall n \geq 1.$$

  Show that

  $$\mathbb{E}Z_n = \theta.$$

- The jackknife described above involves summing over all ways to delete one of the samples from $X_1, \ldots, X_n$. Write a formula for a term to add to $(*)$ that also sums over all ways to delete exactly two of the samples from $X_1, \ldots, X_n$ in $t_{n-2}$.

*Solution.*

$$\begin{aligned}
\mathbb{E}Z_n &= n\theta + a - \frac{n-1}{n} \sum_{i=1}^{n} \mathbb{E}t_{n-1}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \\
&\overset{(*)}{=} n\theta + a + -\frac{n-1}{n} \sum_{i=1}^{n} \left( \theta + \frac{a}{n-1} \right) \\
&= (n - (n-1))\theta + (a - a) = \theta.
\end{aligned}$$

For the second part, we could use

$$\sum_{1 \leq i < j \leq n} t_{n-2}(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_{j-1}, X_{j+1}, \ldots, X_n)$$

## 2. Question 2

Write down the generalized likelihood ratio estimate for the following alpha particle data, as we did in class for a slightly different data set. The corresponding test treats individual counts of alpha particles as independent Poisson random variables, versus the alternative that the probability of a count appearing in each box of data is a sequence of nonnegative numbers that sum to one.

| $m$ | 0,1,2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | $\geq 17$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| # Ints | 16 | 26 | 58 | 102 | 125 | 146 | 163 | 164 | 120 | 100 | 72 | 54 | 20 | 12 | 10 | 4 |

Suppose we wanted to plot the MLE for the Poisson statistic (i.e. plot the denominator of the generalized likelihood ratio test statistic $\frac{\sup_{\theta \in \Theta} f_\theta(X)}{\sup_{\theta \in \Theta_0} f_\theta(X)}$) as a function of $\lambda$.

Describe in detail how you would plot this MLE on a computer, with particular detail on how to avoid outputting zeros or infinities that should not occur.

[1]November 3, 2025, 

*Solution.* As discussed in class, the denominator of the GLR statistic is

$$\begin{aligned}
&\sup_{\theta\in\Theta_0} f_\theta(x)\\
&= \sup_{\lambda>0} 1207!\Big(\prod_{j=2}^{15}\frac{[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}}{x_j!}\Big)\cdot\frac{(e^{-\lambda}[1+\lambda+\lambda^2/2])^{x_1}}{x_1!}\cdot\frac{[e^{-\lambda}\sum_{i=17}^{\infty}\frac{\lambda^i}{i!}]^{x_{16}}}{x_{16}!}\\
&= \sup_{\lambda>0} 1207!\Big(\prod_{j=2}^{15}\frac{[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}}{x_j!}\Big)\cdot\frac{(e^{-\lambda}[1+\lambda+\lambda^2/2])^{x_1}}{x_1!}\cdot\frac{[e^{-\lambda}(e^{\lambda}-\sum_{i=0}^{16}\frac{\lambda^i}{i!})]^{x_{16}}}{x_{16}!}\\
&= \sup_{\lambda>0} 1207!\Big(\prod_{j=2}^{15}\frac{[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}}{x_j!}\Big)\cdot\frac{(e^{-\lambda}[1+\lambda+\lambda^2/2])^{x_1}}{x_1!}\cdot\frac{[1-e^{-\lambda}\sum_{i=0}^{16}\frac{\lambda^i}{i!})]^{x_{16}}}{x_{16}!}.
\end{aligned}$$

where $x_1,\dots,x_{16}$ are the table data values (that is $s_1=16$, $x_2=26$, …, $x_{16}=4$). In order to find the MLE on a computer, we could just "plug in" this formula as a single variable function of $\lambda$, plot it, and find its maximum value. However, this does not work since both the numerators and denominators are extremely large numbers, i.e. just plugging in the formula would lead to either zero or infinite values for $f_\theta(x)$. To get around this issue, note that the denominators are not functions of $\lambda$, so for the purpose of computing the MLE, we can ignore the $x_j!$ terms, i.e. it suffices to maximize the following function of $\lambda$:

$$\Big(\prod_{j=2}^{15}[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j}\Big)\cdot(e^{-\lambda}[1+\lambda+\lambda^2/2])^{x_1}\cdot[1-e^{-\lambda}\sum_{i=0}^{16}\lambda^i i!)]^{x_{16}}.$$

(We have dropped the 1207! term for a similar reason; it evaluates to infinity in double precision arithmetic, and it does not matter for the purpose of optimizing $f_\theta(x)$)

Now the $x_i$ exponents are too large and could lead to overflow. To ameliorate this issue we take this function to a small power (we find that 1/200 suffices), i.e. we plot

$$\Big(\prod_{j=2}^{15}[e^{-\lambda}\lambda^{j+1}/(j+1)!]^{x_j/200}\Big)\cdot(e^{-\lambda}[1+\lambda+\lambda^2/2])^{x_1/200}\cdot[1-e^{-\lambda}\sum_{i=0}^{16}\lambda^i i!)]^{x_{16}/200}.$$

This function can now be plotted on a computer without any overflow issues.

## 3. Question 3

Let $X_1,\dots,X_n$ be a random sample from a Gaussian distribution with known variance $\sigma^2>0$ and unknown mean $\mu\in\mathbb{R}$. Fix $\mu_0\in\mathbb{R}$. Suppose we want to test the hypothesis $H_0$ that $\mu=\mu_0$ versus the alternative $H_1$ that $\mu\neq\mu_0$. That is, $\Theta=\mathbb{R}$, $\Theta_0=\{\mu_0\}$ and $\Theta_0^c=\{\mu\in\mathbb{R}\colon \mu\neq\mu_0\}$. Also, for any $x=(x_1,\dots,x_n)\in\mathbb{R}^n$,

$$f_\mu(x)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

- Explicitly describe the rejection region of the generalized likelihood ratio test.
- Denote $X=(X_1,\dots,X_n)$. If $H_0$ is true, describe the distribution of

$$2\log\frac{\sup_{\theta\in\Theta}f_\theta(X)}{\sup_{\theta\in\Theta_0}f_\theta(X)}$$

(Hint: you can freely use the identity $\sum_{i=1}^n[(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2] = n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - 2n(\frac{1}{n}\sum_{i=1}^n x_i - \mu_0)(\frac{1}{n}\sum_{j=1}^n x_j - \mu_0)$)

*Solution.* From Example 2.63, the MLE is the sample mean, i.e. for any $x \in \mathbb{R}^n$,

$$\sup_{\mu\in\Theta} f_\mu(x) = f\left(\frac{x_1+\cdots+x_n}{n}\right)(\bar{x}).$$

Since $\Theta_0$ is just a single point, we can then write the rejection region of the generalized likelihood ratio test as

$$\begin{aligned}
C &= \{x\in\mathbb{R}^n\colon \sup_{\mu\in\Theta_0} f_\mu(x) \geq k \sup_{\mu\in\Theta} f_\mu(x)\} \\
&= \left\{x\in\mathbb{R}^n\colon \prod_{i=1}^n e^{-\frac{(x_i-\mu_0)^2}{2\sigma^2}} \geq k\prod_{i=1}^n e^{-\frac{(x_i-\bar{x})^2}{2\sigma^2}}\right\} \\
&= \left\{x\in\mathbb{R}^n\colon e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n[(x_i-\mu_0)^2-(x_i-\bar{x})^2]} \geq k\right\} \\
&= \left\{x\in\mathbb{R}^n\colon \sum_{i=1}^n\left[(x_i-\frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2\right] \leq -2\sigma^2\log k\right\} \\
&= \left\{x\in\mathbb{R}^n\colon -n\left(\frac{1}{n}\sum_{j=1}^n x_j - \mu_0\right)^2 \leq -2\sigma^2\log k\right\} \\
&= \left\{x\in\mathbb{R}^n\colon \left|\frac{1}{n}\sum_{j=1}^n x_j - \mu_0\right| \geq \sqrt{-2n^{-1}\sigma^2\log k}\right\}.
\end{aligned}$$

So, the test rejects the null hypothesis, unless $\frac{1}{n}\sum_{j=1}^n X_j$ is close to $\mu_0$. As anticipated by Proposition 3.27, the hypothesis test corresponds to confidence intervals for the sample mean. (Above we used the identity $\sum_{i=1}^n[(x_i - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2] = \sum_{i=1}^n(x_i - \mu_0 + \mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - (x_i-\mu_0)^2 = n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - 2n(\frac{1}{n}\sum_{i=1}^n(x_i-\mu_0))(\frac{1}{n}\sum_{j=1}^n x_j - \mu_0) = n(\mu_0 - \frac{1}{n}\sum_{j=1}^n x_j)^2 - 2n(\frac{1}{n}\sum_{i=1}^n x_i - \mu_0)(\frac{1}{n}\sum_{j=1}^n x_j - \mu_0)$).

Finally, note that

$$2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} = \frac{n}{\sigma^2}\left(\frac{1}{n}\sum_{j=1}^n X_j - \mu_0\right)^2 = \left(\frac{1}{\sigma/\sqrt{n}}\sum_{j=1}^n[X_j-\mu_0]\right)^2$$

has a chi-squared distribution with one degree of freedom. In fact, this holds asymptotically

## 4. Question 4

Let $X_1,\ldots,X_{16}$ denote real valued random variables with $\sum_{j=1}^{16} X_j = 1207$. Denote $X = (X_1,\ldots,X_{16})$. Suppose we know that

$$Z := 2\log\frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} = 2\cdot 1207\sum_{j=1}^{16}\frac{X_j}{1207}\log\left(\frac{X_j/1207}{p_j}\right),$$

for some constants $p_1,\ldots,p_{16} > 0$ with $\sum_{j=1}^{16} p_j = 1$. Suppose we also know that $X_j/1207 \approx p_j$ for all $1 \leq j \leq 16$.

Using a Taylor expansion of the function $h(a) = a\log(a/b)$, show that $Z$ is approximately equal to Pearson's chi-squared statistic

$$\sum_{j=1}^{16} \frac{(X_j - 1207p_j)^2}{1207p_j}$$

*Solution.* We use a Taylor expansion round $b > 0$ for $h(a) := a\log(a/b)$, we have $h(b) = 0$, $h'(b) = 1$ and $h''(b) = 1/b$, so

$$a\log(a/b) \approx (a-b) + \frac{1}{2b}(a-b)^2.$$

Substituting into the above with $a = X_j/1207$ and $b = p_j$), we get

$$2\log\frac{\sup_{\theta\in\Theta_0^c} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \approx 2\cdot 1207\sum_{j=1}^{16}\left[\left(\frac{X_j}{1207} - p_j(\lambda)\right) + \frac{1}{2}\frac{\left(\frac{X_j}{1207} - p_j(\lambda)\right)^2}{p_j(\lambda)}\right].$$

The first term in the sum is zero since $\sum_{j=1}^{16} X_j = 1207$ and $\sum_{j=1}^{16} p_j(\lambda) = 1$. So,

$$2\log\frac{\sup_{\theta\in\Theta_0^c} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)} \approx 1207\sum_{j=1}^{16}\frac{\left(\frac{X_j}{1207} - p_j(\lambda)\right)^2}{p_j(\lambda)} = \sum_{j=1}^{16}\frac{(X_j - 1207p_j(\lambda))^2}{1207p_j(\lambda)}.$$

## 5. Question 5

Let $X_1, X_2, X_3$ be i.i.d. continuous random variables. Let $W_1, W_2, W_3$ be a bootstrap sample from $X_1, X_2, X_3$. Let $Y$ denote the sample median of $X_1, X_2, X_3$. (That is, $Y$ is the middle value among $X_1, X_2, X_3$, which is unique with probability one since the random variables are continuous.)

- Describe the distribution of $(W_{(1)}, W_{(2)}, W_{(3)})$.
- Describe the bootstrap estimator of $Y$.

Solution. Since $X_1, X_2, X_3$ are all distinct with probability one, we have

$$\mathbf{P}(W_1 = X_i, W_2 = X_j, W_3 = X_k \mid X_1, X_2, X_3) = (1/3)^3, \quad \forall i \neq j, j \neq k \leq 3.$$

That is, in describing the distribution of $(W_{(1)}, W_{(2)}, W_{(3)})$, we may as well assume that $X_{(1)} = 1, X_{(2)} = 2, X_{(3)} = 3$, and $W_1, W_2, W_3$ are i.i.d. uniform in $\{1, 2, 3\}$. (We are satisfied with this description of the distribution of $(W_{(1)}, W_{(2)}, W_{(3)})$.)

Now, as covered e.g. in Exercise 2.19 in the notes, by considering $Y$ which is the number of indices $1 \leq j \leq 3$ such that $W_j \leq X_{(i)}$, we have

$$\mathbf{P}(W_{(2)} \leq X_{(i)} \mid X_1, X_2, X_3) = \sum_{k=2}^{3}\binom{3}{k}p_i^k(1-p_i)^{n-k},$$

where $p_i = i/3$ for all $1 \leq i \leq 3$. (This follows since $Y$ is a binomial random variable with parameters 3 and $p_i$.) That is,

$$\mathbf{P}(W_{(2)} \leq X_{(i)} \mid X_1, X_2, X_3) = \sum_{k=2}^{3}\binom{3}{k}(i/3)^k(1-i/3)^{n-k},$$

Therefore, for all $1 \leq i \leq 3$, we have

$$\mathbf{P}(W_{(2)} = X_{(i)} \mid X_1, X_2, X_3) = \mathbf{P}(W_{(2)} \leq X_{(i)} \mid X_1, X_2, X_3) - \mathbf{P}(W_{(2)} \leq X_{(i-1)} \mid X_1, X_2, X_3)$$
$$= \sum_{k=2}^{3} \binom{3}{k} (i/3)^k (1 - i/3)^{n-k} - \left( \sum_{k=2}^{3} \binom{3}{k} ((i-1)/3)^k (1 - (i-1)/3)^{n-k} \right).$$

The bootstrap estimator of $Y$ is then

$$\mathbf{E}[W_{(2)} \mid X_1, X_2, X_3] = \sum_{i=1}^{3} X_{(i)} \mathbf{P}(W_{(2)} = X_{(i)} \mid X_1, X_2, X_3)$$

$$= \sum_{i=1}^{3} X_{(i)} \left( \sum_{k=2}^{3} \binom{3}{k} (i/3)^k (1 - i/3)^{n-k} - \left( \sum_{k=2}^{3} \binom{3}{k} ((i-1)/3)^k (1 - (i-1)/3)^{n-k} \right) \right).$$