Midterm Exam: CS 215

Attempt all questions. You have 120 minutes for this exam. Clearly mark out rough work. No calculators or phones are allowed (or required :-)). You may directly use results/theorems we have stated or derived in class, unless the question explicitly mentions 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 $a > 0$. This is Markov's inequality.
3. For a random variable X with mean μ and variance σ^2, we have $P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2}$. This is Chebyshev's inequality.
4. Gaussian PDF: $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$, MGF $\phi_X(t) = e^{\mu t+\sigma^2t^2/2}$
5. Poisson PMF: $P(X=i) = \frac{e^{-\lambda}\lambda^i}{i!}$, MGF $\phi_X(t) = e^{\lambda(e^t-1)}$
6. Integration by parts: $\int udv = uv - \int vdu$

---

1. Consider a permutation of the first n positive integers, generated uniformly randomly (i.e. each of the $n!$ different permutations are equally likely). The ordered pair (i,j) in the permutation is called an inversion if $i < j$ but j precedes i (i.e. occurs earlier than i) in the permutation. Determine the expected number of inversions in a uniformly randomly generated permutation of the first n positive integers. [10 points]

2. This problem concerns the design of a spam filter based on knowledge of basic discrete probability. You have a 'training set' of 2000 spam messages and 1000 non-spam messages. A word 'ABC' appears in 400 spam and 60 non-spam messages in the training set. Likewise, the word 'PQR' appears in 200 spam and 25 non-spam messages. Multiple occurrences of a word in the same message are counted as just one. Let E_1 and E_2 denote the events that a message contains the words 'ABC' and 'PQR' respectively. Let S be the event that a message is spam. Assume (i) that E_1 and E_2 are independent, (ii) that $E_1|S$ and $E_2|S$ are also independent, and (iii) that $P(S) = P(S^c)$ where S^c is the set-complement. Estimate the probability that a new message (not in the training set) that contains both the words 'ABC' and 'PQR' is a spam message. *(You can use the training set to estimate certain probability values).* [10 points]

3. Consider independently drawn sample values $x_1, x_2, ..., x_n$, each from Poisson(λ/n) where n is known. What is the maximum likelihood estimate for λ? Derive the bias, variance, MSE of this estimator. Is this a consistent estimator? Why (not)? [10 points]
(There is a physical significance to this question, even though one needn't understand it to answer the question. The noise in an image pixel is typically Poisson in nature. The values $x_1, x_2, ..., x_n$ correspond to n images of the same scene acquired in quick succession with acquisition time T/n per image, instead of acquiring one image in time T.)