STSCI 5080 Homework 5

Pete Rigas (pbr43 cornell.edu)

November 27, 2020

1 Problem: Verification of covariance equality

To verify that the term $\text{Cov}(Y_j, Y_l) = -np_jp_l$, we recall what was mentioned in lecture, namely that that other term in the covariance, namely the expectation $\mathbf{E}(Y_jY_l)$, for the multinomially sampled variables Y_j, Y_l, is precisely 0, because Y_j, Y_l, from the definition of multinomially sampled random variables, achieve values with probability 1 for different vectors e_j, e_l, where e_j, e_l contain values of 1 in the j, l th column of the vector, respectively. Furthermore, from the **hint**, we know that the summation of the first term, regardless of having n terms in $(Y_1, \cdots, Y_k)$, will have a 0 for each term so the first term will vanish altogether. So with this fact in mind, we are able to conclude that the covariance of the given expression for these multinomial random variables is just $-np_jp_l$, completing **1**.

2 Rice 273

2.1 Problem: MLE

To find the MLE, we will make use of the likelihood function

$$\mathcal{L}(\theta) = \prod_{i=1}^{3} p_i^{y_i}$$

$$= \log(n!) - \sum_{i=1}^{3} \log(X_i!) + X_i \log(1-\theta)^2 + X_2 \log(2\theta(1-\theta)) + X_3 \log(\theta^2) \ ,$$

and rearrangements give,

$$\log n! - \sum_{i=1}^{2} \log X_i! + (2X_1 + X_2)\log(1-\theta) + 2(X_2 + X_3)\log(\theta) + X_2 \log 2 \ ,$$

from which we can take the log of $\mathcal{L}(\theta)$ to obtain, after differentiating and setting the result to 0,

$$\frac{-2X_1 + X_3}{1-\theta} + \frac{2X_3 + X_2}{\theta} = 0 \ ,$$

$$,$$

implying that, given the stipulation that $\sum_{i=1}^{3} X_i = 1$,

$$\hat{\theta} = \frac{X_2 + 2X_3}{2X_1 + 2X_2 + 2X_3} \ .$$

2.2 showing that the MLE is consistent

To show that our MLE above is consistent and that the convergence in probability holds, we must show that the bias and variance of the given multinomial distribution for the genes approaches 0. For this, we recall that the mean of a multinomial random variable X is,

$$\mathbf{E}(X) = np_i \ ,$$

and also that the variance, is $\mathbf{E}(X) = np_i(1-p_i)$. From these simple facts, we then know that the given MLE is unbiased, therefore implying that each of the given quantities vanishes as the number of samples $n \longrightarrow \infty$, because, in order to find the sample distribution of $\hat{\theta}$ for the given gene distributions of different individuals, we have to make use of histograms to determine whether the estimates that we have gathered are reflective of the histograms used to group together values in the sampling distribution. More specifically, albeit not knowing the exact value of θ, we use bootstrapping which can give us an idea of what $\hat{\theta}$ would be, from which we can form approximate confidence intervals to determine if our confidence intervals are appropriate for the values of θ that we are trying to determine. For example, if we use a fixed number of iterations n we can then determine the standard error of the approximation $\hat{\theta}$, from which a sufficiently small standard error can be computed.

Alternatively, one can directly show that the convergence in probability holds by making use of the fact that $\hat{\theta}$ converges in probability to 2θ, because each one of the terms individually converges in probability.

2.3 Hardy-Weinberg

To test whether HW equilibrium really exists, the likelihood ratio statistic that we want is of the form,

$$\begin{aligned}\mathcal{L}(\mathcal{R}|g) &= \frac{\prod_k P(\mathcal{R}_k|p)}{\prod_k P(\mathcal{R})k|g)} \\ &= \frac{\prod_k \sum_{i,j} P(\mathcal{R}_k, G_{i,j}|p)}{\prod_k \sum_{i,j} P(\mathcal{R}_k, G_{i,j}|g)} \\ &= \frac{\prod_k \sum_{i,j} P(\mathcal{R}_k|G_{i,j}P(G_{i,j}|p)}{\prod_k \sum_{i,j} P(\mathcal{R})k|G_{i,j})P(G_{i,j}|g)} \ ,\end{aligned}$$

where we have taken a ratio of probabilities to reflect the total number of readings $\mathcal{R}$ from an individual, of a locus p of gene g that obeys HW. From this likelihood ratio test statistic, we are able to determine whether HW equilibrium really exists by determining, from the statistic above, whether the genotypes that we are measuring, with slleles i, j, for individuals with the ij subscript from $1, \cdots, N$, obey HW equilibrium.

Moreover, with $v = \frac{n(n-1)}{2}$ degrees of freedom corresponding to n alleles, the quantity obeys the distribution

$$-2\log(\mathcal{L}(\mathcal{R}|g) \sim \chi^2_v \ .$$

3 Problem: Inverse Gamma distribution

To find the pdf of the inverse gamma distribution, we recall from $Y \sim \mathrm{Ga}(\alpha, \beta)$ that

$$1 = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} \exp(-\beta y) \mathrm{d}y = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} (\frac{1}{x})^{\alpha-1} \exp(-\frac{\beta}{x})(-\frac{1}{x^2}) \mathrm{d}x = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \frac{1}{x^{\alpha+1}} \exp(-\frac{\beta}{x}) \mathrm{d}x \ .$$

Now, we observe that the integrand of the expression in the integral above is precisely the pdf of the inverse Gamma distribution. Also, to find the mean of the expression, we will make use of the IG pdf, in which the mean of the IG distribution can be calculated with,

$$\begin{aligned}\mathbf{E}(X^n) = \frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty x^n x^{-\alpha-1}\exp(-\frac{\beta}{x})\mathrm{d}x \\ = \frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(\alpha-n)}{\beta^{\alpha-n}} \\ = \frac{\beta^n\Gamma(\alpha-n)}{(\alpha-1)\cdots(\alpha-n)\Gamma(\alpha-n)}\ ,\end{aligned}$$

implying that $\mathbf{E}(X) = \frac{\beta}{\alpha-1}$ setting $n \equiv 1$, also with $\alpha > 1$.

4 Problem: posterior distribution and Bayes Estimator

To find the posterior distribution and Bayes estimator of η, we observe the following. First of all, we know that the joint distribution of this random variable will be of the Normal-Gamma distribution, which has a pdf of the form

$$f(X_1, \cdots, X_n|\lambda_0, \alpha_0, \beta_0) = \frac{(\lambda_0\tau)^{\frac{1}{2}}}{\sqrt{2\pi}}\exp\bigg(-\frac{1}{2}(\lambda_0\tau)(\mu)^2\bigg)\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}\tau^{\alpha_0-1}\exp(-\beta_0\tau)\ ,$$

following from the facts that

$$f(x_1, \cdots, x_n, \eta) = f(x_1, \cdots, x_n|\eta)f(\eta)\ ,$$

and moreover that,

$$f(x_1, \cdots, x_n) = \frac{f(x_1, \cdots, x_n, \eta)}{\int_{\mathbf{R}^+} f(x_1, \cdots, x_n, \eta)\mathrm{d}\eta}\ .$$

From the lecture notes, we know that if a posterior distribution is proportional to another distribution, then the posterior distribution is itself the distribution that it is proportional, or **conjugate**, to. With this in mind, we then conclude that the posterior distribution is also Normal-Gamma, and this posterior distribution has the parameters, with $\mu_0 \equiv 0$,

$$\begin{aligned}\mu_1 = \frac{\lambda_0\mu_0 + n\bar{x}_n}{\lambda_0 + n}\ , \\ \lambda_1 = \lambda_0 + n\ , \\ \alpha_1 = \alpha_0 + \frac{1}{2}\ , \\ \beta_1 = \beta_0 + \frac{1}{2}\bigg(\sum_{i=1}^n (x_i - \bar{x}_n\bigg)^2 + \frac{n\lambda_0(\bar{x}_n - \mu_0)^2}{2(\lambda_0 + n)}\ .\end{aligned}$$

The Bayes estimator for η is precisely the mean of the Gamma distribution, namely the posterior mean, which is of the form,

$$f(X_1, \cdots, X_n|\eta) = \frac{1}{(2\pi n)^{\frac{n}{2}}}\frac{1}{\Gamma(\alpha)\beta^\alpha}\eta^{-\alpha-1}\prod_{i=1}^n \exp\bigg(-\frac{1}{2}\frac{x_i^2}{\eta} - \frac{\beta}{\eta}\bigg)\ ,$$

implying that the mean for the Gamma distribution above with the given parameters is $\alpha\eta$, which is precisely the Bayes Estimator.

5 Problem: Jeffreys prior

5.1 Part A

By definition, the Jeffreys prior is of the form,

$$\sqrt{I(\theta)} = \sqrt{-\mathbf{E}\left(\frac{\partial^2}{\partial\theta^2}\mathrm{log}f(y|\theta)|\theta\right)}$$
$$\Rightarrow f(y_n|\mu) = \mu^{\sum y_i}e^{-n\mu}\prod_{i=1}^{n}\frac{1}{(y_i)!} \ ,$$

which is the square root of the Fischer information. From here, we proceed by taking the logarithms, so that we can differentiate more easily,

$$\mathrm{log}(f(y_n|\mu) = \sum y_i \mathrm{log}\mu - n\mu - \mathrm{log}\sum_{i=1}^{n}(y_i)! \ ,$$

giving the result, after computing the first derivative of the expression above,

$$\begin{aligned}\frac{\partial \mathrm{log}f(y_n|\mu}{\partial\mu} &= \frac{\partial}{\partial\mu}\sum y_i\mathrm{log}\mu - \frac{\partial}{\partial\mu}n\mu \\ &= \sum y_i\frac{1}{\mu} - n \ .\end{aligned}$$

Next, the second derivative gives

$$\sqrt{\mathbf{E}\left(\frac{\partial^2\mathrm{log}f(y_n|\mu)}{\partial\mu^2}\middle|\mu\right)} = \sqrt{\frac{n}{\mu}} \propto \frac{1}{\sqrt{\mu}} \ .$$

So $\frac{1}{\sqrt{\mu}}$ is what we want. However, π_J is not a proper prior because

$$\int_\Theta \pi_J(\lambda)\mathrm{d}\lambda = \int_\Theta \frac{1}{\sqrt{\lambda}}\mathrm{d}\lambda = +\infty \ .$$

5.2 Part B: Bayes Estimator

Given $X_1, \cdots X_n \sim \mathrm{Po}(\lambda)$, the distribution $-c\mathrm{exp}(-n\lambda)\lambda^{\sum x_i}\lambda^{-\frac{1}{2}} = ce^{-n\lambda}\lambda^{\sum_i x_i - \frac{1}{2}} \propto \lambda^{\sum_i x_i + \frac{1}{2}}e^{\frac{-\lambda}{\frac{1}{n}}} \sim \mathrm{Ga}(\sum_i X_i + \frac{1}{2}, \frac{1}{n})$. This gives that the Bayes estimator is of the form $\frac{\sum X_i + \frac{1}{2}}{n}$.

6 Problem: Likelihood ratio statistic

To find LRT, we begin by obtaining the likelihood

$$l(n,\lambda) = n\mathrm{log}(\lambda) - \lambda\sum_i X_i - m\mathrm{log}(\mu) \ ,$$

from which taking the first partials and solving for the FOCs with respect to each variable gives

$$\frac{\partial l}{\partial \mu} = \frac{m}{\mu} - \sum y_i = 0 \ ,$$

and

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum X_i = 0 \ .$$

Solving the system gives

$$\hat{\mu} = \frac{m}{\sum X_i} \ ,$$

and

$$X = \frac{n}{\sum X_i} \ .$$

For $\mathcal{H}_0$,

$$l(\mu_0) = (m+n)\log(\mu_0) - \mu_0(\sum X_i + \sum Y_i) \ ,$$

implying that

$$\Lambda_n = \frac{l(\mu, \lambda)}{l(\mu_0)} = \frac{\exp(-\lambda \sum X_i)\exp(-\mu \sum Y_i)}{\exp(-\mu_0)(\sum X_i + \sum Y_i)} \ .$$

Rearrangements give

$$\begin{aligned}
&\frac{\exp(\sum X_i)^{-\lambda}\exp(\sum Y_i)^{-\mu}}{\left(\exp(\sum X_i + \sum Y_i)\right)^{-\mu_0}} \ , \\
&= \frac{\left(\exp(\sum X_i + \sum Y_i)\right)^{\mu_0}}{\exp(\sum X_i)^{\lambda}\exp(\sum Y_i)^{\mu}} \ , \\
\Rightarrow \Lambda_n &= \frac{(\sum X_i + \sum Y_i)^{n+m}}{(\sum X_i)^n(\sum Y_i)^m} \ , \\
&= \frac{1}{\left(\frac{\sum X_i}{\sum X_i + \sum Y_i}\right)^n \left(\frac{\sum Y_i}{\sum X_i + \sum Y_i}\right)^m} \ , \\
&= z^{-n}(1-z)^{-m} \ , \\
\Rightarrow Z &= \frac{\sum X_i}{\sum X_i + \sum Y_i} \ .
\end{aligned}$$

7 Problem: Testing H_0

7.1 Part A: convergence in distribution

We can show that the given convergence in distribution holds by immediately observing that if $\sqrt{n}(\bar{X}_n - \mu) = S_n T_n \longrightarrow^D N(0, \sigma^2)$, then $\frac{S_n T_n}{S_n} \longrightarrow^D N(0, \frac{\sigma^2}{S_n^2}) = N(0, 1)$.

7.2 Part B: LRT convergence in distribution

We observe that if $T_n \longrightarrow^D N(0,1)$, then $\frac{T_n^2}{n-1} \longrightarrow^D N(0,1)$. But because $\mathbf{E}(\frac{T_n^2}{n-1}) = 0$, and given that $\log(1+\frac{T_n^2}{n-1}) = \frac{T_n^2}{n-1}$ from the **first** hint, we can apply these observations to conclude that $2\log\Delta_n = \frac{n}{n-1}T_n^2 \longrightarrow^D \chi^2(1) \Rightarrow T_n = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{S_n}$.