Swarthmore Honors Exam 2025: Statistics

Joseph Blitzstein (Harvard University)

Instructions: This is a 3-hour closed-book, closed-note exam. Electronic devices, including calculators, are not allowed. Show your work and explain your reasoning. The last page contains a table of important distributions.

1. Two teams, Team A and Team B, are going to play a 10-game match. For each individual game, the probability is p that Team A will win and $1-p$ that Team B will win, with p unknown. The games are independent. The organizers are wondering whether they made a mistake by having an even number of games, and would like to know how likely it is that the match will end up tied (with 5 wins for each team). Let θ be the probability that the 10-game match ends up tied.

The data are the outcomes $y_1, y_2, \ldots, y_n$ from n games that these two teams have played against each other already, where $y_j = 1$ if Team A won the jth game and $y_j = 0$ if Team B won the jth game. The Y_j's that resulted in the y_j's were i.i.d. with the same parameter p as in the future 10-game match.

(a) Find the maximum likelihood estimator (MLE) $\hat{\theta}$ of θ.

(b) Now suppose instead that a Bayesian approach is used, with prior

$$p \sim \text{Beta}(a, b),$$

where a and b are positive integers (specified *before* observing $y_1, \ldots, y_n$). We now assume that the game outcomes are conditionally independent given p, rather than unconditionally independent. Find the posterior mean of θ.

2. Robin, an ornithologist, is studying how quiet or loud the chirps are that a particular species of bird makes. She measures the loudness of n chirps (in decibels), resulting in i.i.d. observations $Y_1, Y_2, \ldots, Y_n$. Let θ be the loudest possible chirp (in decibels) that it is physically possible for a bird of this species to make.

Robin decides to model Y_j as having the following cumulative distribution function, with the parameters β, θ unknown and positive:

$$P(Y_j \leq y; \beta, \theta) = \begin{cases} 0, \text{ if } y \leq 0; \\ (y/\theta)^\beta, \text{ if } 0 < y < \theta; \\ 1, \text{ if } y \geq \theta. \end{cases}$$

(a) Find the likelihood function $L(\beta, \theta)$.

(b) Find the MLE of θ (you don't have to find the MLE of β). Briefly describe why we should be skeptical of this estimator.

(c) Give a system of two equations that could be solved to obtain method of moments (MoM) estimators for β, θ, based on the first two moments of Y_1. To save time, you don't have to solve the system, but the two equations should be explicit and simplified, so it would just require some algebra to solve the system.

3. We observe i.i.d. random variables

$$Y_1, \ldots, Y_n \sim \text{Expo}(\lambda),$$

where the rate parameter λ needs to be estimated.

(a) Suppose for this part only that we take a Bayesian approach, with prior

$$\lambda \sim \text{Gamma}(\alpha, \beta).$$

Find a 95% credible interval for λ, in terms of the quantile function of a named distribution.

(b) Give a clear, precise explanation for how we could use the bootstrap to obtain an approximate 95% confidence interval for λ.

4. Person 0 has a new virus, and will infect a Poisson number of people. The *first wave* is the people infected by Person 0. Each person in the first wave will then infect a Poisson number of people. The *second wave* is the people infected by someone from the first wave. Each person in the second wave then infects a Poisson number of people, etc. Suppose that no one gets infected more than once, and that the Poisson random variables are i.i.d. with an unknown parameter θ.

We observe $Y_1, Y_2, \ldots, Y_n$, where Y_j is the number of people in the jth wave. Let y_j be the observed value of Y_j.

(a) Find the likelihood function $L(\theta)$.

Hint: First find $P(Y_{j+1} = y_{j+1} | Y_1 = y_1, \ldots, Y_j = y_j)$.

(b) Find a two-dimensional statistic (S, T) that suffices for determining the likelihood function. That is, if we find out the values of S and T then we know the likelihood function from (a), without needing to know the individual values $y_1, \ldots, y_n$.

5. A certain course has two exams: a midterm and a final exam. There are n students in the course. Let x_j be the midterm score of student j, where the scores have been *standardized* so that the sample mean of $\mathbf{x}$ is 0 and the sample variance of $\mathbf{x}$ is 1. Similarly, let y_j be the final exam score of student j, where the scores have been *standardized* so that the sample mean of $\mathbf{y}$ is 0 and the sample variance of $\mathbf{y}$ is 1.

(a) Let $\hat{\alpha}$ be the slope of the fitted regression line in a linear regression model that uses $\mathbf{x}$ to predict $\mathbf{y}$. Let $\hat{\beta}$ be the slope of the fitted regression line in a linear regression model that uses $\mathbf{y}$ to predict $\mathbf{x}$. What is the relationship between $\hat{\alpha}$ and $\hat{\beta}$?

(b) Explain intuitively why $\hat{\beta}$ is *not* the reciprocal of $\hat{\alpha}$, even though intuitively at first it may seem that if we use $y = \alpha x$ to predict y using x, then, solving for x, we should use $x = (1/\alpha)y$ to predict x using y.

(c) For a student who scored one standard deviation above the mean on the midterm, would the linear regression model predict that they would score one standard deviation above the mean on the final exam, higher than that, or lower than that? Give both a mathematical explanation and an intuitive explanation.

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	$\binom{n}{k} p^k q^{n-k},\ \text{for } k \in \{0, 1, \dots, n\}$	np	npq
FS	p	$pq^{k-1},\ \text{for } k \in \{1, 2, \dots\}$	$1/p$	q/p^2
Geom	p	$pq^k,\ \text{for } k \in \{0, 1, 2, \dots\}$	q/p	q/p^2
NBin	r, p	$\binom{r+k-1}{r-1} p^r q^k, k \in \{0, 1, 2, \dots\}$	rq/p	rq/p^2
HGeom	w, b, n	$\frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},\ \text{for } k \in \{0, 1, \dots, n\}$	$\mu = \frac{nw}{w+b}$	$(\frac{w+b-n}{w+b-1})\mu(1 - \frac{\mu}{n})$
Poisson	λ	$\frac{e^{-\lambda}\lambda^k}{k!},\ \text{for } k \in \{0, 1, 2, \dots\}$	λ	λ
Uniform	$a < b$	$\frac{1}{b-a},\ \text{for } x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ, σ^2	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$	μ	σ^2
Log-Normal	μ, σ^2	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-(\log x - \mu)^2/(2\sigma^2)}, x > 0$	$\theta = e^{\mu+\sigma^2/2}$	$\theta^2(e^{\sigma^2} - 1)$
Expo	λ	$\lambda e^{-\lambda x},\ \text{for } x > 0$	$1/\lambda$	$1/\lambda^2$
Gamma	a, λ	$\Gamma(a)^{-1}(\lambda x)^a e^{-\lambda x} x^{-1},\ \text{for } x > 0$	a/λ	a/λ^2
Beta	a, b	$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1 - x)^{b-1},\ \text{for } 0 < x < 1$	$\mu = \frac{a}{a+b}$	$\frac{\mu(1-\mu)}{a+b+1}$
Chi-Square	n	$\frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2},\ \text{for } x > 0$	n	$2n$

The function Γ is given by

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx$$

for all $a > 0$. For any $a > 0$, we have $\Gamma(a + 1) = a\Gamma(a)$.