Statistics 582, Problem Set 5 Solutions

Wellner; 2/11/2015

1. Consider Example 5.5.4 on pages 16 and 17 of the Chapter 5 notes.
(a) Show that the variance of $\hat{\psi}$ is given by

$$Var(\hat{\psi}_n) = \frac{1}{n}\left\{\frac{1}{B}\sum_{j=1}^{B}\frac{\theta_j}{\xi_j} - \psi(\theta)^2\right\}.$$

[Hint: use the formula $Var(Y) = EVar(Y|X) + Var[E(Y|X)]$ twice.]
(b) Use the result of (a) to show that

$$Var(\hat{\psi}_n) \le \frac{1}{n\delta}$$

under the assumption that $\xi_j \ge \delta > 0$ for all $1 \le j \le B$.
(c) What if the sampling probabilities $\xi_i = 1$ for all i: do the conclusions of Example 5.5.4 and (a) and (b) above still hold?

Solution: (a) Since the (X_i, R_i, Y_i)'s are i.i.d.,

$$\begin{aligned}
Var(\hat{\psi}_n) &= n^{-1}Var\left(\frac{R_1Y_1}{\xi_{X_1}}\right) \\
&= n^{-1}\left\{EVar\left(\frac{R_1Y_1}{\xi_{X_1}}|R_1, X_1\right) + Var\left(E\left(\frac{R_1Y_1}{\xi_{X_1}}|R_1, X_1\right)\right)\right\} \\
&= n^{-1}\left\{E\left(\frac{R_1^2}{\xi_{X_1}^2}\theta_{X_1}(1-\theta_{X_1})\right) + Var\left(\frac{R_1}{\xi_{X_1}}\theta_{X_1}\right)\right\} \\
&= n^{-1}\left\{EE\left(\frac{R_1^2}{\xi_{X_1}^2}\theta_{X_1}(1-\theta_{X_1})|X_1\right)\right. \\
&\qquad \left. + EVar\left(\frac{R_1}{\xi_{X_1}}\theta_{X_1}|X_1\right) + Var\left(E\left(\frac{R_1}{\xi_{X_1}}\theta_{X_1}|X_1\right)\right)\right\} \\
&= n^{-1}\left\{E\left(\frac{\theta_{X_1}(1-\theta_{X_1})}{\xi_{X_1}}\right)\right. \\
&\qquad \left. + E\left(\frac{\theta_{X_1}^2}{\xi_{X_1}^2}\xi_{X_1}(1-\xi_{X_1})\right) + Var(\theta_{X_1})\right\} \\
&= n^{-1}\left\{\frac{1}{B}\sum_{j=1}^{B}\frac{\theta_j(1-\theta_j)}{\xi_j} + \frac{1}{B}\sum_{j=1}^{B}\theta_j^2\frac{1-\xi_j}{\xi_j} + \frac{1}{B}\sum_{j=1}^{B}(\theta_j - \overline{\theta})^2\right\} \\
&= n^{-1}\left\{\frac{1}{B}\sum_{j=1}^{B}\frac{\theta_j}{\xi_j} - \psi(\theta)^2\right\}.
\end{aligned}$$

(b) Since $\xi_j \geq \delta$ and $\theta_j \leq 1$ for all j, it follows that

$$Var(\hat{\psi}_n) \leq n^{-1} \frac{1}{B} \sum_{j=1}^{B} \frac{1}{\delta} = \frac{1}{n\delta}.$$

(c) When $\xi_i = 1$ for all $i \in \{1, \ldots, B\}$ the calculations of Example 4.5.4 and (a) and (b) above remain valid. The estimator becomes simply $\hat{\psi}_n = \overline{Y}_n$, and the variance computed in (a) reduces to

$$Var(\hat{\psi}_n) = n^{-1} \left\{ \psi(\theta) - \psi^2(\theta) \right\} = n^{-1} \psi(\theta)(1 - \psi(\theta)) \leq \frac{1/4}{n}.$$

Thus the missing data aspect of the problem (and the introduction of the Horvitz-Thompson estimator) is not crucial to the essence of the problem which is the high-dimensionality of the parameter space.

2. A random variable X takes on values in the set $\{1, 2, 3, 4\}$ with probability distributions $p_0(x)$ or $p_1(x)$ given in the following table.

x	1	3	3	4
$p_0(x)$	.1	.3	.4	.2
$p_1(x)$	.2	.2	.2	.4

(a) Find a most powerful test of size $\alpha = .2$ for testing p_0 versus p_1 and determine its power.
(b) Find a test ϕ which minimizes the sum of risks $a + b$ where $a \equiv E_0\phi$ and $b = E_1(1 - \phi)$.
(c) Compute $d_{TV}(P_0, P_1)$, $H(P_0, P_1)$, and the affinity $\rho(P_0, P_1)$. For the product laws P_0^n and P_1^n (corresponding to observation of $X_1, \ldots, X_n$ i.i.d. P_0 or P_1 respectively), compute $\rho(P_0^n, P_1^n)$ and $H(P_0^n, P_1^n)$ for $n = 8, 32, 128$. What does this imply about the test ϕ_n based on $X_1, \ldots, X_n$ which minimizes the sum of risks?

Solution: (a) Now $p_1(x)/p_0(x) = 2,\ 2/3,\ 1/2,\ 2$ according as $x = 1,\ 2,\ 3,\ 4$, so a most powerful test of size $\alpha = .2$ is given by

$$\phi(x) = \begin{cases} 2/3 & \text{if } x = 1 \text{ or } 4 \\ 0 & \text{if } x = 2 \text{ or } 3. \end{cases}$$

Then

$$E_0\phi(X) = \frac{2}{3}\{P_0(X = 1) + P_0(X = 4)\} = \frac{2}{3}\{.1 + .2\} = .2,$$

while

$$\text{Power} \;=\; E_1\phi(X) = \frac{2}{3}\{P_1X = 1) + P_1(X = 4)\} = \frac{2}{3}\{.2 + .4\} = .4.$$

In fact, the whole family of MP tests is given by

$$\phi_\gamma(x) \;=\; \begin{cases} \gamma & \text{if } \; x = 1, \\ 1 - \gamma/2 & \text{if } \; x = 4, \\ 0 & \text{if } \; x = 2, 3. \end{cases}$$

for some $0 \le \gamma \le 1$. The first test above is just $\phi_{2/3}$. Then

$$E_0\phi_\gamma(X) = \gamma P_0(X = 1) + (1 - \gamma/2)P_0(X = 4) = .1\gamma + .2 - .2\gamma/2 = .2$$

while

$$\begin{aligned} Power \;&=\; E_1\phi_\gamma(X) = \gamma P_1(X = 1) + (1 - \gamma/2)P_1(X = 4) \\ &=\; \gamma(.2) + (1 - \gamma/2)(.4) = .4. \end{aligned}$$

(b) The test which minimizes $a + b$ is given by

$$\phi(x) = \left\{ \begin{array}{ll} 1, & \text{if } \; p_1(x) > p_0(x) \\ 0, & \text{if } \; p_1(x) < p_0(x) \end{array} \right\} = \left\{ \begin{array}{ll} 1, & \text{if } \; x = 1 \text{ or } 4 \\ 0, & \text{if } \; x = 2 \text{ or } 3. \end{array} \right.$$

Then $a = E_0\phi(X) = .3$ and $b = E_1(1 - \phi(X)) = .4$ and for the test minimizing $a + b$ we have $a + b = .7$. Note that $1 - d_{TV}(P_0, P_1) = \int p_0 \wedge p_1 d\mu = .7$.
(c) We compute

$$d_{TV}(P_0, P_1) = \frac{1}{2}\{.1 + .1 + .2 + .2\} = \frac{1}{2}(.6) = .3.$$

Furthermore,

$$\rho(P_0, P_1) = \sqrt{.02} + \sqrt{.06} + \sqrt{.08} + \sqrt{.08} = .9520558...$$

so that

$$H^2(P_0, P_1) = 1 - \rho(P_0, P_1) = .0479442,$$

and $H(P_0, P_1) = .2189616$. Note that the inequalities of proposition 2.1.15 (Chapter 2, notes) are indeed satisfied:

$$\begin{aligned} H^2(P_0, P_1) = .0479 < .3 \;&=\; d_{TV}(P_0, P_1) \\ &<\; H(P_0, P_1)(2 - H^2(P_0, P_1))^{1/2} \\ &=\; .2189616(2 - .0479442)^{1/2} = .3059244 \\ &<\; \sqrt{2}H(P_0, P_1) = .30996585. \end{aligned}$$

For $n = 8,\ 32,\ 128$ we have:

n	$\rho(P_0^n, P_1^n)$	$H(P_0^n, P_1^n)$
1	.952058	.218962
8	.67499	.57009
32	.20759	.89018
128	.0001857	.999071

Since the test $\phi = \phi(X)$ which minimizes the sum of risks has

$$\begin{aligned} E_0\phi(\underline{X}) + E_1(1-\phi(\underline{X})) &= \int p_0(\underline{x}) \wedge p_1(\underline{x}) d\mu(\underline{x}) \\ &\leq \rho(P_0^n, P_1^n) \leq \rho^n(P_0, P_1) \to 0. \end{aligned}$$

From the table above we see that this happens quite rapidly.

3. Suppose that $X_1, \ldots, X_n$ are i.i.d. $N_k(\theta, \Sigma)$ with Σ known. Suppose that $\theta \sim N_k(\mu, \tau^2 I)$.
(a) Find the Bayes estimator for estimating θ with squared error loss $L(\theta, a) = \|\theta - a\|^2 \equiv \sum_{j=1}^k (\theta_j - a_j)^2$.
(b) Use the result of (a) to show that $\overline{X}_n$ is a minimax estimator of θ.

Solution: If $X_1, \ldots, X_n$ are $N_k(\theta, \Sigma)$, and $\theta \sim N_k(\mu, \tau^2 I)$, then the posterior distribution of θ is

$$(\theta | X_1, \ldots, X_n) \sim N_k(V(n\Sigma^{-1}\overline{X}_n + \tau^{-2} I \mu), V)$$

where $V \equiv (n\Sigma^{-1} + \tau^{-2}I)^{-1}$; this follows from straightforward calculation, assuming that Σ^{-1} exists. Hence the Bayes estimator of θ is

$$d_\Lambda(X) = V(n\Sigma^{-1}\overline{X}_n + \tau^{-2}\mu).$$

This follows from Theorem 5.5.1 since for any rule $d = d(\underline{X})$

$$\begin{aligned} \int |\theta - d(\underline{X})|^2 d\Lambda(\theta|\underline{X}) &= \int |\theta - E(\theta|\underline{X})|^2 d\Lambda(\theta|\underline{X}) + |E(\theta|\underline{X}) - d(\underline{X})|^2 \\ &\leq \int |\theta - E(\theta|\underline{X})|^2 d\Lambda(\theta|\underline{X}) \end{aligned}$$

with equality if and only if $d(\underline{X}) = E(\theta|\underline{X})$.
(b) Now

$$\begin{aligned} \mathcal{R}(\Lambda, d_\Lambda) &= E\{(\theta - d_\Lambda(\underline{X}))^T(\theta - d_\Lambda(\underline{X}))\} \\ &= EE\{(\theta - d_\Lambda(\underline{X}))^T(\theta - d_\Lambda(\underline{X}))\Big|\underline{X}\} \\ &= E\{\text{trace}(n\Sigma^{-1} + \tau^{-2}I)^{-1}\} \\ &= \text{trace}(n\Sigma^{-1} + \tau^{-2}I)^{-1} \\ &\to \text{trace}(n^{-1}\Sigma) \quad \text{as} \quad \tau^2 \to \infty \\ &= R(\theta, \overline{X}) \quad \text{for all} \quad \theta. \end{aligned}$$

Hence by Theorem 5.6.7, $\overline{X}_n$ is minimax.

4. **Optional bonus problem 1:** (a) Suppose that $X \sim F$, and let $m = F^{-1}(1/2)$, $\mu = E(X)$, $\sigma^2 = Var(X)$, and we assume that $E(X^2) < \infty$. Show that

$$|m - \mu| \le \sqrt{2\sigma^2}\,.$$

Hint: use Chebychev's inequality.
(b) Show that the inequality in (a) can be improved to $|m - \mu| \le \sigma$ by using convexity of the function $\psi(x) \equiv |m - x|$ and concavity of $\varphi(x) = \sqrt{x}$.

Solution: (a) By Chebychev's inequality,

$$P(|X - \mu| > \sqrt{2}\sigma) \le \frac{\sigma^2}{2\sigma^2} = \frac{1}{2}.$$

or equivalently $P(|X - \mu| \le \sqrt{2}\sigma) > 1/2$. But this implies that $|m - \mu| \le \sqrt{2}\sigma$ since m satisfies $P(X \le m) \ge 1/2$ and $P(X \ge m) \ge 1/2$.
(b) Since ψ is convex it follows that

$$|m - \mu| \;\;=\;\; |m - E(X)| \le E|m - X| \le E(|X - \mu|)$$

since m minimizes $f(b) \equiv E|X - b|$. But $E|X - \mu| = E\{(|X - \mu|^2)^{1/2}\} \le \{E|X - \mu|^2\}^{1/2}$ since $\varphi(x) = \sqrt{x}$ is concave.

5. **Optional bonus problem 2:** Suppose that $X \sim P_\theta$ for $\theta \in \Theta \subset R^k$ has well-defined Fisher information matrix $I(\theta)$ for θ. The *Jeffreys prior* distribution Λ_J has density $\lambda_J(\theta) = \det(I(\theta))^{1/2}$ with respect to Lebesgue measure on Θ. Note that Λ_J may not be a finite measure, and even if Λ_J is a finite measure, it may not have total mass 1. If a prior distribution is a finite measure, then call it a *proper prior distribution*, and correspondingly if it is not a finite measure, call it an *improper prior distribution*. If the resulting posterior distribution is a finite measure, call it a *proper posterior distribution*, and (by convention) normalize it to have total mass 1. See Lehmann and Casella, TPE, pages 230, 234, 287, 305.
(a) Suppose that $X \sim$ Bernoulli(θ). Find the Jeffrey's prior density λ_J for θ. Is Λ_J a finite measure? If it is finite, what is $\Lambda_J((0,1))$? Find the corresponding posterior distribution of Θ starting with the Jeffrey's prior.
(b) Suppose that $X \sim$ Poisson(θ) with $\theta \in (0, \infty)$. Find the Jeffrey's prior density λ_J for θ. Is Λ_J a finite measure? If it is finite, what is $\Lambda_J((0,\infty))$? Find the corresponding posterior distribution of Θ starting with the Jeffrey's prior. Is it ever a proper posterior distribution?
(c) Suppose that $X \sim$ Geometric(θ), i.e. the number of trials until the first success in i.i.d. Bernoulli trials with probability θ of success for each trial – recall

Chapter 1, section 1. Find the Jeffrey's prior density λ_J for θ. Is Λ_J a finite measure? If it is finite, what is $\Lambda_J((0,1))$? Find the corresponding posterior distribution of Θ starting with the Jeffrey's prior. If we observe $X_1, \ldots, X_n$ i.i.d. Geometric(θ), so that $\sum X_i \sim$ Negative Binomial(n, θ) is the posterior distribution "proper" for some n?
(d) Suppose that $X \sim$ Weibull(θ) with $\theta = (\alpha, \beta) \in (0,\infty)\times(0,\infty)$ as in chapters 3 and 4. Find the Jeffrey's prior density λ_J for θ. Is Λ_J a finite measure? If it is finite, what is $\Lambda_J((0,\infty)^2)$? Find the corresponding posterior distribution of Θ starting with the Jeffrey's prior.

Solution: (a) When $X \sim$ Bernoulli(θ), the Information for θ is $I(\theta) = \{\theta(1-\theta)\}^{-1}$, so the Jeffrey's prior for θ is has density

$$\lambda_J(\theta) = \frac{1}{\sqrt{\theta(1-\theta)}}\,.$$

This density is proportional to the Beta($1/2, 1/2$) density

$$\lambda(\theta) = \frac{\Gamma(1)}{\Gamma(1/2)\Gamma(1/2)}\theta^{1/2-1}(1-\theta)^{1/2-1}1_{(0,1)}(\theta)$$

(which is also known as the "arcsin" distribution because the corresponding distribution function is $\Lambda(\theta) = (2/\pi)\arcsin(\sqrt{\theta})$; it arises naturally as the limiting distribution of the proportion of time a random walk stays positive). Thus $\lambda_J((0,1)) = \Gamma(1/2)^2 = \pi$. The corresponding posterior distribution is proportional to

$$\theta^{x-1/2}(1-\theta)^{1-x-1/2} = \theta^{(x+1/2)-1}(1-\theta)^{(3/2-x)-1}\,;$$

i.e. the posterior density is Beta($x+1/2, 3/2-x$) if $X = x$ is observed.
(b) When $X \sim$ Poisson(θ) with $P_\theta(X = x) = \theta^x e^{-\theta}/x!$ for $x = 0, 1, 2, \ldots$, the score function is

$$\dot{l}_\theta(x) = \frac{x}{\theta} - 1,$$

and the information for θ is $I(\theta) = 1/\theta$. Thus the Jeffrey's prior for θ has density λ_J proportional to $\theta^{-1/2}$. Hence

$$\int_0^\infty \lambda_J(\theta)d\theta = \int_0^\infty \lambda^{-1/2}d\theta = \infty,$$

because the integral diverges at ∞. The posterior density is proportional to

$$p(x|\theta)\lambda_J(\theta) = \frac{\theta^x}{x!}\exp(-\theta)\cdot\theta^{-1/2} = \frac{\theta^{x-1/2}}{x!}\exp(-\theta),$$

which has a finite integral for every $x \in \{0, 1, 2, \ldots\}$, namely

$$\int_0^\infty \frac{\theta^{x-1/2}}{x!} \exp(-\theta) d\theta = \frac{\Gamma(x+1/2)}{x!}.$$

Thus the posterior density is Gamma $(x + 1/2, 1)$.

(c) When $X \sim$ Geometric(θ) with $P_\theta(X = x) = (1-\theta)^{x-1}\theta$ for $x = 1, 2, \ldots$, the score function is

$$\dot{l}_\theta(x) = \frac{1}{1-\theta}\left(\frac{1}{\theta} - X\right)$$

and the information for θ is

$$I(\theta) = E_\theta(\dot{l}_\theta(X))^2 = -E_\theta \ddot{l}_{\theta\theta}(X) = \frac{1}{\theta^2(1-\theta)}.$$

Hence the Jeffrey's prior for θ has density

$$\lambda_J(\theta) = \frac{1}{\theta\sqrt{(1-\theta)}}.$$

Here we have

$$\int_0^1 \lambda_J(\theta) d\theta = \int_0^1 \frac{1}{\theta\sqrt{(1-\theta)}} d\theta = \infty$$

because the integral diverges at 0. Hence this density does not correspond to a finite measure. It corresponds in some sense to a Beta$(0, 1/2)$ density; but recall that the Beta(α, β) densities are defined for $\alpha, \beta > 0$. In spite of this the posterior density is proportional to

$$\theta^0(1-\theta)^{x-1-1/2} = (1-\theta)^{x-1/2-1}$$

which corresponds to Beta$(1, x - 1/2)$ if $X = x$ is observed. This is completely proper for any $x \in \{1, 2, \ldots\}$ since $x - 1/2 > 0$.

When $X_1, \ldots, X_n$ are i.i.d. Geometric(θ) so that $\sum_1^n X_i \sim$ Negative Binomial(n, θ), then the density of the data is proportional to $\theta^n(1-\theta)^{y-n}$ with $y \geq n$, and since $I_n(\theta) = n/[\theta^2(1-\theta)]$, the posterior is proportional to

$$\theta^{n-1}(1-\theta)^{y-n-1/2} = \theta^{n-1}(1-\theta)^{y-n+1/2-1}$$

which has finite mass for all $y \geq n$. Hence the posterior is proper for all $n \geq 1$.

(d) From Example 3.2.7 the information matrix for the Weibull density is

$$I(\theta) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{a}{\alpha} \\ \frac{a}{\alpha} & \frac{b^2}{\beta^2} \end{pmatrix}$$

where $a = -(1-\gamma)$ and $b^2 = \pi^2/6 + (1-\gamma)^2$. Therefore the Jeffrey's prior is given by

$$\lambda_J(\theta) = \det(I(\theta))^{1/2} = \frac{\pi/\sqrt{6}}{\alpha}.$$

This is not the density of a finite measure:

$$\int_0^\infty \int_0^\infty \lambda_J(\alpha,\beta) d\alpha d\beta = \frac{\pi}{\sqrt{6}} \int_0^\infty \int_0^\infty \frac{1}{\alpha} d\alpha d\beta = \infty.$$

The posterior density is proportional to

$$\begin{aligned} p_\theta(x)\lambda_J(\theta) &= \frac{\beta}{\alpha}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp(-(x/\alpha)^\beta)\frac{\pi/\sqrt{6}}{\alpha} 1_{(0,\infty)}(\alpha) 1_{(0,\infty)}(\beta) \\ &= \frac{\pi}{\sqrt{6}} \frac{\beta}{\alpha^2}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp(-(x/\alpha)^\beta) 1_{(0,\infty)}(\alpha) 1_{(0,\infty)}(\beta). \end{aligned}$$

When we try to compute the normalizing constant to form a density (namely the marginal density $p(x)$), we find, however, that (using the change of variables $u = \alpha^{-\beta}$ in the second line so that $du = -\beta\alpha^{-\beta-1} d\alpha$)

$$\begin{aligned} p(x) &= \frac{\pi}{\sqrt{6}} \int_0^\infty \int_0^\infty \frac{\beta}{\alpha^2}\left(\frac{x}{\alpha}\right)^{\beta-1} \exp(-(x/\alpha)^\beta) d\alpha d\beta \\ &= \frac{\pi}{\sqrt{6}} \int_0^\infty \frac{x^{\beta-1}}{x^\beta} \int_0^\infty x^\beta \exp(-x^\beta u) du d\beta \\ &= \frac{\pi}{\sqrt{6}} \int_0^\infty \frac{1}{x} d\beta = \infty. \end{aligned}$$

Hence the posterior (for one observation) is not proper for any x.

For n observations we compute

$$p_\theta(\underline{x})\lambda_J(\theta) = \frac{\beta^n}{\alpha^n} \frac{(\prod x_i)^{\beta-1}}{\alpha^{n(\beta-1)}} \exp(-(\sum x_i^\beta/\alpha^\beta)\frac{\pi/\sqrt{6}}{\alpha} 1_{(0,\infty)}(\alpha) 1_{(0,\infty)}(\beta).$$

Therefore, using the same change of variables as before, $u = \alpha^{-\beta}$,

$$\begin{aligned} p(\underline{x}) &= \frac{\pi}{\sqrt{6}} \int_0^\infty \beta^n (\prod x_i)^{\beta-1} \int_0^\infty \frac{1}{\alpha^{n+1}} \frac{1}{\alpha^{n(\beta-1)}} \exp(-\sum x_i^\beta/\alpha^\beta) d\alpha d\beta \\ &= \frac{\pi}{\sqrt{6}} \int_0^\infty \beta^{n-1} (\prod x_i)^{\beta-1} \int_0^\infty u^{n-1} \exp(-\sum x_i^\beta u) du d\beta \\ &= \frac{\pi}{\sqrt{6}} \Gamma(n) \int_0^\infty \beta^{n-1} \frac{(\prod x_i)^{\beta-1}}{(\sum x_i^\beta)^n} d\beta = \frac{\pi}{\sqrt{6}} \frac{\Gamma(n)}{\prod x_i} \int_0^\infty \beta^{n-1} \frac{(\prod x_i)^\beta}{(\sum x_i^\beta)^n} d\beta \end{aligned}$$

This converges if the x_i's are not all equal. This can be seen by writing the last integral as follows:

$$\begin{aligned}\int_0^\infty \beta^{n-1}\frac{(\prod x_i)^\beta}{(\sum x_i^\beta)^n}d\beta &= \int_0^\infty \beta^{n-1}\frac{1}{\left(\frac{\sum_1^n x_i^\beta}{(\prod x_i^\beta)^{1/n}}\right)^n}d\beta \\ &= \int_0^\infty \beta^{n-1}\frac{1}{(\sum_1^n y_i^\beta)^n}d\beta \\ &= \int_0^1 \beta^{n-1}\frac{1}{(\sum_1^n y_i^\beta)^n}d\beta + \int_1^\infty \beta^{n-1}\frac{1}{(\sum_1^n y_i^\beta)^n}d\beta.\end{aligned}$$

where $y_i \equiv x_i/(\prod x_i)^{1/n}$, $i = 1, \ldots, n$. But for $\beta \geq 1$ we have

$$\left(\frac{1}{n}\sum_1^n y_i^\beta\right)^{1/\beta} \geq \frac{1}{n}\sum_1^n y_i,$$

by Liapunov's inequality, so

$$\sum_1^n y_i^\beta \geq n\left(\frac{1}{n}\sum_1^n y_i\right)^\beta$$

where $1 = (\prod_1^n y_i)^{1/n} \leq n^{-1}\sum_1^n y_i$ with strict inequality if the y_i's are not all equal to 1, or, equivalently, if the x_i's are not all equal. Thus the integrals in the last display are bounded above by

$$\begin{aligned}&\int_0^1 \beta^{n-1}\frac{1}{(ny_{(1)}^\beta)^n}d\beta + \int_1^\infty \beta^{n-1}\frac{1}{n^n\overline{y}^{n\beta}}d\beta \\ &= \int_0^1 \beta^{n-1}\frac{1}{(ny_{(1)}^\beta)^n}d\beta + \int_1^\infty \frac{\beta^{n-1}}{n^n}\exp(-n\beta\log(\overline{y}))d\beta \\ &< \infty\end{aligned}$$

where $\log(\overline{y}) > 0$ since $\overline{y} > 1$ if the x_i's are not all equal. Thus the resulting posterior distribution is proper for $n \geq 2$ as long as all the observations are not all equal.