Felix Céard-Falkenberg
7174020

Problem 1

(a) *Express $P(X_i = 1)$ as a function of a and π.*

We have

$$\begin{aligned} P(X_i = 1) &= P(X_i = 1, \text{Group} = A) + P(X_i = 1, \text{Group} = B) \\ &= P(X_i = 1 \mid \text{Group} = A) \cdot P(\text{Group} = A) + P(X_i = 1 \mid \text{Group} = B) \cdot P(\text{Group} = B) \\ &= a \cdot P(\text{Group} = A) + (1-a) \cdot P(\text{Group} = B) \\ &= a \cdot \pi + (1-a) \cdot (1-\pi) \ . \end{aligned}$$

Naturally, we have that

$$\begin{aligned} P(X_i = 0) &= P(X_i = 0, \text{Group} = A) + P(X_i = 0, \text{Group} = B) \\ &= P(X_i = 0 \mid \text{Group} = A) \cdot P(\text{Group} = A) + P(X_i = 0 \mid \text{Group} = B) \cdot P(\text{Group} = B) \\ &= (1-a) \cdot P(\text{Group} = A) + a \cdot P(\text{Group} = B) \\ &= (1-a) \cdot \pi + a \cdot (1-\pi) \ . \end{aligned}$$

(Optional) Because the random variable X_i can only have 2 potential values, we can now easily check whether our expressions are correct as $P(X_i = 1) + P(X_i = 0) = 1$. We have

$$\begin{aligned} P(X_i = 1) + P(X_i = 0) &= a \cdot \pi + (1-a) \cdot (1-\pi) + (1-a) \cdot \pi + a \cdot (1-\pi) \\ &= a \cdot \pi + (1-\pi) - a + \pi \cdot a + \pi - \pi \cdot a + a - \pi \cdot a \\ &= a \cdot \pi + 1 - \pi + \pi - \pi \cdot a \\ &= 1 \ . \end{aligned}$$

(b) *Give expressions for the joint probability of the observed data $X_1, X_2, \ldots, X_n$, and the corresponding likelihood function and loglikelihood function.*

We aim to express the computation for

$$P(X_1, X_2, \ldots, X_n) \ .$$

Because we know that the n samples are drawn with the i.i.d. assumption, it holds that

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i) \ .$$

Hence,

$$
\begin{aligned}
P(X_1, X_2, \ldots, X_n) &= \prod_{i=1}^{n} P(X_i) \\
&= \prod_{i=1}^{n} \begin{cases} P(X_i = 1) & \text{if } X_i = 1 \\ P(X_i = 0) & \text{if } X_i = 0 \end{cases} \\
&= P(X_i = 1)^m \cdot P(X_i = 0)^{n-m} \\
&= (a \cdot \pi + (1-a) \cdot (1-\pi))^m \cdot ((1-a) \cdot \pi + a \cdot (1-\pi))^{n-m}
\end{aligned}
$$

which is the likelihood function of the observed samples.

Let us now derive the experession for the log likelihood of the data:

$$
\begin{aligned}
\log P(X_1, X_2, \ldots, X_n) &= \log \left[(a \cdot \pi + (1-a) \cdot (1-\pi))^m \cdot ((1-a) \cdot \pi + a \cdot (1-\pi))^{n-m} \right] \\
&= m \cdot \log \left[a \cdot \pi + (1-a) \cdot (1-\pi) \right] + (n-m) \cdot \log \left[(1-a) \cdot \pi + a \cdot (1-\pi) \right] \qquad (1)
\end{aligned}
$$

(c) *Derive an expression for the maximum likelihood estimator π_{ML} of π.*

We now aim to estimate π. To this end, we replace π with π_{ML} in Equation 1 and derive the log likelihood w.r.t. π_{ML}. We have

$$
\begin{aligned}
&\frac{\partial}{\partial \pi_{ML}} \log P(X_1, X_2, \ldots, X_n) \\
=&\frac{\partial}{\partial \pi_{ML}} m \cdot \log \left[a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML}) \right] + (n-m) \cdot \log \left[(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML}) \right] \\
=&m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a - (1-a)) + (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot ((1-a) - a) \,.
\end{aligned}
$$

We can now find the critical points by setting the gradient to zero:

$$
\begin{aligned}
\frac{\partial}{\partial \pi_{ML}} \log P(X_1, X_2, \ldots, X_n) &\overset{!}{=} 0 \\
m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a - (1-a)) + (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot ((1-a) - a) &= 0 \qquad (2)
\end{aligned}
$$

where we observe that the the equation holds whenever $(1-a) = a$, because then both additive terms are zero. It is easy to see that $a = 0.5$ is the only real valued solution for this condition. However, as we are interested in the maximum likelihood estimator of π_{ML}, setting $a = 0.5$ prohibits us from finding the maximum value.

Explain why setting $a = 0.5$ is problematic in more details...

Therefore, we have to consider the case where $a \neq 0.5$.

Felix Céard-Falkenberg
7174020

$$\begin{aligned}
m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a-(1-a)) &= -(n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot ((1-a)-a) \\
m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a-(1-a)) &= (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot (a-(1-a)) \\
m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} &= (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \\
m \cdot \frac{1}{(2a-1) \cdot \pi_{ML} + 1 - a} &= (n-m) \cdot \frac{1}{(1-2a) \cdot \pi_{ML} + a} \\
m \cdot \frac{1}{(2a-1) \cdot \pi_{ML} + 1 - a} &= (n-m) \cdot \frac{1}{a - (2a-1) \cdot \pi_{ML}} \\
m \cdot (a - (2a-1) \cdot \pi_{ML}) &= (n-m) \cdot ((2a-1) \cdot \pi_{ML} + 1 - a) \\
am - (2a-1) \cdot m \cdot \pi_{ML} &= (n-m) \cdot (2a-1) \cdot \pi_{ML} + (n-m) \cdot (1-a) \\
am - (n-m) \cdot (1-a) &= (n-m) \cdot (2a-1) \cdot \pi_{ML} + (2a-1) \cdot m \cdot \pi_{ML} \\
am - (n-m) \cdot (1-a) &= \pi_{ML} \cdot (2a-1)((n-m)+m) \\
am - (n-m) + a \cdot n - am &= \pi_{ML} \cdot (2a-1)(n) \\
n \cdot (a-1) + m &= \pi_{ML} \cdot (2a-1)(n) \\
\frac{n \cdot (a-1) + m}{(2a-1)(n)} &= \pi_{ML} \\
\frac{n \cdot (a-1) + m}{n} \cdot \frac{1}{2a-1} &= \pi_{ML} \\
\left(a - 1 + \frac{m}{n}\right) \cdot \frac{1}{2a-1} &= \pi_{ML} \qquad (3)
\end{aligned}$$

We verify that this is indeed a maximum by plugging the result from Equation 3 into the second derivative:

$$\begin{aligned}
&\frac{\partial^2}{\partial {\pi_{ML}}^2} \log P(X_1, X_2, \ldots, X_n) \\
=&\frac{\partial}{\partial \pi_{ML}} m \cdot \frac{1}{a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML})} \cdot (a-(1-a)) + (n-m) \cdot \frac{1}{(1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML})} \cdot ((1-a)-a) \\
=&m \cdot (a-(1-a)) \cdot \frac{1}{(a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML}))^2} \cdot (a-(1-a)) \\
&+ (n-m) \cdot ((1-a)-a) \cdot \frac{1}{((1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML}))^2} \cdot ((1-a)-a) \\
=&2m \cdot (a-(1-a)) \cdot \frac{1}{(a \cdot \pi_{ML} + (1-a) \cdot (1-\pi_{ML}))^2} + 2(n-m) \cdot ((1-a)-a) \cdot \frac{1}{((1-a) \cdot \pi_{ML} + a \cdot (1-\pi_{ML}))^2}
\end{aligned}$$

This will be loooong I think...

(d) *Can you think of an easier way to arrive at the same estimator as the maximum likelihood estimator?*

(e) *Show that π_{ML} is an unbiased estimator of π.*

We aim to show that $\mathbb{E}[\pi_{ML}] = \pi$.

(f) *Derive an expression for the variance of π_{ML}. Analyse its dependence on a.*

(g) *Give an expression for the bias of this naive estimator. Is it asymptotically unbiased?*

(h) *Derive expressions for the mean squared error (MSE) of the randomized response estimator and the naive estimator. Analyse the expressions to draw conclusions about when, depending on the relevant parameters (π, a, ℓ, n), one estimator should be favored over the other.*

Problem 2

		$N = 500$	$N = 1000$	$N = 5000$
$\hat{N}$		590.45±311	1096.62±286	5152.14±583
Error (%)		18.09	9.66	3.04
$\hat{\sigma}_{\hat{N}}$		239.58±258	284.17±116	566.36±98
C.I.	$\alpha = 0.05$	[299.68±85; 1322.77±1298]	[687.34±137; 1842.05±609]	[4179.35±423; 6414.78±804]
	$\alpha = 0.01$	[250.36±59; 1740.30±2079]	[601.62±110; 2185.30±773]	[3922.46±383; 6884.52±890]

Table 1: Result from 1000 MC simulations for estimating $\hat{N}$. For each value, we report on the mean (black) and standard deviation (gray). We omit the digits after the decimal point for the standard deviation for brevity. The Error (%) is the absolute difference between the estimated and the true value of N, divided by the true value of N. Hence, it measures the relative error of the estimated value to the true value.

Observation. We implement the formulas and run the Monte-Carlo simulations for 1000 runs and report the results in Table 1. We first observe that, although the absolute error between the estimated and the true value of N increases with N, the relative error decreases when the true N is larger.

Interestingly, the given formula proposed for estimating the variance of $\hat{N}$ is very accurate, as its average value is very close to the computed variance of the Monte-Carlo simulation — in our case, we report on the standard deviation, which is simply the square root of the variance.

Lastly, we computed the averaged confidence intervals across all runs for $\alpha = 0.05$ and $\alpha = 0.01$. Naturally, the confidence intervals when using $\alpha = 0.01$ are larger than when using $\alpha = 0.05$, we notice that the standard deviation of the lower confidence decreases when increasing the size of the Confidence Interval (C.I.), whereas the standard deviation of the upper-confidence increases when decreasing α.

Discussion.

Statement on the Useage of Gen. AI

Unless stated otherwise, all code, text, and math derivations have been entirely thought and written by me with no external help — both for gen. AI and wolframalpha anc co.