PSYCH-GA.2211/NEURL-GA.2201 – Fall 2025
Mathematical Tools for Neural and Cognitive Science

Homework 5

Due: 25 Nov 2025
(late homeworks penalized 10% per day)

See the course web site for submission details. For each problem, show your work - if you only provide the answer, and it is wrong, then there is no way to assign partial credit! And, please don't procrastinate until the day before the due date... *start now*!

1. **Comparing two estimators**. A common method of estimating the size of biological populations is the "capture-mark-recapture" method. One proceeds by repeatedly capturing animals, putting a tag on them, and releasing them. After marking M animals, you then capture a new group of C animals and find that K of them are tagged. One assumes the C 2nd-round captures are *independent* of the previous captures (i.e., "sampling with replacement"). If the full population size is N, then the proportion of the population that is tagged is M/N, and thus the expected value of the proportion tagged in the 2nd-round captures (K/C) should be the same as the population proportion. Setting these equal and solving for N gives an estimate of the population size: $\hat{N} = MC/K$.

   (a) Check whether $\hat{N}$ is a maximum likelihood estimator. For a few triplets $\{K, C, M\}$, plot the likelihood $L(N) = p(K|C, M, N)$ for a range of values of N (e.g., from $\hat{N} - 5$ to $\hat{N} + 5$. For example, try $\{K, C, M\} = \{10, 200, 4000\}$. Note that N has to be an integer, so note whether $\hat{N}$ should be rounded or truncated to the nearest lower integer to maximize likelihood, or whether your results suggest a consistent pattern with regard to the non-integer $\hat{N}$.

   (b) Check whether the estimator is unbiased. For a few triplets $\{C, M, N\}$, simulate the 2nd capture value K 1,000 times (i.e., generate appropriately distributed samples of K), and compute the population estimate from each. Is the mean of the population estimates close to the correct answer?

   (c) Write the precise distribution for samples K (when $\{C, M, N\}$ are known and fixed), so you can compute the exact probability distribution of estimates $\hat{N}$. Do that for $N = 1000$, $M = 100$, $C = 100$. Plot the distribution and compute its mean and standard deviation. Does this calculation indicate that the estimator is biased? (Note: do this for the non-integer value of $\hat{N}$).

   (d) Some authors have suggested an alternative estimator: $\hat{N}' = \lfloor (M + 1)(C + 1)/(K + 1) \rfloor$, where $\lfloor \cdot \rfloor$ indicates truncation to the next lower integer. Repeat part (c) for this estimator and compare the bias and variance of this estimator to the original one.

2. **Bayesian inference of binomial proportions.** Poldrack (2006) published an influential attack on the practice of "reverse inference" in fMRI studies, i.e., inferring that a cognitive process was engaged on the basis of activation in some area. For instance, if Broca's area was found to be activated using standard fMRI statistical-contrast techniques, researchers might infer that the subjects were using language. In a search of the literature, Poldrack found that

Broca's area was reported activated in 90 out of 840 fMRI contrasts involving engagement of language, but this area was also active in 215 out of 2754 contrasts not involving language.

(a) Assume that the conditional probability of activation given language, as well as that of activation given no language, each follow a Bernoulli distribution (i.e., like coin-flipping), with parameters x_l and x_{nl}. Compute the likelihoods of these parameters, given Poldrack's observed frequencies of activation. Plot the likelihoods for `x=[0:.001:1]` as a bar chart.

(b) Find the value of x that maximizes each discretized likelihood function. Compare these to the exact maximum likelihood estimates given by the formula for the ML estimator of a Bernoulli probability.

(c) Using the likelihood functions computed for discrete x, compute and plot the discrete posterior distributions $P(x \mid \text{data})$ and the associated cumulative distributions $P(X \leq x \mid \text{data})$ for both processes. For this, assume a uniform prior $P(x) \propto 1$ and note that it will be necessary to compute the normalizing constant for Bayes' rule (i.e., $P(\text{data})$). Use the cumulative distributions to compute (discrete approximations to) upper and lower 95% confidence bounds on each proportion.

(d) *Are these frequencies different from one another?* Consider the joint posterior distribution over x_l and x_{nl}, the Bernoulli probability parameters for the language and non-language fMRI contrasts. Given that these two frequencies are independent, the (discrete) joint distribution is given by the outer product of the two marginals. Plot it (with `imagesc`). Compute (by summing the appropriate entries in the joint distribution) the posterior probabilities that $x_l > x_{nl}$ and, conversely, that $x_l \leq x_{nl}$.

(e) *Is this difference sufficient to support reverse inference?* Compute the probability $P(\text{language} \mid \text{activation})$. This is the probability that observing activation in Broca's area implies engagement of language processes. To do this use the estimates from part (b) as the relevant conditional probabilities, and assuming the prior that a contrast engages language, $P(\text{language}) = 0.5$. Hint: To calculate this probability, you will need to "marginalize", i.e., integrate over the unknown values of x_l and x_{nl}. Poldrack's critique said that we cannot simply conclude that activation in a given area indicates that a cognitive process was engaged without computing the posterior probability. Is this critique correct? To answer this, compute the posterior odds $\left(\frac{P(\text{language}|\text{activation})}{P(\text{not language}|\text{activation})}\right)$ using the maximum-likelihood estimates of x_l and x_{nl} from Poldrack's data and compare the posterior odds to the prior odds before running your experiment $\left(\frac{p(\text{language})}{p(\text{not language})}\right)$.