Stat 574B: Bayesian Methods, HW 2

- Due Mon Oct 2. Make the presentation nice

Problem 1

Let us revisit the clinical trial accrual problem in which it required 12 months to recruit the first 9 patients. Given the posterior distribution in that analysis $p(\theta|y)$, we evaluated the predictive distribution for $\tilde{Y}$, the time needed to recruit the next 39 patients, assuming $\tilde{Y} \mid \theta \sim \text{Gamma}(39, \theta)$, which follows from assuming that the individual waiting times are exponentially distributed.

As mentioned in the course notes, there is an alternative way to look at this problem. In particular, if we define $\tilde{M}$ to be number of future patients recruited in the next 36 months, then using the relationship between Poisson and exponential waiting times, we can show that the distribution for $\tilde{M} \mid \theta$ implied by the Gamma model for $\tilde{Y} \mid \theta$ is that $\tilde{M} \mid \theta \sim \text{Poisson}(36\,\theta)$.

a. Based on the posterior in the notes, what is the PPD of $\tilde{M} \mid y$? Be explicit.

b. Plot the PPD of $\tilde{M} \mid y$, and compute relevant summaries of this distribution.

c. Compute $\Pr(\tilde{M} \geq 39 \mid y)$ and comment on the potential for completing the recruitment on time.

Problem 2

Assume that the joint distribution of (X, Y) is given by

$$p(x, y) \propto \binom{n}{x} y^{x+\alpha-1}(1-y)^{n-x+\beta-1},$$

where X takes on discrete values $x = 0, 1, ..., n$, and Y is a continuous random variable on [0,1]. The parameters $\alpha > 0$ and $\beta > 0$ are fixed.

(a) Show that X given $Y = y$ is Binomial(n, y) and that Y given $X = x$ is Beta$(\alpha + x, \beta + n - x)$. No formal derivations are needed here - just point out the functional forms for the kernels of the two conditional distributions.

(b) Devise a Gibbs sampling routine to simulate pairs (X, Y) from this distribution. Examine the behavior of the algorithm as a function of the starting value, and whether the start is taken from X or Y. Does convergence appear to depend on these features, or on α and β? Describe some salient features of the iterative scheme (i.e. need for burn-in, auto-correlation, effective sample sizes), making suitable plots for both variables. You can summarize what you found and give a few selected examples, rather than showing everything you did.

(c) It is known that the marginal distribution of X is $\text{Beta} - \text{Binomial}(n, \alpha, \beta)$. Provide graphical and numerical evidence for or against the Gibbs routine converging to the correct target distribution for X.

Problem 3

Two therapists are asked to rate a sample of n patients on an ordinal severity scale, with possible responses $1, 2, ..., I$. Let Y_1 and Y_2 denote the therapist's ratings for a randomly selected individual and set $Pr(Y_1 = l, Y_2 = k|p) = p_{lk}$ for all possible values of l and k. The marginal distributions of Y_1 and Y_2 are given by

$$Pr(Y_1 = l|p) = p_{l1} + p_{l2} + ... + p_{lI} \equiv p_{l+}$$

and

$$Pr(Y_2 = k|p) = p_{1k} + p_{2k} + ... + p_{Ik} \equiv p_{+k}.$$

Let $p = (p_{11}, p_{12}, ..., p_{II})$ be the $I^2 \times 1$ vector of joint probabilities.

Let $Z = (Z_{11}, Z_{12}, ..., Z_{II})$ be the $I^2 \times 1$ vector with elements Z_{lk} where Z_{lk} is the number of the n patients that receive score l from therapist 1 and score k from therapist 2 (i.e. $Y_1 = l$ and $Y_2 = k$). A simple model for the data assumes that Z given p is Multinomial with sample size n and probability vector p. In practice, the data are summarized as an $I \times I$ table of counts.

As outlined in the attached Section 7.2.6 from Simonoff(1992), a common interest with such data is to provide a measure of agreement between the raters. A standard measure is Cohen's kappa, or κ, defined as

$$\kappa = \frac{\sum_{k=1}^{I} p_{kk} - \sum_{k=1}^{I} p_{k+}p_{+k}}{1 - \sum_{k=1}^{I} p_{k+}p_{+k}}.$$

If rater's responses were independent, then $p_{kk} = p_{k+}p_{+k}$ for each k, implying $\kappa = 0$. If there is perfect agreement between the raters then $\sum_{k=1}^{I} p_{kk} = 1$. So κ scales the excess probability of agreement over that implied by independence (i.e. the numerator) by the maximum possible excess, which is obtained under perfect agreement. Typically $\kappa > 0$. If there is perfect agreement then $\kappa = 1$. Simonoff also discusses weighted kappa which differentially weights discordant ratings as a function of their difference.

a. Read Simonoff's Section 7.2.6 and download and install the **fmsb** package from CRAN. Using a suitably selected function from fmsb, compute a frequentist point estimate and 95% confidence interval for κ for the Sock Test study.

b. Assume a Dirichlet prior on p worth $M = 2$ observations with a best guess for each p_{lk} being 1/16. Use simulation methods to estimate the prior on κ. Create suitable graphical and numerical summaries of this prior and describe the results.

c. Estimate the posterior distribution of κ, giving graphical and numerical summaries, including a 95% posterior credible interval. Discuss these results and contrast them with the frequentist analysis.

7.2.6 *Rater Agreement Tables*

A particular form of matched pairs data occurs when the same object is rated by (two) different evaluators. For example, in a clinical context, a patient might be evaluated for disease status by two different physicians. In this context, independence is not of any great interest — the presumption would be that the ratings of the two physicians are related to each other.

Rather, it is agreement between the two raters that is important, which requires association, but is a stronger condition (for example, the ratings of two physicians who always differ by one rating scale in the same direction are strongly associated, but agree poorly with each other).

A popular summary of the amount of agreement in a table is *Cohen's kappa*. This measure is based on comparing the observed agreement in the table to what would be expected if the ratings were independent. Since agreement corresponds to being on the main diagonal, the excess probability of agreement over that implied by independence is

$$\sum_{i=1}^{I} p_{ii} - \sum_{i=1}^{I} p_{i\cdot}p_{\cdot i}.$$

Kappa scales this value by its maximum possible value (perfect agreement), yielding

$$\kappa = \frac{\sum_i p_{ii} - \sum_i p_{i\cdot}p_{\cdot i}}{1 - \sum_i p_{i\cdot}p_{\cdot i}}. \qquad (7.14)$$

The sample version of this measure, $\hat{\kappa}$, substitutes the observed frequencies for the probabilities in (7.14), and estimates the difference between observed and expected agreement as a proportion of the maximum possible difference. Note that negative values of κ are possible if agreement is worse than would be expected by chance, but this does not typically happen in practice. When $\kappa = 0$, agreement is what would be expected by random chance, and increasing values of κ represent increasing agreement. An asymptotic confidence interval for κ can be constructed using the estimated standard error of $\hat{\kappa}$,

$$\widehat{\text{s.e.}}(\hat{\kappa}) = \left[\frac{A + B - C}{n\left(1 - \sum_i \hat{p}_{i\cdot}\hat{p}_{\cdot i}\right)^2}\right]^{1/2}, \qquad (7.15)$$

where

$$A = \sum_i \hat{p}_{ii}[1 - (\hat{p}_{i\cdot} + \hat{p}_{\cdot i})(1 - \hat{\kappa})]^2,$$

$$B = (1 - \hat{\kappa})^2 \sum_{i \neq j} \hat{p}_{ij}(\hat{p}_{j\cdot} + \hat{p}_{\cdot i})^2,$$

$\hat{P}_{ij}$: observed frequency

and

$$C = \left[\hat{\kappa} - \left(\sum_i \hat{p}_{i\cdot}\hat{p}_{\cdot i}\right)(1 - \hat{\kappa})\right]^2.$$

ɔected by chance is 70% of the

:mperature

Fever	Total
1	50
36	50
37	100

correct measurement, the sen- ɔ0 = .72, while the specificity · thermometer is not identify- ·red. This is consistent with a e pacifier thermometer, which ꓩ accepted that rectal temper- hough the precise relationship (1997) suggested adjusting the ıe observed mean difference in ː following agreement table:

ier temperature

Fever	Total
12	50
46	50
58	100

ıore similar, although there is · ($X^2 = 4.0$, $p = .046$). Cohen's since disagreements from lack ıperature readings (resulting in l with disagreements from lack lting in false positive readings). ıggested that while rectal ther- children (90 days and younger), s crucial, for older children the nstitutes a reasonable noninva- in low-volume ambulatory care ng time for a steady-state tem- ot a major problem.

Cohen's kappa treats all disagreements as being of equal importance, which might not be appropriate, especially for ordinal data (where presumably disagreements in neighboring categories would be considered less serious than ones separated by multiple categories). A weighted version of κ can address this by penalizing some disagreements more than others. Let $\{w_{ij}\}$ be a set of weights such that $0 \le w_{ij} \le 1$, $w_{ij} = w_{ji}$, and $w_{ii} = 1$. The weighted kappa is then defined as

$$\kappa_w = \frac{\sum_i \sum_j w_{ij} p_{ij} - \sum_i \sum_j w_{ij} p_{i\cdot} p_{\cdot j}}{1 - \sum_i \sum_j w_{ij} p_{i\cdot} p_{\cdot j}}.$$

Weights that decrease with increasing degree of disagreement are appropriate for an ordinal table; a common choice of weights is

$$w_{ij} = 1 - (i-j)^2/(I-1)^2. \qquad (7.16)$$

The estimated asymptotic standard error of $\hat{\kappa}_w$ is

$$\widehat{\text{s.e.}}(\hat{\kappa}_w) = \left(\left\{\sum_{ij} \hat{p}_{ij} \left[w_{ij} - (\overline{w}_{i\cdot} + \overline{w}_{\cdot j})(1 - \hat{\kappa}_w)\right]^2 - \left[\hat{\kappa}_w - \left(\sum_{ij} w_{ij} p_{i\cdot} p_{\cdot j}\right)(1 - \hat{\kappa}_w)\right]^2\right\} \Bigg/ \left\{n\left[1 - \left(\sum_{ij} w_{ij} p_{i\cdot} p_{\cdot j}\right)\right]^2\right\}\right)^{\frac{1}{2}}, \qquad (7.17)$$

where $\overline{w}_{i\cdot} = \sum_j \hat{p}_{\cdot j} w_{ij}$ and $\overline{w}_{\cdot j} = \sum_i \hat{p}_{i\cdot} w_{ij}$.

The Sock Test for Evaluating Activity Limitation

Musculoskeletal pain, including back and lower body pain related to arthritis or traumatic injury, is common in both adults and children. The standard treatments for this pain are rest, medication, and physical therapy. Physical therapists and physicians working with patients require methods of measuring impairment, pain, and ability to perform daily activities for diagnosis, classification and prediction of ultimate outcome. While self-evaluation by patients is an important component of rehabilitation, clinician-derived assessments of physical function are also desirable. Thus, there is a need for a simple assessment tool that can be used by physical therapists to determine performance of a daily living task that is probably important to most patients with musculoskeletal pain.

investigated such a tool, the Sock
tting on a sock. As would be true
bility to reflect perceived activity
it to evaluate intertester reliability
sults should agree when applied to

. The patient sits on a high bench,
floor. He or she then lifts up one
s down toward the lifted foot with
fingertips of both hands. Scores on
ale from 0 (can grab the toes with
ase) to 3 (can hardly, if at all, reach
erances on each side of the ankle)).
is assigned the score. The following
on 21 patients with musculoskeletal
ns from two therapists.

e for
pist 2

2	3	Total
0	0	5
0	0	8
3	0	5
0	3	3
3	3	21

ilar marginal distributions for the
r this table is somewhat problem-
le symmetry model fits adequately
of course consistent with marginal
g weights (7.16), is $\hat{\kappa}_w = .89$, with
, the study implies that there is very
k Test.

greement is controversial, for several
ter, it is difficult to interpret. Most
nstitutes strong agreement and $\kappa <$
t (with values in between moderate-
imprecise guideline. The value of

kappa is also strongly dependent on the marginal distributions. For this reason, applications of the same rating system to populations with different characteristics can lead to very different values of κ.

The weighted kappa also has an arbitrariness to it from the choice of weights. For example, if weights $w_{ij} = 1-|i-j|/(I-1)$ are used in the Sock Test example, $\hat{\kappa}_w = .78$, with asymptotic 95% confidence interval (.61, .97), with noticeably different implications for the strength of agreement. Model-based analyses of rater agreement tables do not suffer from these drawbacks, which is a distinct advantage over measures such as kappa.

7.3 Conditional Analyses

The test statistics and confidence intervals discussed in this chapter are asymptotic in nature, and are not necessarily valid for tables with small expected cell counts. Conditional analysis (that is, the construction of exact tests) is thus useful for these models. Just as was true in Section 6.3, the key is to condition on appropriate sufficient statistics, resulting in a null distribution free of any nuisance parameters, allowing exact inference. This is a harder problem here, since the row and column marginal totals are no longer necessarily the appropriate sufficient statistics.

In an important sense, conditional analysis is less important for the tables discussed in this chapter than the unstructured ones described in Chapter 6. The reason for this is that much of the testing here has been based on focused subset tests. For example, for ordered categorical data, the focused test of independence given the fit of the uniform association model is more powerful than the omnibus independence test. Similarly, in square tables, tests of marginal homogeneity are subset tests: symmetry given quasi-symmetry, or symmetry given ordinal quasi-symmetry. These subset tests, being based on (for large tables, far) fewer degrees of freedom, are much more likely to follow their asymptotic null distributions, so p-values are likely to be valid even in tables with small expected cell counts.

This does not mean, of course, that all asymptotic analyses of square or ordered tables are correct. A full discussion of how to adapt exact methods to such tables is beyond the scope of this book, but the flavor of the problem can be seen from one situation that is readily amenable to exact analysis: marginal homogeneity for matched binary data. Recall from Section 7.2.5 that McNemar's test of marginal homogeneity in a 2 × 2 table is actually just a test of equality of the two off-diagonal probabilities, $p_{12} = p_{21}$, and that it does not involve the diagonal cell counts at all. Let $K = n_{12} + n21$ be the total number of observations in the off-diagonal cells. Under the null hypothesis of marginal homogeneity, these K observations should be split evenly between the two cells. That is, under the null hypothesis, $n_{12} \sim \text{Bin}(K, .5)$ (obviously, n_{21} also has this null distribution). The exact

Strand and Wie (1999) proposed and investigated such a tool, the Sock Test, which simulates the activity of putting on a sock. As would be true for any new clinical test, besides an ability to reflect perceived activity limitation in patients, it is also important to evaluate intertester reliability when the test is applied. That is, test results should agree when applied to the same person by different testers.

The Sock Test is performed as follows. The patient sits on a high bench, with his or her feet not touching the floor. He or she then lifts up one leg at a time and simultaneously reaches down toward the lifted foot with both hands, grabbing the toes with the fingertips of both hands. Scores on the Sock Test are given on an ordinal scale from 0 (can grab the toes with fingertips and perform the action with ease) to 3 (can hardly, if at all, reach as far as the malleoli (the rounded protuberances on each side of the ankle)). The leg with more limited performance is assigned the score. The following table summarizes the results of a study on 21 patients with musculoskeletal pain in Bergen, Norway, using evaluations from two therapists.

Score for	*Score for therapist 2*				
therapist 1	0	1	2	3	Total
0	5	0	0	0	5
1	3	5	0	0	8
2	0	2	3	0	5
3	0	0	0	3	3
Total	8	7	3	3	21

A reliable test should result in similar marginal distributions for the two therapists. While formal testing for this table is somewhat problematic, given the small sample, the simple symmetry model fits adequately ($G^2 = 6.9, df = 6, p = .33$), which is of course consistent with marginal homogeneity. The weighted kappa, using weights (7.16), is $\hat{\kappa}_w = .89$, with 95% confidence interval $(.78, .99)$. Thus, the study implies that there is very strong intertester reliability for the Sock Test.

The use of kappa to measure rater agreement is controversial, for several reasons. Since κ is not a model parameter, it is difficult to interpret. Most recommendations agree that $\kappa > .75$ constitutes strong agreement and $\kappa < .4$ corresponds to fair-to-poor agreement (with values in between moderate-to-good), but this is obviously a very imprecise guideline. The value of

kappa is also strongly c
reason, applications of tl
characteristics can lead

The weighted kappa ;
weights. For example, if
Test example, $\hat{\kappa}_w = .78$,
with noticeably different
based analyses of rater ag
which is a distinct advar

7.3 Conditional

The test statistics and
asymptotic in nature, ai
expected cell counts. Con
tests) is thus useful for t
key is to condition on a
distribution free of any n
is a harder problem here
longer necessarily the ap

In an important sense,
discussed in this chapter
ter 6. The reason for this
on focused subset tests.]
cused test of independen
is more powerful than th
tables, tests of marginal
quasi-symmetry, or symn
tests, being based on (fo
much more likely to follo
are likely to be valid ever

This does not mean, of
ordered tables are correct
to such tables is beyond tl
can be seen from one situ
marginal homogeneity for
that McNemar's test of n
just a test of equality of
that it does not involve
$n21$ be the total number
the null hypothesis of ma
be split evenly between t
$n_{12} \sim \text{Bin}(K, .5)$ (obvious