PSYCH-GA.2211/NEURL-GA.2201 – Fall 2018
Mathematical Tools for Neural and Cognitive Science

Homework 6

Due: 18 Dec 2018
(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. **Psychopathy.** You are interested in causes and treatment options for Psychopathy. You obtained a dataset, contained in the file `psychopathy.mat` obtained from a prison for violent offenders in upstate New York (not everyone in the prison is a psychopath, but they are more prevalent than in the general population). All study participants underwent a structural scan with a mobile, truck-mounted MRI. Each row of the matrix represents data from one prisoner. The first column contains the estimated cortical volume of paralimbic areas, relative to the population median, in cm^3. The second column contains the Hare Psychopathy Checklist (PCL-R) scores, which range from 0 to 40 (the higher the score, the more psychopathic traits someone exhibits). These scores are *not* distributed normally in either the general population (median = 4) or this prison subpopulation (median = 20). The third column indicates whether they already participated in an experimental treatment program known as decompression therapy (0 = did not yet participate, 1 = did already participate). To avoid self-selection effects, everyone in this dataset agreed to the therapy, but prisoners were randomly assigned to an earlier and a later treatment group, so that the untreated prisoners could serve as a control group.

   (a) Use polynomial regression to model PCL-R scores as a function of relative volume of paralimbic areas. (Note, you can use your code from HW2.) Use cross-validation to determine the best polynomial degree.

   (b) Use bootstrapping methods to estimate the 95% confidence interval of the average paralimbic volume of the decompression treatment group vs. the control group. If the random assignment worked, the confidence intervals should overlap. Do they? Also, do these data suggest that there is a statistically reliable difference from the general population, in terms of paralimbic volume?

   (c) Use a suitable t-test to compare the mean PCL-R score of prisoners who did and did not undergo decompression therapy. What is the p-value? Assuming an alpha-level of 0.05, is this difference significant? Can you reject the null hypothesis that decompression therapy is ineffective in terms of decreasing PCL-R scores?

   (d) Do a permutation test to assess whether decompression therapy has an effect. Designate an appropriate test statistic and calculate its p-value.

2. **Simulating a 2AFC experiment.** Consider a two-alternative forced choice (2AFC) psychophysical experiment in which a subject sees two stimulus arrays of some intensity on a trial and must say which one contains the target. (One and only one contains the target.) Her probability of being correct on a trial is:

$$p_c(I) = 1/2 + 1/2\Phi(I; \mu, \sigma)$$

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 function $p_c(I)$ is known as the *psychometric function*. (Minor note, somewhat subtle: This setup only makes sense if I is on a logarithmic scale, e.g., $I = k \log C$, where C is stimulus contrast.)

   (a) Plot two psychometric functions, for $\{\mu, \sigma\}$ equal to $\{5, 2\}$ and $\{4, 3\}$. (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 $p_c(I)$? Explain why this range is appropriate.

   (b) Write a function `C=simpsych(mu,sigma,I,T)` which takes two vectors `(I,T)` of the same length, containing a list of intensities and the number of trials for each intensity, respectively, simulates draws from $p_c(I)$, and returns a vector, C, of the same length as I and T, which contains the number of trials correct out of T, at each intensity I.

   (c) Illustrate the use of `simpsych` with `T=ones(1,7)*100` and `I=1:7` for $\mu = 4$ and $\sigma = 1$. Plot `C ./ T` vs `I` (as points) and plot the psychometric function $p_c(I)$ (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?

3. **Fitting a psychometric function.** Now we'll simulate the inverse (scientific) side of the problem, and use this probabilistic model as a means of fitting/analyzing a simulated data set.

   (a) Write a function `nll = nloglik(mu,sigma,I,T,C)` that returns the negative log likelihood of parameters `mu` and `sigma`, for data set `I,T,C` (we're negating it because we will be *minimizing* this function to solve for the optimal parameters).

   (b) Generate a contour plot (function `contour`, using 50 lines) of the negative log likelihood of the data set from part (c) of the previous problem, for all pairs of `mu` from `muall = [2:0.2:10]` and a `sigma` from `sigmaall = [0.5:0.2:6]`. What is the approximate location of the best fitting pair of parameters from this plot?

   (c) Use the function `fminsearch` to get a more precise estimate of values `mu,sigma` that minimize the function `nloglik(mu,sigma,)`. Two comments: first, the syntax for calling `nloglik` within `fminsearch` is a bit odd:
   `fminsearch(@(x) nloglik(x(1),x(2),I,T,C), <startpoint>).`
   Here, the `@` notation is used to create a temporary function, with argument `x` a vector containing the two variables being optimized (mean and stdev). Second, you'll need to specify a start point for the search – for this problem, `[2,2]` is a reasonable choice.

   (d) A variant of `fminsearch, fminunc`, also returns the *Hessian* (the matrix of second derivatives) of the negative log likelihood at the optimal `mu` and `sigma`. (Note: `fminunc` is less robust than `fminsearch`, and if the optimizer strays too far from the true values, there may be numerical problems due to overflow of the likelihood; in this case, try a

different starting point.) The inverse of the Hessian provides an estimate of the covariance matrix of the parameter estimates. Use this to determine 95% confidence intervals on each parameter (Hint: a 95% confidence interval is the mean ± 1.96 standard deviations of the parameter estimate. Compute the standard deviation of a marginal of the 2-D Gaussian that has covariance equal to the inverse Hessian.) Do the true parameter values (4 and 1) fall within these confidence intervals?

(e) Produce a second set of confidence intervals for the parameters using a bootstrap method. For each of the 7 intensities, resample 100 trials (correct or incorrect) from the 100 trials of that intensity in the original data, with replacement. Refit the model to the resampled data using `fminsearch`. Plot the histograms (function hist) of `mu` and `sigma` estimates obtained over 500 such resampled datasets, and define your confidence intervals as the region between the 2.5th and 97.5th percentiles of these distributions. How well do these values agree with those from the previous exercise?

4. **Comparing two psychometric functions.** Suppose we repeat the psychophysical experiment before and after giving the subject an experimental drug. Do the parameters change?

(a) Simulate the experiment using `simpsych` twice, once using the original parameters and again using the parameters `mu=5, sigma=1`. Fit each dataset using `fminsearch` to recover estimated parameters, and make note of the difference between the two estimates of `mu` and `sigma`.

(b) Now construct a permutation test of the null hypothesis (i.e., the hypothesis that there has been no change in the parameters). For each intensity, combine the 100 trials from each condition into a total of 200, then randomly partitioning this into two groups of 100. Fit both resampled datasets again, noting the difference between the two `mu`s and the two `sigma`s. Repeat this process 500 times to produce a null distribution of the differences in each parameter. How likely (at what quantile; one-tailed p-value) is the actual difference in `mu` from 3A ***? What about for `sigma`? Do these results make sense given the true parameter values from which you simulated the datasets?