Problem set for Probability and Statistics 1 — 27 April 2020

Summary

- We examine a sequence of i.i.d. random variables with the same distribution, e.g., $\mathsf{Geom}(\theta)$, $\mathsf{Unif}(0,\theta)$, where θ is a parameter.
- We write $X_1,\dots,X_n \sim F_\theta$, called a **random sample** from F_θ (parametric model).
- We measure $X_1 = x_1, X_2 = x_2, \dots$ and want to estimate θ.
- $\hat{\theta}$ is a function of the measured data $(X_1, X_2, \dots)$ to estimate θ, called an *estimator*.
- **Bias:** $\mathbb{E}_\theta[\hat{\theta} - \theta] \dots \theta$ true parameter, $\hat{\theta}$ our estimate (random variable as it depends on observed data).
- An estimator is **unbiased:** $bias = 0$ for all $\theta \in \Theta$.
- An estimator is **asymptotically unbiased:** bias converges to 0, i.e., $\mathbb{E}_\theta[\hat{\theta}] \to \theta$ for all $\theta \in \Theta$.
- An estimator is **consistent**, denoted $\hat{\theta} \xrightarrow{P} \theta$, if for all $\varepsilon > 0$ and all $\theta \in \Theta$, $\mathbb{P}[|\hat{\theta} - \theta| > \varepsilon] \to 0$.

Estimators and their properties

1. For the exponential distribution $\mathsf{Exp}(\theta)$, $\theta \in \Theta = (0,\infty)$ consider the estimator

$$\hat{\theta} = 1/\bar{X_n} = n/(X_1 + \dots + X_n),$$

and recall that $\mathbb{E}(X) = \frac{1}{\theta}$. Using the fact that $\int_0^\infty t^{-n} e^{-\frac{n\theta}{t}}\,\mathrm{d}t = \frac{\Gamma(n-1)}{(n\theta)^{n-1}}$, determine whether it is unbiased. (See the hints at the end for help.)

2. The variables X_i are independent and identically distributed according to an unknown cdf $F_X(x)$. After generating n samples, for each $x \in \mathbb{R}$ we estimate $F_X(x)$ by

$$\hat{F}_n(x) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{[X_i,\infty)}(x)$$

For a fixed value of x, compute its bias and variance. Is $\hat{F}_n(x)$ consistent?

3. Assume a sample of continuous random variables: $X_1, X_2, \dots, X_n$, where $\mathbb{E}[X_i] = \mu$, $Var(X_i) = \sigma^2 > 0$. Consider the following estimators: $\hat{\mu}_{1,n} = X_n$, $\hat{\mu}_{2,n} = \frac{1}{n+1}\sum_{i=1}^{n} X_i$.

(a) Are $\hat{\mu}_{1,n}$ and $\hat{\mu}_{2,n}$ unbiased?
(b) Are $\hat{\mu}_{1,n}$ and $\hat{\mu}_{2,n}$ consistent?

Practice problems

Recall:

- **Convolution formula:** For continuous independent random variables X, Y, the variable $Z = X + Y$ has the density
$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx.$$
- The *covariance* of random variables X, Y is given by $Cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$; this simplifies as $Cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.
- **Probabilistic Cauchy-Schwarz inequality:** for $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, we have $XY \in L(\Omega, \mathcal{F}, \mathbb{P})$ and
$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$
- The *Pearson correlation coefficient* of $X, Y \in L^2(\Omega, \mathcal{F}, \mathbb{P})$ with $Var(X), Var(Y) > 0$ is given by
$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$
- **Markov's inequality:** $\mathbb{P}[X \geq a\mathbb{E}[X]] \leq \frac{1}{a}$ for $X \geq 0$.
- **Chebyshev's inequality:** $\mathbb{P}[|X - \mathbb{E}[X]| \geq t\sigma_X] \leq \frac{1}{t^2}$.
- **Central limit theorem:** Let $X_1, X_2, \ldots$ a sequence of i.i.d. $L^2(\Omega, F, P)$ random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 > 0$. Then for the sequence $(S_n^*)_{n=1}^{\infty}$, where
$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$
we have
$$S_n^* \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$

distribution	pdf	cdf	expectation	variance
$\mathsf{Unif}(a, b)$	$\frac{1}{b-a}\mathbb{1}_{[a,b]}(x)$	$\min\{\frac{x-a}{b-a}, 1\}\mathbb{1}_{[a,\infty)}(x)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$\mathsf{Exp}(\lambda)$	$\lambda e^{-\lambda x}\mathbb{1}_{[0,\infty)}(x)$	$(1 - e^{-\lambda x})\mathbb{1}_{[0,\infty)}(x)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
$\mathsf{Cauchy}(x_0, \gamma)$	$\frac{1}{\pi\gamma(1+(\frac{x-x_0}{\gamma})^2)}$	$\frac{1}{2} + \frac{1}{\pi}\arctan\frac{x-x_0}{\gamma}$	—	—
$\mathcal{N}(\mu, \sigma^2)$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$	$\Phi(\frac{x-\mu}{\sigma})$	μ	σ^2
$\mathsf{Gamma}(r, \alpha)$	$\frac{\alpha^r x^{r-1} e^{-\alpha x}}{\Gamma(r)}\mathbb{1}_{[0,\infty)}(x)$	$\frac{\gamma(r,\alpha x)}{\Gamma(r)}\mathbb{1}_{[0,\infty)}(x)$	$\frac{r}{\alpha}$	$\frac{r}{\alpha^2}$

z	-4	-3	-2	-1	0	1	2	3	4
$\Phi(z)$	0.00003	0.00135	0.02275	0.15866	0.500000	0.84135	0.97725	0.99865	0.99997

4. Let $X, Y, Z \sim \mathsf{Unif}(0, 1)$ be independent random variables.

(a) What is the distribution of $X + Y$? Determine the density (in two ways) – using the convolution formula and by a geometrical argument.

(b) * What is the distribution of $X + Y + Z$?

5. Let X and Y be two random variables with $\mathbb{E}[X] = 1$, $\sigma_X = 2$, and $\mathbb{E}[Y] = 2$, $\sigma_Y = 1$. Find the maximum possible value for $\mathbb{E}[XY]$. Also, express Y as a function of X for which this maximum is achieved. (See the hints at the end for help.)

6. We know that the average number of points on a test was 40 (out of 100). Estimate the proportion of students with at least 80 points. Improve the estimate if you know that the standard deviation of the number of points is 10.

7. You are throwing a party for 100 guests and wondering how many sandwiches to order. You know from experience that the number of sandwiches eaten by a random guest follows a Poisson distribution with a mean of 3. Approximately how many sandwiches do you need to order so that with probability 0.95 no guest will go hungry?

8. Let $S = \sum_{k=56}^{100} \binom{100}{k}$. Also, let $X = \sum_{i=1}^{100} X_i$, where X_i is 0 or 1, both with probability $\frac{1}{2}$ and the variables $X_1, \ldots, X_n$ are independent. Thus, $X \sim \mathsf{Bin}(100, 1/2)$.

(a) Express S using the cumulative distribution function F_X.

(b) Use CLT to estimate this probability.

(c) We want to estimate whether our coin (and the way we flip it) is fair. If we get more than 55 heads out of a hundred flips, we will say it is not fair. What is the probability of making a mistake? That is, if we have a fair coin, what is the probability that we get more than 55 heads out of a hundred flips? Also compute upper bounds for that probability using Markov's inequality, then Chebyshev's inequality.

9. A statistician wants to estimate the average height h (in meters) of people in a population using n independent samples $X_1, \ldots, X_n$, randomly selected from all possible people. For estimation, he uses the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$. He estimates that the standard deviation of a single measurement is at most 1 meter.

(a) What value of n guarantees that the standard deviation of $\bar{X}_n$ is at most 1 cm?

(b) For what n does Chebyshev's inequality guarantee that $\bar{X}_n$ differs from h by at most 5 cm with at least 99% probability?

(c) The statistician notices that all measured people have heights in the interval $(1.4, 2.1)$. How should he adjust the estimate of the standard deviation? How will the answers to the previous questions change?

For problem 1, recall that for $X_i \sim \mathsf{Exp}(\theta)$ independent, we have $X_1 + \ldots + X_n \sim \mathsf{Gamma}(n, \theta)$. Express the cdf of $\hat{\theta}$ in terms of the cdf for $X_1 + \ldots + X_n$ and use the chain rule and your knowledge of the gamma distribution to take its derivative and obtain its pdf.

For problem 5, the best bound is not obtained by applying the Cauchy-Schwarz inequality directly, but by considering the correlation coefficient.
