Midterm Exam: CS 215

Attempt all six questions. You have a time of 120 minutes for this exam. Clearly mark out rough work. No calculators or phones are allowed (or required :-)). You may use results/theorems we have stated or derived in class, unless explicitly stated otherwise. Avoid writing lengthy answers.

Useful Information

1. The empirical mean of n independent and identically distributed random variables is approximately Gaussian distributed. The approximation accuracy is better when n is larger. If the random variables are Gaussian, the empirical mean is exactly Gaussian distributed.

2. For a non-negative random variable X, we have $P(X \geq a) \leq E(X)/a$ where $a > 0$.

3. For a random variable X with mean μ and variance σ^2, we have $P(|X-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}$.

4. Gaussian PDF: $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$, MGF $\phi_X(t) = e^{\mu t + \sigma^2 t^2/2}$

5. Poisson PMF: $P(X = i) = \dfrac{e^{-\lambda}\lambda^i}{i!}$

6. Taylor series expansion of $f(x)$ about x_0 is given as $f(x) = f(x_0) + (x - x_0)f'(x) + \dfrac{(x-x_0)^2 f^{(2)}(x_0)}{2!} + \dots + \dfrac{(x-x_0)^n f^{(n)}(x_0)}{n!}$

---

1. Consider random variables $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Let Y be a random variable that takes on the value of X_1 with probability p and the value of X_2 with probability $1-p$ where $0 \leq p \leq 1$. Write down an expression for the PDF of Y in terms of the PDFs of X_1 and X_2. If you had access to a program to draw a sample value from $\mathcal{N}(0,1)$, and a program to draw a sample value from Uniform$(0,1)$, then state a procedure to draw a sample from the distribution of Y. [17 points]

2. If $X \sim$ Uniform(a, b) where $0 < a < b$, derive the mean, median, variance, PDF and CDF of $Y = \dfrac{1}{X}$. [17 points]

3. Let $X \sim$ Poisson(λ). In this question, we present a method to prove that the variance of $\sqrt{X}$ is approximately 0.25 when λ is large. (This is popularly called the variance stabilizing transform.) To this end, we define $g(X) = \sqrt{X}$. Write down the second order Taylor series expansion of $g(X)$ about λ. (For the Poisson distribution, the higher order terms can be ignored for large values for λ. This is a fact, which you are not expected to verify here.) Taking expectation on both sides, derive the expression for $E(g(X))$ and hence the expression for the variance of $g(X)$. [4+8+5=17 points]

4. Consider n sample points $\{(x_i, y_i)\}_{i=1}^n$ where for all i, y_i is the value of a random variable $Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$, x_i is known, and α, β are unknown. Assume that the random variables $Y_1, Y_2, ..., Y_n$ are independent. Given these sample points, we have seen in class that the least squares estimate of β and α are respectively given by $\hat{\beta} = \dfrac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$ and $\hat{\alpha} = \dfrac{1}{n}\sum_{i=1}^n Y_i - \hat{\beta}\bar{x}$ where $\bar{x} = \dfrac{1}{n}\sum_{i=1}^n x_i$. Show that these estimates are unbiased. Derive an expression for their variance in terms of $n, \bar{x}, \sigma^2, \{x_i\}_{i=1}^n$. Note that the values $\{x_i\}_{i=1}^n$ are treated as constants. [5+5+4+3=15 points]

5. A storage device contains the annual income of a group G of n families in a country. A computer program has read through these records, and has computed and stored in memory the value $S = \sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - x_j)^2$ where x_i is the annual income of the i^{th} family. Some analysis you wish to perform requires the sample standard deviation of the annual income of the families in G. However, the storage device is incredibly slow and you do not have the option of reading any of the data again. How will you compute the standard deviation given S and n? Derive all required formulae if necessary. [17 points]

6. Given sample values $x'_1, x'_2, ..., x'_n$ respectively from $n > 0$ iid random variables $X_1, X_2, ..., X_n$, each having the CDF $F_X(x)$, the so-called empirical CDF is defined as $F_n(x) = \dfrac{1}{n} \sum_{i=1}^{n} \mathbf{1}(x'_i \leq x)$ where $\mathbf{1}(q)$ is the indicator function which produces 1 if the predicate q is true, and 0 otherwise. Prove that $F_n(x)$ is an unbiased estimate of $F_X(x)$ and derive its variance. Hence prove that $\lim_{n\to\infty} E([F_n(x) - F_X(x)]^2) = 0$. [6+7+4=17 points]