541B Midterm 2 Solutions[1]

1. Question 1

(a) Let X be a Gaussian random variable with unknown mean $\mu_X \in \mathbf{R}$ and variance 1, so that X has the following density (with $\sigma = 1$)

$$f_{\mu_X}(z) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-(z-\mu_X)^2/(2\sigma^2)\Big), \qquad \forall\, z \in \mathbf{R}.$$

Let Y be a Gaussian random variable with unknown mean $\mu_Y \in \mathbf{R}$ and variance 1.

Assume that X and Y are independent. Let $\alpha \in (0,1)$.

Find a $100(1-\alpha)\%$ confidence interval for $\mu_X - \mu_Y$.

Simplify the expression to a reasonable extent.

Solution. Since $X - Y - (\mu_X - \mu_Y)$ is a Gaussian random variable with mean 0 and variance 2, we have, for any $-\infty < a < b < \infty$,

$$\mathbf{P}(a \le X - Y - (\mu_X - \mu_Y) \le b) = \int_a^b \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-z^2/4} dz.$$

That is,

$$\mathbf{P}(X - Y - b \le \mu_X - \mu_Y \le X - Y - a) = \int_a^b \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-z^2/4} dz.$$

Choosing a, b such that $\int_a^b \frac{1}{\sqrt{2}\sqrt{2\pi}} e^{-z^2/4} dz = 1 - \alpha$ completes the proof.

2. Question 2

(a) Complete the statement of the following Theorem from the notes about the limiting distribution of the GLR statistic:

Let $X := (X_1, ..., X_n)$ be a random sample of size n from a family of distributions $\{f_\theta \colon \theta \in \Theta\}$. Fix $\theta_0 \in \Theta \subseteq \mathbf{R}$. Suppose we test the hypothesis H_0 that $\{\theta = \theta_0\}$ versus the alternative $\{\theta \neq \theta_0\}$. Suppose the assumptions of Theorem 2.69 from the notes hold. Let $\lambda(X) := \frac{\sup_{\theta\in\Theta} f_\theta(X)}{\sup_{\theta\in\Theta_0} f_\theta(X)}$ denote the generalized likelihood ratio statistic. If H_0 is true, then as $n \to \infty$:

Solution: $2\log\lambda(X)$ converges in distribution to a chi squared random variable with one degree of freedom.

(b) In the alpha particle examples from class and the homework, we had a table with 16 entries of the following form, corresponding to the results of 1207 independent experiments

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

What did we conclude about the distribution of the random variable

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

Solution. We conclude that its distribution is approximately a chi squared distribution with $14 = 16 - 1 - 1$ degrees of freedom.

(c) Are the conclusions of parts (a) and (b) above different? Explain **in detail** why the conclusions of parts (a) and (b) are the same, or not the same.

[1]November 2, 2026, © 2026 Steven Heilman, All Rights Reserved.

Solution. Parts (a) and (b) have different conclusions. In part (a), the sample size n is large (and goes to infinity). In contrast, in part (b), the sample size (which is the number of columns in the table) is 16, which is nowhere close to infinity. However, the number of independent experiments (1207) is a fairly large number. Parts (a) and (b) are therefore different due to the different large parameters that are present (number of samples being large versus small). As a minor point, the entries in the table in part (b) are not independent, since their sum is constrained to be 1207.

3. Question 3

Let $(x_1, y_1), \ldots, (x_n, y_n)$ in $\mathbf{R} \times \mathbf{R}$.
For any $m, b \in \mathbf{R}$, define

$$f(m, b) := \sum_{i=1}^{n} (y_i - mx_i - b)^2.$$

Using the second derivative test, show that the minimum value of f is achieved when

$$m = \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{j=1}^n y_j\right) - n\left(\sum_{k=1}^n x_k y_k\right)}{\left(\sum_{i=1}^n x_i\right)^2 - n\left(\sum_{j=1}^n x_j^2\right)} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{j=1}^n (x_j - \overline{x})^2}.$$

$$b = \frac{1}{n}\left(\sum_{i=1}^n y_i - m\sum_{j=1}^n x_j\right) = \overline{y} - m\overline{x}.$$

Briefly explain why this is actually the minimum value of $f(m, b)$. (You are allowed to use the inequality $(\sum_{i=1}^n x_i)^2 \leq n(\sum_{i=1}^n x_i^2)$.)

Solution.

Let $X = \sum_{i=1}^n x_i$, $Y = \sum_{j=1}^n y_j$, $A = \sum_{k=1}^n x_k y_k$, $S = \sum_{j=1}^n x_j^2$. We have $\nabla f(m, b) = (\sum_{i=1}^n 2(y_i - mx_i - b)(-x_i), -\sum_{i=1}^n 2(y_i - mx_i - b))$. So, $\nabla f(m, b) = (0, 0)$ when $\sum_{i=1}^n (y_i - mx_i - b)) = 0$ and $\sum_{i=1}^n (y_i - mx_i - b)x_i = 0$. We want to solve for m and b. We have $Y - mX - nb = 0$ and $A - mS - bX = 0$. So, $b = (Y - mX)/n$, and then $A - mS - (Y - mX)X/n = 0$. So, $m(-S + X^2/n) = -A + XY/n$, and $m = (XY - An)/(X^2 - Sn)$.

Finally, strict convexity of f implies that the only critical point we found is the global minimum of f. (If $x_1 = \cdots = x_n$, then the minimum will not be unique. For example, if $n = 1$, then the function $f(m, b) = (1 - m - b)^2$ has a minimum on the line $m + b = 1$.)

(Instead of using strict convexity, we could also examine the Hessian of the function f.)

4. Question 4

Let

$$h(t) = \frac{1}{1 + e^{-t}}, \qquad \forall\, t \in \mathbf{R}.$$

In logistic regression, we are given independent random variables $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ and we assume there exists $a, b \in \mathbf{R}$ such that

$$P(Y_i = 1) = h(aX_i + b), \qquad \mathbf{P}(Y_i = 0) = 1 - \mathbf{P}(Y_i = 1), \qquad \forall\, i \geq 1.$$

Suppose $n = 2$ and

$$(X_1, Y_1) = (-1, 0), \qquad (X_2, Y_2) = (1, 1).$$

(a) Write an expression for the log-likelihood $\ell(a, b)$. Simplify it to a reasonable extent.

(b) Verify that $\frac{d^2}{dt^2} \log h(t) < 0$ for all $t \in \mathbf{R}$.

(c) Does ℓ have a maximum value in the plane? Explain why your conclusion does not contradict with part (b).

Solution.

$$\ell(a, b) = \log \prod_{i=1}^{2} h(aX_i + b)^{Y_i} [1 - h(aX_i + b)]^{1-Y_i} = \log(1 - h(b - a)) + \log h(a + b).$$

Using $1 - h(t) = -h(-t)$, we get the slightly simpler expression

$$\ell(a, b) = \log h(a - b) + \log h(a + b).$$

The function $g(t) := \log h(t)$ satisfies

$$g''(t) = -h(t)(1 - h(t)) < 0.$$

As stated in the problem, this implies that ℓ is strictly concave.

Nevertheless,

$$\ell(a, 0) = 2 \log h(a) \longrightarrow 0 \quad \text{as } a \to \infty,$$

while $\ell(a, b) < 0$ for every $(a, b) \in \mathbf{R}$. Hence no maximum value of ℓ exists. Strictly concave functions are not guaranteed to have global maxima.

5. Question 5

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

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

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,

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