STAT 345-01: Nonparametric Statistics

Problem Set 10

Assigned 2018 November 20
Due 2018 November 29

Show your work on all problems! Be sure to give credit to any collaborators, or outside sources used in solving the problems. Note that if using an outside source to do a calculation, you should use it as a reference for the method, and actually carry out the calculation yourself; it's not sufficient to quote the results of a calculation contained in an outside source.

Please hand in parts one and two separately. If you wish to submit your part one electronically, please send it directly to the grader as pdf only.

1 Part One

1.1 Conover Problems on 2×2 Contingency Tables

Exercise 4.1.8, part (a)

Followup: Repeat the calculation of the previous problem using the one-sided test for differences in probability of hiring for male and female candidates. Which test is appropriate if the university has 24 positions to fill? Suppose instead of hiring, the data referred to individual faculty members being promoted in rank. Which test would be appropriate then?

Problem 4.2.2

1.2 Conover Problems on the Median Test

Exercise 4.4.2

Exercise 4.4.6

2 Part Two

2.1 Project Status Report (one per team)

Submit a progress report on your proposal, including any partial results, and a rough draft of your report. (You'll get feedback on this, so the more you can submit, the better.)

2.2 Bayesian Approach

From a Bayesian perspective, the question of whether the row and/or column totals are held fixed when calculating a p-value is an irrelevant one, because Bayesian probabilities concern statements about a model given the actual observed data, not statements about what data might have been observed in a hypothetical repeated experiment. So for a 2×2 contingency table, if we write

$P(H_0|\{O_{ij}\})$ as the posterior probability of H_0 being true given the observations, it doesn't matter if we also condition on the row or column totals, since they are automatically given by the values in the table itself, e.g., $P(H_0|N,\{O_{ij}\}) = P(H_0|\{r_i\},\{O_{ij}\})P(H_0|\{O_{ij}\})$. But it turns out that the context of the experiment does still matter, because it defines the meaning of the hypotheses. One standard quantity in Bayesian hypothesis testing is the *Bayes factor* $p(\mathbf{x}|H_a)/p(\mathbf{x}|H_0)$ which measures how strongly the data favor the alternative hypothesis H_a over H_0. The "evidence" $p(\mathbf{x}|H)$ associated with hypothesis is like a sampling distribution, but it is appropriately averaged over possible parameter values according to a prescription included in H. Suppose the categorical observations $\mathbf{x} = \{(x_I, y_I)|I = 1,\dots,N\}$ are independent (which rules out the "lady tasting tea" scenario) so that the sampling distribution for the sequence of observations (which eliminates combinatorical factors which would cancel out anyway) is

$$p(\mathbf{x}|\{p_{ij}\}) = p_{11}^{O_{11}} p_{12}^{O_{12}} p_{21}^{O_{21}} p_{22}^{O_{22}}$$

Evaluate the following, both for general $\{O_{ij}\}$ (use $\{r_i\}$, $\{c_j\}$, and N as appropriate to simplify your answer), and for the example considered in class, where $O_{11} = 1$, $O_{12} = 6$, $O_{21} = 8$, and $O_{22} = 2$.

(a) The evidence

$$p(\mathbf{x}|H_0) = \int_0^1 \int_0^1 p(\mathbf{x}|\{p_{ij}\})\, dp_{1\bullet}\, dp_{\bullet 1}$$

for a model H_0 in which the probability for an observation to land in row i and column j is $p_{ij} = p_{i\bullet}\, p_{\bullet j}$, where $p_{2\bullet} = 1-p_{1\bullet}$ and $p_{\bullet 2} = 1-p_{\bullet 1}$, and the model assigns a uniform distribution to the parameters $p_{1\bullet}$ and $p_{\bullet 1}$. You may find the Beta function identity $\int_0^1 u^k(1-u)^\ell\, du = \frac{k!\ell!}{(k+\ell+1)!}$ useful.

(b) The evidence

$$p(\mathbf{x}|H_1) = \int_0^1 \int_0^1 \int_0^1 p(\mathbf{x}|\{p_{ij}\})\, dp_{1\bullet}\, dp_1^{(1)}\, dp_1^{(2)}$$

for a model H_1 in which the probability for an observation to land in row i and column j is $p_{ij} = p_{i\bullet}\, p_j^{(i)}$, where $p_{2\bullet} = 1 - p_{1\bullet}$, $p_2^{(i)} = 1 - p_1^{(i)}$, and the model assigns a uniform distribution to the parameters
$p_{1\bullet}$, $p_1^{(1)}$, and $p_1^{(2)}$.

(c) The evidence

$$p(\mathbf{x}|H_2) = 6\int_0^{1-p_{11}-p_{12}} \int_0^{1-p_{11}} \int_0^1 p(\mathbf{x}|\{p_{ij}\})\, dp_{11}\, dp_{12}\, dp_{21}$$

for a model H_2 in which any set of non-negative probabilities satisfying $p_{11}+p_{12}+p_{21}+p_{22} = 1$ is equally likely. You may find the identity
$\int_0^{1-u-v} \int_0^{1-u} 1 \int_0^1 u^k v^\ell w^m (1-u-v-w)^n\, du\, dv\, dw = \frac{k!\ell!m!n!}{(k+\ell+m+n+3)!}$ useful.

(d) The Bayes factor $p(\mathbf{x}|H_1)/p(\mathbf{x}|H_0)$, which is a measure of how much the data favor a model with row-dependent column probabilities over one with row-independent column probabilities.

(e) The Bayes factor $p(\mathbf{x}|H_2)/p(\mathbf{x}|H_0)$, which is a measure of how much the data favor a model of correlated categorical data over one of uncorrleated data.