

# 1 GR6411 - First half: exam solutions

**Q1 Answer:** Since the potential outcomes are discrete, we can use

$$\mathbb{P}(Y_i = y, D_i = d)$$

to denote the probability that the outcome of unit  $i$  equals  $y$  and the treatment status  $D_i$  equals  $d \in \{0, 1\}$ . By definition of  $Y_i$ , any unit  $i \in \{2, \dots, N-1\}$  satisfies:

$$\begin{aligned} \mathbb{P}(Y_i = y, D_i = 1) &= \mathbb{P}(Y_i(1) = y, D_i = 1) \\ &= P(Y_i(1) = y) \cdot p, \end{aligned}$$

where the last line uses the independence between treatments and potential outcomes. Analogously, for any such unit

$$\begin{aligned} \mathbb{P}(Y_i = y, D_i = 0) &= \mathbb{P}(Y_i(0) = y, D_i = 0) \\ &= P(Y_i(0) = y) \cdot (1 - p). \end{aligned}$$

This means we that for any  $i \in \{2, \dots, N-1\}$

$$\mathbb{P}(Y_i = y, D_i = d) = (P(Y_i(1) = y) \cdot p)^d (P(Y_i(0) = y) \cdot (1 - p))^{1-d}$$

Since the first unit is always treated, and the last unit never is, the likelihood function of  $(Y, d)$  equals

$$\mathcal{L}(Y, d|P) \equiv P(Y_1(1) = y_1) \left( \prod_{i=2}^{N-1} \mathbb{P}(Y_i = y_i, D_i = d_i) \right) P(Y_N(0) = y_N).$$

Consequently, the likelihood equals

$$\mathcal{L}(Y, D|P) = \left( \prod_{i=1}^N P(Y_i(1) = y_i)^{d_i} P(Y_i(0) = y_i)^{1-d_i} \right) \left( p^{\sum_{i=2}^{N-1} d_i} (1-p)^{(N-2) - \sum_{i=2}^{N-1} d_i} \right). \quad (1.1)$$

**Q2 answer:** Take two different distributions  $P, P'$ , with the same marginals over potential outcomes. Equation (1.1) shows that

$$\mathcal{L}(Y, d|P) = \mathcal{L}(Y, d|P'),$$

for any  $(Y, d)$ . Thus, there exists  $P \neq P'$  that induce the same distribution over the data. This

means that  $P$  is not identified.

To show that the marginals are identified, take any two pairs of distributions  $\theta \equiv (P_1, P_0)$ ,  $\theta' \equiv (P'_1, P'_0)$ ,  $\theta \neq \theta'$ . We want to argue that  $\theta$  and  $\theta'$  induce different distributions over the data. To see this, assume that there is  $y$  such that  $P_1(Y_i(1) = y) \neq P'_1(Y_i(1) = y)$ . Then,

$$\mathbb{P}_\theta(Y_i = y, d = 1) = P(Y_i(1) = y)p \neq P'(Y_i(1) = y)p = \mathbb{P}_{\theta'}(Y_i = y, d = 1).$$

**Q3 Answer**<sup>1</sup>: Let  $\mathcal{Y}_1 = \{y_1^{(1)}, y_1^{(2)}, \dots, y_1^{(J_1)}\}$  and  $\mathcal{Y}_0 = \{y_0^{(1)}, y_0^{(2)}, \dots, y_0^{(J_0)}\}$  denote the support of the marginals of potential outcomes, for treatment and control respectively. From the answer above, the parameters of the statistical model are

$$\begin{aligned} \mathbf{q}_1 &= q_1^{(1)}, q_1^{(2)}, \dots, q_1^{(J_1)} \\ \mathbf{q}_0 &= q_0^{(1)}, q_0^{(2)}, \dots, q_0^{(J_0)}, \end{aligned}$$

where  $q_d^{(j)} = \mathbb{P}(Y_i = y^{(j)}, d_i = d)$ . For a given sample, we denote the number of times potential outcome  $j$  shows up under the treatment as

$$n_1^{(j)}(Y, D) = \#\{i \in \{1, 2, \dots, N\} : y_i = y_1^{(j)}; d_i = 1\}$$

for  $j \in \{1, 2, \dots, J_1\}$ . Similarly,

$$n_0^{(j)}(Y, D) = \#\{i \in \{1, 2, \dots, N\} : y_i = y_0^{(j)}; d_i = 0\}$$

for  $j \in \{1, 2, \dots, J_0\}$  counts the number of times outcome  $j$  happens in the control group. Above, we emphasize the dependence of  $n_d^{(j)}$  on data because this means they are random variables. This is something that many of you got wrong!

Note that with the above notation, log-likelihood of the data can be expressed in the following way (dropping dependence on data for notational convenience):

$$\ln \mathcal{L}(Y, D|P) = \sum_{j=1}^{J_1} n_1^{(j)} \ln q_1^{(j)} + \sum_{j=1}^{J_0} n_0^{(j)} \ln q_0^{(j)} + \text{stuff that only depends on } p$$

---

<sup>1</sup>The proofs here are partly inspired by the solutions given by Nicolas and Lucas.

The log-likelihood maximization problem is thus

$$\begin{aligned}
& \max_{\mathbf{q}_0, \mathbf{q}_1} \quad \sum_{j=1}^{J_1} n_1^{(j)} \ln q_1^{(j)} + \sum_{j=1}^{J_0} n_0^{(j)} \ln q_0^{(j)} \\
& \text{s.t.} \quad q_d^{(j)} \geq 0 \quad \forall d \in \{0, 1\}, \forall j \in \{1, \dots, J_d\} \\
& \quad \sum_{j=1}^{J_d} q_d^{(j)} = 1 \quad \forall d \in \{0, 1\}
\end{aligned}$$

It should be clear from the above that solutions won't depend  $p$  (the probability of being treated). Also, the problem above is separable in  $\mathbf{q}_0$  and  $\mathbf{q}_1$ . Because of that we specialize the claim below to the treatment group, but everything goes through for  $\mathbf{q}_0$ .

**Claim 1.** *The likelihood is maximized by setting the probability of  $y_1^{(j)}$ , and  $j = 1, \dots, J_1$  to its relative frequency. That is,*

$$\hat{q}_1^{(j)} = \frac{n_1^{(j)}(Y, D)}{n_1(D)}$$

where  $n_1(D)$  is the total number of observations for which  $d_i = 1$ . That is,

$$n_1(D) = \# \{i : d_i = 1\} = \sum_{j=1}^{J_1} n_1^{(j)}(Y, D)$$

We will give three proofs of this claim, but before that note that any solution to the problem will have  $\hat{q}_1^{(k)} = 0$  if potential outcome  $k$  never shows up. I.e.,

$$n_1^{(k)} = 0 \implies \hat{q}_1^{(k)} = 0$$

We would be able to improve the likelihood by moving all the mass from those  $k$  to some  $j$  such that  $n_1^{(j)} > 0$ .<sup>2</sup>

Conversely, if  $n_1^{(k)} > 0$ , the MLE will feature  $\hat{q}_1^{(k)} > 0$  – otherwise the log-likelihood will be minus infinity, which is improved upon by, for example, assigning point mass to a single outcome that shows up.

*Proof 1.* From the concavity of log, we have for  $x > 0$  that

$$\ln(x) \leq x - 1 \tag{1.2}$$

---

<sup>2</sup>Note that we can always do that due to the assumption that at least one observation shows up in control and treatment groups.

Define  $\hat{q}_1^{(j)} = n_1^{(j)}/n_1$ . Take any other candidate for MLE  $\tilde{\mathbf{p}}_1$  and note that

$$\begin{aligned} \sum_{j=1}^{J_1} n_1^{(j)} \ln \left[ \frac{\tilde{q}_1^{(j)}}{\hat{q}_1^{(j)}} \right] &\leq \sum_{j=1}^{J_1} n_1^{(j)} \left[ \frac{\tilde{q}_1^{(j)}}{\hat{q}_1^{(j)}} - 1 \right] = \sum_{j=1}^{J_1} [n_1 \tilde{q}_1^{(j)} - n_1^{(j)}] \\ &= n_1 \sum_{j=1}^{J_1} [\tilde{q}_1^{(j)} - 1] = 0 \end{aligned}$$

Thus  $\sum_{j=1}^{J_1} n_1^{(j)} \ln \tilde{q}_1^{(j)} \leq \sum_{j=1}^{J_1} n_1^{(j)} \ln \hat{q}_1^{(j)}$  as we wanted to show. □

*Proof 2.* We begin by showing that

$$\frac{\hat{q}_1^{(j)}}{n_1^{(j)}} = \frac{\hat{q}_1^{(j')}}{n_1^{(j)'}}$$

must hold at an MLE whenever  $n_1^{(j)} > 0$  and  $n_1^{(j')} > 0$ . Indeed, suppose otherwise; take any  $\hat{\mathbf{q}}_1$  such that  $\sum_{j=1}^{J_1} \hat{q}_1^{(j)} = 1$ . Without loss of generality suppose the first two outcomes satisfy,  $n_1^{(1)} > 0$ ,  $n_1^{(2)} > 0$  but

$$\frac{\hat{q}_1^{(1)}}{n_1^{(1)}} \neq \frac{\hat{q}_1^{(2)}}{n_1^{(2)}}$$

Define  $\tilde{q}_1^{(j)}$  in the following way. For all  $j \notin \{1, 2\}$ , set  $\tilde{q}_1^{(j)} = \hat{q}_1^{(j)}$ . Define

$$\tilde{q}_1^{(1)} = \frac{n_1^{(1)}}{n_1^{(1)} + n_1^{(2)}} (\hat{q}_1^{(1)} + \hat{q}_1^{(2)})$$

$$\tilde{q}_1^{(2)} = \frac{n_1^{(2)}}{n_1^{(1)} + n_1^{(2)}} (\hat{q}_1^{(1)} + \hat{q}_1^{(2)})$$

This way,  $\sum_{j=1}^{J_1} \tilde{q}_1^{(j)} = 1$  so it is a valid estimator. Also,  $\tilde{q}_1^{(1)}/n_1^{(1)} = \tilde{q}_1^{(2)}/n_1^{(2)}$ .

Moreover, by Jensen's inequality,

$$n_d^{(1)} \ln \left( \frac{\hat{q}_1^{(1)}}{n_1^{(1)}} \right) + n_d^{(2)} \ln \left( \frac{\hat{q}_1^{(2)}}{n_1^{(2)}} \right) < (n_1^{(1)} + n_1^{(2)}) \ln \left( \frac{\tilde{q}_1^{(1)}}{n_1^{(1)}} \right) \quad (1.3)$$

Adding and subtracting  $\sum_j n_1^{(j)} \ln n_1^{(j)}$  to the log likelihood, and applying (1.3), we get

$$\begin{aligned} \sum_{j=1}^{J_1} n_1^{(j)} \ln \hat{q}_1^{(j)} &= \sum_{j=1}^{J_1} n_1^{(j)} \ln n_1^{(j)} + \sum_{j=1}^{J_1} n_1^{(j)} \ln \left( \frac{\hat{q}_1^{(j)}}{n_1^{(j)}} \right) \\ &< \sum_{j=1}^{J_1} n_1^{(j)} \ln n_1^{(j)} + \sum_{j=1}^{J_1} n_1^{(j)} \ln \left( \frac{\tilde{q}_1^{(j)}}{n_1^{(j)}} \right) \\ &= \sum_{j=1}^{J_1} n_1^{(j)} \ln \tilde{q}_1^{(j)} \end{aligned}$$

whence  $\hat{\mathbf{p}}_1$  can't be an MLE. This establishes that at any MLE, there exists a constant  $\eta$  such that  $\hat{q}_1^{(j)}/n_1^{(j)} = \eta$  for every  $j$ . Adding over  $j$ , we get

$$1 = \sum_{j=1}^{J_1} \hat{q}_1^{(j)} = \eta \sum_{j=1}^{J_1} n_1^{(j)} = \eta n_1$$

therefore  $\hat{q}_1^{(j)} = \frac{n_1^{(j)}}{n_1}$ . □

*Proof 3.* Form the Lagrangian:

$$\Lambda = \sum_{j=1}^{J_1} n_1^{(j)} \ln q_1^{(j)} + \lambda_1 \left[ 1 - \sum_{j=1}^{J_1} q_1^{(j)} \right]$$

Given the above discussion, we can restrict our search to  $n_1^{(j)} > 0$ , and discard corner solutions in this restricted set. The first order condition with respect to such  $j$  (e.g, in the treatment group) is

$$n_1^{(j)} = \lambda_1 \hat{q}_1^{(j)}$$

for  $n_1^{(j)} > 0$ . Summing up over all  $(j)$ , we get  $n_1 = \lambda_1$ . Hence  $\hat{q}_1^{(j)} = n_1^{(j)}/n_1$ .

Because the objective is a strictly concave function and the constraint is a compact convex set, the first order condition is sufficient. □

**Q4 Answer:** This estimator is unbiased, independently of the value of  $p \in (0, 1)$ . We show first that

$$\begin{aligned}
\mathbb{E} \left[ \frac{1}{n_1} \sum_{i \in \{i | d_i = 1\}} Y_i \right] &= \mathbb{E} \left[ \frac{1}{n_1} \sum_{i=1}^N Y_i 1\{d_i = 1\} \right] \\
&= \mathbb{E} \left[ \mathbb{E} \left[ \frac{1}{n_1} \sum_{i=1}^N Y_i 1\{d_i = 1\} | d_1, \dots, d_N \right] \right] \\
&\quad \text{(by the LIE)} \\
&= \mathbb{E} \left[ \mathbb{E} \left[ \frac{1}{n_1} \sum_{i=1}^N Y_i(1) 1\{d_i = 1\} | d_1, \dots, d_N \right] \right] \\
&\quad \text{(by definition of } Y_i) \\
&= \mathbb{E} \left[ \left[ \frac{1}{n_1} \sum_{i=1}^N \mathbb{E}[Y_i(1) | d_1, \dots, d_N] \cdot 1\{d_i = 1\} \right] \right] \\
&\quad \text{(since } n_1 \text{ and } d_i \text{ are constant given the conditioning set)} \\
&= \mathbb{E} \left[ \left[ \frac{1}{n_1} \sum_{i=1}^N \mathbb{E}[Y_i(1) | d_i] \cdot 1\{d_i = 1\} \right] \right] \\
&\quad \text{(by independence across units)} \\
&= \mathbb{E} \left[ \left[ \frac{1}{n_1} \sum_{i=1}^N \mathbb{E}[Y_i(1)] \cdot 1\{d_i = 1\} \right] \right] \\
&\quad \text{(by the RCT assumption)} \\
&= \mathbb{E}[Y_i(1)].
\end{aligned}$$

The proof for  $\mathbb{E}_{P_0}[Y_i(0)]$  is analogous.

**Q5 Answer:** Let  $\pi(\theta)$  be the researcher's prior. The statistical model is

$$\hat{\theta} | \theta \sim \mathcal{N}(\theta, D)$$

so that given  $\pi$ , we can find the posterior density by Bayes rule, (assuming all  $\sigma^2$  are positive and  $\pi$  absolutely continuous)

$$f_\pi(\theta | \hat{\theta}) \propto \varphi(\hat{\theta} - \theta; D) \pi(\theta)$$

where  $\varphi(\cdot; D)$  denotes the density of a multivariate normal with mean zero and variance matrix  $D$ .

The posterior utility associated with action  $a_{ab}$  is then

$$\begin{aligned}\tilde{U}(a_{ab}) &= \mathbb{E}[U(a_{ab}, \theta) | \hat{\theta}] = \int U(a_{ab}, \theta) f_{\pi}(\theta | \hat{\theta}) d\theta \\ &= \int \theta_{ab} f_{\pi}(\theta | \hat{\theta}) d\theta \\ &= \mathbb{E}[\theta_{ab} | \hat{\theta}]\end{aligned}$$

where the third equality follows from the given utility specification. A Bayes decision rule in this setting (given  $\pi$ ) is any

$$(a^*, b^*) \in \operatorname{argmax}_{a \in \{0,1\}, b \in \{0,1\}} \mathbb{E}[\theta_{ab} | \hat{\theta}]$$

where  $\mathbb{E}[\theta_{00} | \hat{\theta}] = 0$ .