PSYCH-GA.2211/NEURL-GA.2201 – Fall 2021
Mathematical Tools for Neural and Cognitive Science

Homework 5

Due: 24 Nov 2021
(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. **Dueling estimators**. In this problem, we use simulation to compare three estimators of the mean of a Normal (Gaussian) distribution.

   (a). First consider the *average*, which minimizes the sum of squared deviations, and is also the Maximum Likelihood estimator. Generate 10,000 samples, each of size 10, from the Normal(0,1) distribution (a 10x10000 matrix). Compute the average of each of the 10,000 samples. Plot a histogram of the resulting estimates (use 50 bins, and set the plot range to [-2.3,2.3]). What shape should the histogram have (explain why)? What is the (theoretical) variance of the average of 10 values drawn from a univariate Gaussian (derive this)? Is the empirical variance of your 10,000 estimates close to this?

   (b). Now consider the *median*, which minimizes the sum of absolute deviations. Compute the median of each of the 10,000 samples, and again plot a histogram. What shape does this one have? Compare it to a normal distribution using the function `normplot`, which plots the quantiles of a sample of data versus the normal quantiles (known as a Q-Q plot: if data are normally distributed, the points shuld fall nearly on a straight line.) Does the distribution of estimated values deviate significantly from a Normal distribution? Specifically, compare the Q-Q plot for the median estimator to that for the mean from part (a).

   (c). Finally, consider an estimator that computes the average of the minimum and maximum over the sample (as shown in class, this one minimizes the $L_\infty-$norm). Again, compute this estimate for each of your 10,000 samples, plot the histogram, and examine and comment on the Q-Q plot, just as in part (b).

   (d). All three of these estimators are unbiased (because of the symmetry of the distribution), so we can use variance as the sole criterion for quality. Generate a new set of 10,000 samples, this time of dimension 256. Apply each estimator to sub-matrices of samples of size $\{8, 16, 32, 64, 128, 256\}$, and compute the variance of each estimator for each. Plot these (on a single log-log plot), along with a line showing the theoretically-computed variance of the average estimator. Does the variance of all three estimators converge at the same rate $(1/N)$? How much larger is the variance of the median estimator than the average estimator? How large a sample would you need for the average and median estimators to achieve the same variance as the average-extrema estimator (from part (c)) on samples of size 256?

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 103 out of 869 fMRI contrasts involving engagement of language, but this area was also active in 199 out of 2353 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. Compute these functions at the values `x=[0:.001:1]` and plot them 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 $P(x \mid data)$ and the associated cumulative distributions $P(X \leq x \mid data)$ for both processes. For this, assume a uniform prior $P(x) \propto 1$ and note that it will be necessary to compute (rather than ignore) the normalizing constant for Bayes' rule. 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 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 $P(language \mid activation)$. 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, $P(language) = 0.5$. 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, compare the Bayes factor ($\frac{p(\text{language}|\text{activation})}{p(\text{not language}|\text{activation})}$) using the maximum-likelihood estimates from Poldrack's data of activation probabilities, compared to the prior odds before running your experiment ($\frac{p(\text{language})}{p(\text{not language})}$).

3. **Simulating a 2AFC experiment.** Consider a two-alternative forced-choice psychophysical experiment (fancy name: heterochromatic brightness matching). Subjects are shown a blue spot and a red spot and must decide which appears brighter. The intensity of the blue spot is fixed, and that of the red spot is randomly varied over trials. The purpose of the experiment is to estimate the intensity of red that matches the blue. For a red spot of brightness I, the probability of the observer saying "The red spot is brighter" is:

$$p(I) = \lambda * \frac{1}{2} + (1 - \lambda) * \Phi(I; \mu, \sigma),$$

where $\Phi(I; \mu, \sigma)$ is the cumulative distribution function of the Gaussian (`normcdf` in matlab) with mean μ and standard deviation σ, evaluated at I. The parameter λ is called the "lapse rate" and is the proportion of trials the observer didn't pay attention and just guessed. The function $p(I)$ is known as the *psychometric function.*

(a) Plot two psychometric functions, for $\{\lambda, \mu, \sigma\}$ equal to $\{.05, 5, 2\}$ and $\{.05, 4, 3\}$ (use $I = [1 : 10]$). Describe the difference between these. If you increase μ, how does the curve change? If you increase σ, how does the curve change? (If you are not sure, make more plots with different parameter values.) What is the range of $p(I)$? Explain why this range is appropriate.

(b) Write a function `B=simpsych(lambda,mu,sigma,I,T)` to simulate an experiment. The arguments `(I,T)` are vectors of equal length, the first containing a list of intensities and the number of trials to be run for each corresponding intensity. The function should generate draws from $p(I)$, and returns a vector, B, (of the same length as I and T), containing the number of trials for which the simulated observer responded that the red spot was brighter, for each intensity I.

(c) Illustrate the use of `simpsych` with `T=ones(1,7)*100` and `I=1:7` for $\lambda = 0.05$, $\mu = 4$ and $\sigma = 1$. Plot `B ./ T` vs `I` (as points) and plot the psychometric function $p(I)$ (as a curve) on the same graph.

(d) Do the same with `T=ones(1,7)*10` and plot the results (including the psychometric function). What is the difference between this and the plot of the previous question?

(e) For each simulated dataset, assume you know that $\lambda = 0.05$ and $\sigma = 1$ and compute the likelihood of the data (or, more easily, its log) for values of μ ranging from 1 to 7 in steps of 0.1. What is the (approximate) maximum-likelihood estimate of μ?

(f) Next, assume you don't know the value of σ (but still know that $\lambda = 0.05$) and compute the likelihood of the data for a grid of (μ, σ) pairs, where μ varies as before, and σ ranges from 0.1 to 2.5 in steps of 0.1. What are the (approximate) maximum-likelihood estimates of μ and σ?

4. **Signal Detection Theory.** Consider an experiment where a moving-dot visual stimulus is presented to a subject. The difficulty of detecting the motion is varied by changing the *coherence* of the moving dots, which is the fraction of dots moving to the right (at zero coherence, the dots move randomly, and at 100% coherence, all of the dots move to the right). Suppose we want to decide whether the stimulus is random or is moving to the right, based on the response of a single neuron that fires at a random rate, whose mean is 5 spikes/s in response to a 0% coherence noisy stimulus and 8 spikes/s for 10% coherence. Suppose also that the distribution of firing rates is Gaussian with a standard deviation of 1 spikes/s for both stimuli.

(a) For the "no coherence"' stimulus, generate 1000 trials of the firing rate of the neuron in response to these stimuli (i.e., draw 1000 random samples from a Gaussian with $\mu = 5$ and $\sigma = 1$). Since we cannot have negative firing rates, set all rates that are below zero to zero. Now do the same thing for the 10% coherence stimulus. On the same figure, plot the histograms of the firing rates for each stimulus type.

(b) The success of the decoder (assuming this model of Gaussian noise) is determined by two things, the separation of the mean firing rates and the standard deviation of the neuron. From class, we know that this is captured in the measure known as d'. Calculate d' for this task and pair of stimuli (ignoring the fact that you are clipping firing rates at zero).

(c) Explain why the maximum likelihood decoder for this problem involves comparing the measurement to a threshold. For various thresholds t, calculate the hit and false-alarm rates using your sample data from (a), and plot these against each other (this is an ROC curve, defined in class). What threshold would you pick based on this curve to maximize the percentage-correct of the decoder, assuming that 0% and 10% coherence stimuli occur equally often. Plot this threshold as a point on the ROC curve and as a vertical line on your histogram from part (a). Next, suppose that 10% coherence stimuli occur 75% of the time. Determine and plot the threshold that maximizes percentage correct for this new prior.

(d) Consider now a neuron with a more "noisy" response so that the mean firing rates are the same but the standard deviation is 2 spikes/s instead of 1 spike/s. What is the new value of d'. Recompute and plot the optimal (maximum accuracy) thresholds for this noisy neuron for both the 50-50 and 75-25 priors. How do they differ from those in the previous part?