Swarthmore Honors Exam 2013: Statistics

Joseph Blitzstein (Harvard University)

Instructions: This is a 3-hour closed-book, closed-note exam. You may use a calculator that does not do algebra or calculus. Show your work and explain your reasoning. The last page contains a table of important distributions.

1. Two i.i.d. Normal observations are made, $y_1, y_2 \sim \mathcal{N}(\mu, 1)$, with μ unknown. "Student", in his paper introducing the t-distribution 105 years ago, states:

"If two observations have been made and we have no other information, it is an even chance that the mean of the (Normal) population will lie between them."

(a) Verify Student's claim, showing how to interpret it as providing a 50% (frequentist) confidence interval for μ.

(b) Compute the expected length of the interval from (a) exactly (simplify). How does the average length compare ($=, <$, or $>$) to that of the usual 50% confidence interval, $(\bar{y} - z_{0.75}/\sqrt{2}, \bar{y} + z_{0.75}/\sqrt{2})$, where $z_{0.75} \approx 0.67$ is the 0.75 Normal quantile?

2. You have k independent unbiased estimators of an unknown parameter θ, where the jth, denoted $\hat{\theta}_j$, has mean θ and known variance $V_j > 0$. Consider linear combinations of the $\hat{\theta}_j$, with the constraint that the weights assigned to these estimators be such that the resulting combination is unbiased.

(a) Find the best constants, in the sense of minimizing the mean squared error.

(b) Explain intuitively why your answer to (a) makes sense, in terms of Fisher information and/or an example.

3. Let $y = (y_1, \dots, y_k)$ be data and $\mu = (\mu_1, \dots, \mu_k)$ be parameters, connected through the Normal hierarchical model where, for $i = 1, 2, \dots, k$,

$$\begin{aligned} y_i | \mu_i &\overset{\text{ind.}}{\sim} \mathcal{N}(\mu_i, V_i) \\ \mu_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, A). \end{aligned}$$

The variances V_i and the hyperparameters μ_0 and A are known constants.

(a) The parameters μ_i are independent a priori (i.e., before observing the data). Are they also independent a posteriori (i.e., after observing the data)? You can give either a mathematical proof or a convincing intuitive argument.

(b) Find the posterior distribution of μ_i given the data y.

4. A widget-making company wants to study the reliability of their supposedly water-resistant widgets. The *survival time* of a widget that gets wet is defined as the length of time from when the widget gets wet until it stops working. Suppose that such survival times are i.i.d. Exponential r.v.s, with rate parameter λ and mean $\mu = 1/\lambda$, with μ measured in days.

The CEO hires you as a consultant, and hands you a data set $(t_1, \dots, t_{10})$ of survival times (in days) of 10 widgets that got wet. The following conversation ensues.

CEO: "We need some number-crunching help. Can you analyze our data?"

You: "Sure, but before we get to the data analysis, we should clarify some key issues. First of all, what is your scientific goal?"

CEO: "The goal is to figure out the average survival time of a widget that gets wet. Now can you analyze our data?"

You: "It is essential for me to know more about how the data were *sampled.* Can you tell me precisely what the data-collection process was?"

CEO: "A technician poured water on some widgets and then measured their survival times. Now can you analyze our data?"

You: "So there were 10 working widgets to start with, which all got wet?"

CEO: "I don't see why that matters, but there may have been more than 10 initially. The technician poured water on some widgets on a Friday at noon, and then went away for the weekend, returning on the following Monday at noon. While he was gone, some of the widgets may have stopped working and accidentally been disposed of by someone else. The technician forgot to record how many widgets he had initially, but I gave you the survival times for all the widgets that were present when he returned. Now can you analyze our data?"

You: "The statistician R.A. Fisher once said, 'To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.' But I will try."

(a) Find the likelihood function for λ. Note that any widget with a survival time $t < 3$ would have been discarded *without you even knowing of its existence*; the data you have are conditioned on having values of at least 3 (this is called *truncated data*).

(b) Find the MLEs of μ and of λ, and give a simple explanation in words for how and why the MLE of μ differs from the sample mean of $(t_1, \dots, t_{10})$.

A follow-up experiment is performed, this time with you involved from the start. You get 30 widgets wet, and carefully monitor them. But 7 days after you start the experiment, the CEO gets impatient and demands immediate results. At this point in time, 21 widgets have stopped working, and you have recorded their survival times, but for the other 9 widgets, you know only that their survival times will be *at least* 7 days (the survival times for these 9 widgets are said to have been *censored*).

(c) Find the MLEs of μ and of λ (just based on the data from the follow-up experiment), and give a simple explanation in words for how and why the MLE of μ differs from the sample mean of the 21 observed survival times.

Hint: a widget's contribution to the likelihood function for λ is the PDF evaluated at t if the widget was observed to have stopped working at time t, and is the probability of still being working after 7 days if its survival time was censored.

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 $k \in \{0,1,\ldots,n\}$	np	npq
FS	p	pq^{k-1}, for $k \in \{1,2,\ldots\}$	$1/p$	q/p^2
Geom	p	pq^k, for $k \in \{0,1,2,\ldots\}$	q/p	q/p^2
NBinom	r, p	$\binom{r+n-1}{r-1}p^r q^n, n \in \{0,1,2,\ldots\}$	rq/p	rq/p^2
HGeom	w, b, n	$\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$, for $k \in \{0,1,\ldots,n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1})n\frac{\mu}{n}(1-\frac{\mu}{n})$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!}$, for $k \in \{0,1,2,\ldots\}$	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 $n > 2$