Swarthmore Honors Exam 2017: Statistics

Joseph Blitzstein (Harvard University)

Instructions: This is a 3-hour closed-book, closed-note exam. No calculators are allowed. Show your work and explain your reasoning. The last page contains a table of important distributions.

1. (**Sunny days and rainy days**)

A scientist is planning to make n measurements outdoors to estimate an unknown parameter μ, by making one measurement each day for the next n days (with $n \geq 2$ fixed in advance). Each day will either be sunny or rainy (but not both), with probability p of being sunny.

On the jth day, the scientist's measurement will be

$$y_j \sim \begin{cases} \mathcal{N}(\mu, \sigma_1^2), & \text{if the } j\text{th day is sunny;} \\ \mathcal{N}(\mu, \sigma_0^2), & \text{if the } j\text{th day is rainy.} \end{cases}$$

Let I_j be the indicator r.v. for the jth day being sunny. So the observed data are y, I, where $y = (y_1, \dots, y_n)$ and $I = (I_1, \dots, I_n)$. Assume independence across days, and that the parameters p, σ_0^2, and σ_1^2 are known, with $\sigma_0^2 > \sigma_1^2$. It is observed that at least one day is sunny and at least one day is rainy.

(a) Find the maximum likelihood estimator $\hat{\mu}$ for μ.

(b) In quantifying the uncertainty in estimating μ with $\hat{\mu}$, does it make more sense to report the unconditional standard deviation of $\hat{\mu}$ or the conditional standard deviation given I? Explain briefly.

2. **Bayesian logic.**

Let Y_1, Y_2, θ, α be continuous random variables and f, g, h be probability density functions with the following structure:

$$\begin{aligned} Y_i|\theta, \alpha &\sim f(y|\theta) \text{ independently, for } i = 1, 2 \\ \theta|\alpha &\sim g(\theta|\alpha) \\ \alpha &\sim h(\alpha). \end{aligned}$$

The interpretation is that Y_1 is observed, θ is a parameter of interest, α is a hyperparameter, and Y_2 is a future observation. Provide expressions in terms of f, g, h for the conditional distribution of $\theta|Y_1$ and that of $Y_2|Y_1$.

3. **How rare is your book?**

You are given a limited edition, collector's item book, of which only N copies exist; you have little information about how large N is. To assess how rare the book is, you wish to estimate N. The jth book printed was inscribed with the number j (so the books are numbered $1, 2, \dots, N$). Your book turns out to be Number 180. Assume that your copy is equally likely to be any of the N copies of the book.

(a) Find the maximum likelihood estimator for N (if it exists).

(b) Find the method of moments estimator for N (if it exists).

(c) Using the improper prior $\pi(N) \propto 1/N$ (for $N = 1, 2, 3, \dots$), show that the posterior distribution for N is proper (i.e., the posterior probability mass function can be normalized so that it sums to 1).

(d) Using the prior from (b), find the posterior mode of N (if it exists).

(e) Using the prior from (b), find the posterior mean of N (if it exists).

(f) Using the prior from (b), find a good approximation to the posterior median of N (if it exists).
Hint: Find a good approximation to $P(N \geq n|\text{data})$. It may help to approximate certain sums using certain integrals.

(g) Which of the estimators from the earlier parts would you prefer, and why? (A short, intuitive explanation suffices for this part. Also, if you prefer another estimator to any of the above, you can argue for that estimator instead.)

4. **Probability of no typos.**

Each page of an n-page book has a Pois(λ) number of typos, where λ is unknown (but is not treated as an r.v.). Typos on different pages are independent. Thus we have i.i.d. $X_1, \dots, X_n \sim \text{Pois}(\lambda)$, where X_j is the number of typos on page j. Suppose we are interested in estimating the probability θ that a page has no typos:

$$\theta = P(X_j = 0) = e^{-\lambda}.$$

(a) Let $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Show that $T_n = e^{-\bar{X}_n}$ is biased for estimating θ.

(b) Show that as $n \to \infty$, $T_n \to \theta$ with probability 1.

(c) Show that the following estimator is unbiased for estimating θ:

$$W = \frac{1}{n}\left(I(X_1 = 0) + \cdots + I(X_n = 0)\right)$$

(d) Let $\tilde{W} = E(W|X_1 + \cdots + X_n)$. Find a simplified expression for $\tilde{W}$. Then determine whether $\tilde{W}$ is also unbiased for θ.

Hint: A handy fact, which you do not need to check, is that

$$X_1|(X_1 + \cdots + X_n = s) \sim \mathrm{Bin}(s, 1/n).$$

(e) Determine which of W and $\tilde{W}$ has lower variance.

Table of Important Distributions

Let $0 < p < 1$ and $q = 1 - p$.

Name	Param.	PMF or PDF	Mean	Variance
Bernoulli	p	$P(X = 1) = p, P(X = 0) = q$	p	pq
Binomial	n, p	$\binom{n}{k} p^k q^{n-k}$, for $k \in \{0, 1, \dots, n\}$	np	npq
FS	p	pq^{k-1}, for $k \in \{1, 2, \dots\}$	$1/p$	q/p^2
Geom	p	pq^k, for $k \in \{0, 1, 2, \dots\}$	q/p	q/p^2
NBinom	r, p	$\binom{r+n-1}{r-1} p^r q^n, n \in \{0, 1, 2, \dots\}$	rq/p	rq/p^2
HGeom	w, b, n	$\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$, for $k \in \{0, 1, \dots, n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1}) n \frac{\mu}{n}(1 - \frac{\mu}{n})$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!}$, for $k \in \{0, 1, 2, \dots\}$	λ	λ
Uniform	$a < b$	$\frac{1}{b-a}$, for $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ, σ^2	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2
Expo	λ	$\lambda e^{-\lambda x}$, for $x > 0$	$1/\lambda$	$1/\lambda^2$
Gamma	a, λ	$\Gamma(a)^{-1}(\lambda x)^a e^{-\lambda x} x^{-1}$, for $x > 0$	a/λ	a/λ^2
Beta	a, b	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}$, for $0 < x < 1$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{a+b+1}$
χ^2	n	$\frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$, for $x > 0$	n	$2n$
Student-t	n	$\frac{\Gamma((n+1)/2)}{\sqrt{n\pi}\Gamma(n/2)}(1 + x^2/n)^{-(n+1)/2}$	0 if $n > 1$	$\frac{n}{n-2}$ if $n > 2$