PROBLEMS IN BASIC PROBABILITY

MANJUNATH KRISHNAPUR

Note: These are problems I gave as homeworks in the many times I taught the first course in UG probability and statistics at IISc. They are taken from various sources and there are also some that I made up.

Problem 1. (Feller, I.8.1) Among the digits 1,2,3,4,5 first one is chosen, and then a second selection is made among the remaining four digits. Assume that all twenty possible results have the same probability. Find the probability that an odd digit will be selected *(a)* the first time, *(b)* the second time, *(c)* both times.

Problem 2. (*) (Feller, II.10.3) In how many ways can two rooks of different colours be put on a chessboard so that they can take each other?

Problem 3. (Feller, II.10.5) The numbers $1, 2, \ldots, n$ are arranged in random order. Find the probability that the digits *(a)* 1 and 2, *(b)* 1, 2, and 3,   appear as neighbours in the order named.

Problem 4. (Feller, II.10.8) What is the probability that among k digits *(a)* 0 does not appear; *(b)* 1 does not appear; *(c)* neither 0 nor 1 appears; *(d)* at least one of the two digits 0 or 1 does not appear? Let A and B represent the events in *(a)* and *(b)*. Express the other events in terms of A and B.

Problem 5. (Feller, II.10.11) A man is given n keys of which only one fits his door. He tries them successively (sampling without replacement). The number of trials required may be $1, 2, \ldots, n$. Show that each of these outcomes has probability $1/n$.

Problem 6. (*) (Feller, II.10.20) From a population of N elements a sample of size k is taken. Find the probability that none of m prescribed elements will be included in the sample, assuming the sample to be *(a)* without replacement,   *(b)* with replacement. Compare the numerical values for the two methods when *(i)* $N = 100$, $m = k = 3$, and *(ii)* $N = 100$, $m = k = 10$.

Problem 7. (Feller, II.10.28) A group of $2N$ boys and $2N$ girls is divided into two equal groups. Find the probability p that each group has equal number of boys and girls. Estimate p using Stirling's approximation.

Problem 8. (*) (Feller, II.10.39) If r_1 indistinguishable red balls and r_2 indistinguishable blue balls are placed into n cells, find the number of distinguishable arrangements.

Problem 9. (Feller, II.10.40) If r_1 dice and r_2 coins are thrown, how many results can be distinguished?

Problem 10. In how many ways can two bishops be put on a chessboard so that they can take each other?

Problem 11. A deck of n cards labelled $1, 2, \ldots, n$ is shuffled well. Find the probability that the digits *(a)* 1 and 2, *(b)* 1, 2, and 3, appear as neighbours in the order named.

Problem 12. A deck of n cards labelled $1, 2, \ldots, n$ is shuffled well. Find the probability that the digits *(a)* 1 and 2, *(b)* 1, 2, and 3, appear as neighbours in the order named.

Problem 13. (Feller, II.12.1) Prove the following identities for $n \geq 2$. [*Convention:* Let n be a positive integer. Then $\binom{n}{y} = 0$ if y is not an integer or if $y > n$].

$$1 - \binom{n}{1} + \binom{n}{2} - \ldots = 0$$

$$\binom{n}{1} + 2\binom{n}{2} + 3\binom{n}{3} + \ldots = n2^{n-1}$$

$$\binom{n}{1} - 2\binom{n}{2} + 3\binom{n}{3} - \ldots = 0$$

$$2.1\binom{n}{2} + 3.2\binom{n}{3} + 4.3\binom{n}{4} + \ldots = n(n-1)2^{n-2}$$

Problem 14. (Feller, I.12.10) Prove that

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \ldots + \binom{n}{n}^2 = \binom{2n}{n}.$$

Problem 15. (Feller, I.12.20) Using Stirling's formula, prove that $\frac{1}{2^{2n}}\binom{2n}{n} \sim \frac{1}{\sqrt{\pi n}}$. [*Convention:* $a_n \sim b_n$ is shorthand for $\lim\limits_{n \to \infty} \frac{a_n}{b_n} = 1$].

Problem 16. (*) Two positive integers $m \leq N$ are fixed. A box contains N coupons labelled $1, 2, \ldots, N$. A sample of m coupons is drawn.

(1) Write the probability space in the following two ways of drawing the sample.

   (a) (Sampling without replacement). A coupon is drawn uniformly at random, then a second coupon is drawn uniformly at random, and so on, till we have m coupons.

   (b) (Sampling with replacement). A coupon is drawn uniformly at random, its number is noted and the coupon is replaced in the box. Then a coupon is drawn at random from the box, the number is noted, and the coupon is returned to the box. This is done m times.

(2) Let $N = k + \ell$ where k, ℓ are positive integers. We think of $\{1, 2, \ldots, k\}$ as a sub-population of the whole population $\{1, 2, \ldots, N\}$. For each of the above two schemes of sampling (with and without replacement), calculate the probability that the sample of size m contains no elements from the sub-population $\{1, 2, \ldots, k\}$.

Problem 17. (*) Write the probability spaces for the following experiments. Coins and dice may not be fair!

(1) A coin is tossed till we get a head followed immediately by a tail. Find the probability of the event that the total number of tosses is at least N.

(2) A die is thrown till we see the number 6 turn up five times (not necessarily in succession). Find the probability that the number 1 is never seen.

(3) A coin is tossed till the first time when the number of heads (strictly) exceeds the number of tails. What is the probability that the number of tosses is at least 5.

(4) (Extra exercise for fun! Do not submit this part) In the previous experiment, find the probability that the number of tosses is more than N.

Problem 18. (*) A particular segment of the DNA in a woman is $ATTAGCGG$ and the corresponding segment in her husband is $CTAAGGCG$. Write the probability space for the same DNA segment in the future child of this man-woman pair. Assume that all possible combinations are equally likely, and ignore the possibility of mutation.

Problem 19. In a party there are N unrelated people. Their birthdays are noted (ignore leap years and assume that a year has 365 days). Find the probability of the event that no two of them have the same birthday. Get the numerical value for $N = 20$ and $N = 30$.

Problem 20. A deck of 52 cards is shuffled well and 3 cards are dealt. Find the probability of the event that all three cards are from distinct suits.

Problem 21. Place r_1 indistinguishable blue balls and r_2 indistinguishable red balls into m bins uniformly at random. Find the probability of the event that the first bin contains balls of both colors.

Problem 22. A coin with probability p of turning up H (assume $0 < p < 1$) is tossed till we get a TH or a HT (i.e., two consecutive tosses must be different, eg., TTH or $HHHT$). Find the probability of the event that at least 5 tosses are required.

Problem 23. A drunken man return home and tries to open the lock of his house from a bunch of n keys by trying them at random till the door opens. Consider two cases: (1) He is so drunk that he may try the same key several times. (2) He is moderately drunk and remembers which keys he has already tried.In both cases, find the probability of the event that he needs n or more attempts to open the door.

Problem 24. Let $\mathbf{x} = (0, 1, 1, 1, 0, 1)$ and $\mathbf{y} = (1, 1, 0, 1, 0, 1)$. A new 6-tuple $\mathbf{z}$ is created at random by choosing each z_i to be x_i or y_i with equal chance, for $1 \le i \le 6$ (A toy model for how two DNA sequences can recombine to give a new one). Find the probability of the event that $\mathbf{z}$ is identical to $\mathbf{x}$.

Problem 25. From a group of W women and M men, a team of L people is chosen at random (of course $L \le W + M$). Find the probability of the event that the teams consists of exactly k women.

Problem 26. Place r distinguishable balls in m labelled bins in such a way that each bin contains at most one ball. All *distinguishable* arrangements are deemed equally likely (this is known as Fermi-Dirac statistics). Find the probability that the first bin is empty.

Problem 27. A box contains $2N$ coupons labelled $1, 2, \ldots, 2N$. Draw k coupons (assume $k \le N$) from the box one after another (1) with replacement, (2) without replacement. Find the probability of the event that no even numbered coupon is in the sample.

Problem 28. In a class with 108 people, one student gets a joke by e-mail. He/she forwards it to one randomly chosen classmate. The recipient does the same - chooses a classmate at random (could be the sender too) and forwards it to him/her. The process goes on like this for 20 steps and stops. What is the probability that the first person to get the mail does not get it again? What is the chance that no one gets the e-mail more than once?

Problem 29. (Feller, III.6.3) Find the probability that in five tossings a coin falls head at least three times in succession.

Problem 30. (*) (Feller, III.6.1) Ten pairs of shoes are in a closet. Four shoes are selected at random. Find the probability that there will be at least one pair among the fours shoes selected.

Problem 31. (*) Let $A_1, A_2, A_3, \ldots$ be events in a probability space. Write the following events in terms of $A_1, A_2, \ldots$ using the usual set operations (union, intersection, complement).

(1) An infinite number of the events A_i occur.

(2) All except finitely many of the events A_i occur.

(3) Exactly k of the events A_i occur.

Problem 32. (Feller, I.8.1) Let $A_1, \ldots, A_n$ be events in a probability space (Ω, p) and let $0 \leq m \leq n$. Let B_m be the event that at least m of the events $A_1, \ldots, A_n$ occur. Mathematically,

$$B_m = \bigcup_{1 \leq i_1 < i_2 < \ldots < i_m \leq n} (A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_m}).$$

Show that

$$\mathbf{P}(B_m) = S_m - \binom{m}{1} S_{m+1} + \binom{m+1}{2} S_{m+2} - \binom{m+2}{3} S_{m+3} + \ldots$$

where $S_k = \sum_{1 \leq i_1 < i_2 < \ldots < i_k \leq n} \mathbf{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k})$.

Problem 33. (*) Recall the problem of matching two shuffled decks of cards, but with n cards in each deck, so that $\Omega_n = S_n \times S_n$ and $p_{(\pi, \sigma)} = \frac{1}{(n!)^2}$ for each $(\pi, \sigma) \in \Omega$. Let A_m be the event that there are exactly m matches between the two decks[1].

(1) For fixed $m \geq 0$, show that $\mathbf{P}(A_m) \to e^{-1} \frac{1}{m!}$ as $n \to \infty$.

(2) Assume that the approximations above are valid for $n = 52$ and $m \leq 10$. Find the probability that there are at least 10 matches.

[**Remark:** You may use the result of the previous problem to solve this one].

---

[1]Strictly speaking, we should write $A_{n,m}$, since the $A_{n,m} \subseteq \Omega_n$ but for ease of notation we omit the subscript n. Similarly, it would be appropriate to write p_n and $\mathbf{P}_n$ for the probabilities, but again, we simplify the notation when there is no risk of confusion.

Problem 34. Place r_n distinguishable balls in n distinguishable urns. Let A_n be the event that at least one urn is empty[2].

(1) If $r_n = n^2$, show that $\mathbf{P}(A_n) \to 0$ as $n \to \infty$.

(2) If $r_n = Cn$ for some fixed constant C, show that $\mathbf{P}(A_n) \to 1$ as $n \to \infty$.

(3) Can you find an increasing function $f(\cdot)$ such that if $r_n = f(n)$, then $\mathbf{P}(A_n)$ does not converge to 0 or 1? [**Hint:** First try $r_n = n^\alpha$ for some α, not necessarily an integer].

Problem 35. A box contains N coupons labelled $1, 2, \ldots, N$. Draw m_N coupons at random, with replacement, from the box. Let A_N be the event that every coupon from the box has appeared at least once in the sample.

(1) If $m_N = N^2$, show that $\mathbf{P}(A_N) \to 1$ as $N \to \infty$.

(2) If $m_N = CN$ for some fixed constant C, show that $\mathbf{P}(A_N) \to 0$ as $N \to \infty$

(3) Can you find an increasing function $f(\cdot)$ such that if $m_N = f(N)$, then $\mathbf{P}(A_N)$ does not converge to 0 or 1? [**Hint:** See if you can relate this problem to the previous one].

Problem 36. Let $A_1, \ldots, A_n$ be events in a common probability space. Let B be the event that *at least two* of the A_is occur. Prove that

$$\mathbf{P}(B) = S_2 - 2S_3 + 3S_4 - \ldots + (-1)^m(m-1)S_m$$

where $S_k = \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbf{P}\{A_{i_1} \cap \ldots \cap A_{i_k}\}$ for $1 \le k \le n$.

Not for submission: More generally, you may show that the probability that at least ℓ of the A_is occur is equal to

(1) $$S_\ell - \binom{\ell}{1}S_{\ell+1} + \binom{\ell+1}{2}S_{\ell+2} - \binom{\ell+2}{3}S_{\ell+3} + \ldots$$

Problem 37. Continuing with the notation of the previous problem, assume the formula given in (1) holds. If C is the event that *exactly ℓ* of the events A_is occur, then show that

(2) $$\mathbf{P}(C) = S_\ell - \binom{\ell+1}{\ell}S_{\ell+1} + \binom{\ell+2}{\ell}S_{\ell+2} - \ldots + (-1)^{n-\ell}S_n.$$

[**Hint:** If you want to prove this directly, without using (1), that is also okay.]

Problem 38. If r distinguishable balls are placed at random into m labelled bins, write an expression for the probability that each bin contains at least two balls.

---

[2]Similar to the previous comment, here it would be appropriate to write $\mathbf{P}_n(A_n)$ as the probability spaces are changing, but we keep the notation simple and simply write $\mathbf{P}(A_n)$.

Problem 39. A deck of cards is dealt to four players (13 cards each). Find the probability that at least one of the players has two or more *aces*.

Problem 40. Let p be the probability that in a gathering of 2500 people, there is some day of the year that is not the birthday of anyone in the gathering. Make reasonable assumptions and argue that $0.3 \leq p \leq 0.4$.

Problem 41. Consider the problem of a psychic guessing the order of a deck of shuffled cards. Assume complete randomness of the guesses. Use the formula in (2) to derive an expression for the probability that the number of guesses is exactly ℓ, for $0 \leq \ell \leq 52$. Use meaningful approximation to these probabilities and give numerical values (to 3 decimal places) of the probabilities for $\ell = 0, 1, 2 \ldots, 6$.

Problem 42. Place r_m distinguishable balls in m distinguishable bins. Let A_m be the event that at least one bin is empty[3].

(1) If $r_n = m^2$, show that $\mathbf{P}(A_m) \to 0$ as $m \to \infty$.

(2) If $r_m = Cm$ for some fixed constant C, show that $\mathbf{P}(A_m) \to 1$ as $n \to \infty$.

(3) Can you find an increasing function $f(\cdot)$ such that if $r_m = f(m)$, then $\mathbf{P}(A_m)$ does not converge to 0 or 1? [**Hint:** First try $r_m = m^\alpha$ for some α, not necessarily an integer].

Problem 43. (*) A random experiment is described and a random variable observed. In each case, write the probability space, the random variable and the pmf of the random variable.

(1) Two fair dice are thrown. The sum of the two top faces is noted.

(2) Deal thirteen cards from a shuffled deck and count (a) the number of red cards (i.e., diamonds or hearts), (b) the number of kings, (c) the number of diamonds.

Problem 44. (*) Place r distinguishable balls in m distinguishable bins at random. Count the number of balls in the first bin.

(1) Write the probability space and the random variable described above.

(2) Find the probability mass function of the number of balls in the first bin.

(3) Find the expected value of the number of balls in the first bin.

---

[3]Here it would be appropriate to write $\mathbf{P}_m(A_n)$ as the probability spaces are changing, but we keep the notation simple and simply write $\mathbf{P}(A_n)$.

Problem 45. (*) Find $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$ for the following random variables.

(1) $X \sim \text{Geo}(p)$.

(2) $X \sim \text{Hypergeo}(N_1, N_2, m)$.

Problem 46. Let X be a non-negative integer-valued random variable with CDF $F(\cdot)$. Show that $\mathbf{E}[X] = \sum_{k=0}^{\infty}(1 - F(k))$.

Problem 47. A coin has probability p of falling head. Fix an integer $m \geq 1$. Toss the coin till the mth head occurs. Let X be the number of tosses required.

(1) Show that X has pmf

$$f(k) = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k = m, m+1, m+2, \ldots.$$

(2) Find $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$.

[**Note:** When $m = 1$, you should get the Geometric distribution with parameter p. We say that X has *negative-binomial distribution*. Some books define $Y := X - m$ (the number of tails till you get m heads) to be a negative binomial random variable. Then, Y takes values $0, 1, 2, \ldots$.]

Problem 48. For a pmf $f(\cdot)$, the *mode* is defined as any point at which f attains its maximal value (i.e., t is a mode if $f(t) \geq f(s)$ for any s). For each of the following distributions, find the mode(s) of the distribution and the value of the pmf at the modes.

(1) $\text{Bin}(n, p)$.

(2) $\text{Pois}(\lambda)$.

(3) $\text{Geo}(p)$.

Problem 49. Use MATLAB for the following exercise.

(1) Plot the pmf of Binomial, Poisson and Geometric distributions for various values of the parameters. Observe the plots to say where the maximum is attained, how the shape changes with changes in parameter, etc. [For definiteness

(2) Simulate random numbers (number of samples can be 50 or 100 etc) from the same distributions and plot their histograms. Visually compare the histograms with the plots of the pmf.

(3) Consider the "real-life" data given in Feller's book (chapter 6) and plot their histograms. Compare with the plot of the pmf for the appropriate distribution with appropriate choice of parameters.

Problem 50. Let A, B be events with positive probability in a common probability space. We have seen in class that $\mathbf{P}(A|B)$ and $\mathbf{P}(B|A)$ are not to be confused.

(1) Show that $\mathbf{P}(A|B) = \mathbf{P}(B|A)$ if and only if $\mathbf{P}(A) = \mathbf{P}(B)$.

(2) Show that $\mathbf{P}(A|B) > \mathbf{P}(A)$ if and only if $\mathbf{P}(B|A) > \mathbf{P}(B)$. That is, if occurrence of B makes A more likely than it was before, then the occurrence of A makes B more likely than it was.

Problem 51. There are 10 bins and the kth bin contains k black and $11 - k$ white balls. A bin is chosen uniformly at random. Then a ball is chosen uniformly at random from the chosen bin.

(1) Find the conditional probability that the chosen ball is black, given that the kth bin was chosen. Use this to compute the (unconditional) probability that the chosen ball is white.

(2) Given that the chosen ball is black, what is the probability that the kth bin was chosen?

Problem 52. A fair die is thrown n times. For $1 \leq k \leq n - 1$, let A_k be the event that the kth throw and the $(k + 1)$st throw yield the same result. Are $A_1, \ldots, A_{n-1}$ independent? Are they pairwise independent?

Problem 53. Suppose r distinguishable balls are placed in m labelled bins at random. Each ball has probability p_k of going into the kth bin, where $p_1 + \ldots + p_m = 1$. Let X_k be the number of balls that go into the kth bin.

(1) Find the pmf of X_1.

(2) Find the pmf of the random variable $X_1 + X_2$.

Problem 54. Two fair dice are thrown and let X be the total of the two numbers that show up. Find the pmf of X. What is the most likely value of X?

Problem 55. Two dice (not necessarily identical, and not necessarily fair) are thrown and let X be the total of the two numbers that turn up. Can you design the two dice so that X is equally likely to be any of the numbers $2, 3, \ldots, 12$?

Problem 56. A coin has probability p of falling head. Assume $0 < p < 1$ and fix an integer $m \geq 1$. Toss the coin till the mth head occurs. Let X be the number of tosses required. Show that X has pmf
$$f(k) = \binom{k-1}{m-1} p^m (1-p)^{k-m}, \quad k = m, m+1, m+2, \ldots.$$
Find the CDF of X.

[**Note:** When $m = 1$, this is the Geometric distribution with parameter p. We say that X has *negative-binomial distribution*. Some books define $Y := X - m$ (the number of tails till you get m heads) to be a negative binomial random variable. Then, Y takes values $0, 1, 2, \ldots$.]

Problem 57. A box contains n coupons with one number on each coupon. We do not know the numbers but we know that they are distinct. Coupons are drawn one after another from the box, without replacement (i.e., after choosing a coupon at random, it is not put back into the box before drawing the next coupon). If the kth number drawn is larger than all the previous numbers, what is the chance that it is the largest of the n numbers?

Problem 58. Let X be a random variable with distribution (CDF) F and density f.

(1) Find the distribution and density of the random variable $2X$.

(2) Find the distribution and density of the random variable $X + 5$.

(3) Find the distribution and density of the random variable $-X$.

(4) Find the distribution and density of the random variable $1/X$.

Problem 59. Let X be a random variable with Gamma(ν, λ) distribution. Let F be the CDF of X. When ν is a positive integer, show that for $t \geq 0$,

$$F(t) = 1 - e^{-\lambda t} \sum_{k=0}^{\nu-1} \frac{\lambda^k t^k}{k!}.$$

[**Note:** Observe that this quantity is the same as $\mathbf{P}(N \geq \nu)$ where N is a Poisson random variable with parameter λt. There is a connection here but we cannot discuss it now].

Problem 60. For a pdf $f(\cdot)$, the *mode* is defined as any point at which f attains its maximal value (i.e., t is a mode if $f(t) \geq f(s)$ for any s). For each of the following distributions, find the mode(s) of the distribution and the value of the pmf at the modes.

(1) $N(\mu, \sigma^2)$.

(2) Exp(λ).

(3) Gamma$(\nu, 1)$.

Problem 61. (*) Let F be a CDF. For each $0 < q < 1$, the *q-quantile(s)* of F is any number $t \in \mathbb{R}$ such that $F(s) \leq q$ if $s < t$ and $F(s) \geq q$ if $s > t$.

(1) If F is the CDF of Exp(λ) distribution, find its q-quantile(s).

(2) If F is the $N(0, 1)$ distribution, use the normal tables to find the unique q-quantile for the following values of q: $0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99$.

(3) If F is the Geo(0.02) distribution, find a q-quantile for $q = 0.01, 0.25, 0.5, 0.75, 0.99$.

Problem 62. (*) Give explicit description of how you would simulate random variables from the following distributions.

(1) The standard Cauchy distribution with density $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$.

(2) The Beta(1/2, 1/2) distribution with density $\frac{1}{\pi\sqrt{x(1-x)}}$.

(3) (<u>Do not need to submit this</u>) Draw 100 random numbers from either of these densities (on MATLAB or any other program that gives uniform random numbers) using the above procedure and draw the histograms. Compare the histograms to the plot of the densities.

Problem 63. In each of the following situations, the distribution of the random variable X is given. Find the distributionof Y (it is enough to find the density of Y).

(1) $X \sim \text{Unif}[0, 1]$ and $Y = \sin^{-1}(X)$.

(2) $X \sim \text{Unif}[0, 1]$ and $Y = \cos^{-1}(X)$.

(3) $X \sim N(0, 1)$ and $Y = X^2$.

[**Note:** We define $\sin^{-1}$ to take values in $[-\pi/2, \pi/2]$ and $\cos^{-1}$ to take values in $[0, \pi]$. In the third part, observe that $f(x) = x^2$ is not a one-one function, so the formula given in the notes does not apply directly].

For the next problem: If X is a random variable with density $f(x)$, then its expected value is defined as $\mathbf{E}[X] = \int_{-\infty}^{+\infty} xf(x)dx$ (this integral is defined only if $\int_{-\infty}^{+\infty} |x|f(x)dx$ is finite). We shall study this notion in greater detail in class, but for now take it as an exercise in integration. More generally, if $T : \mathbb{R} \to \mathbb{R}$, the $\mathbf{E}[T(X)] = \int_{-\infty}^{+\infty} T(x)f(x)dx$.

Problem 64. Find $\mathbf{E}[X]$ and $\mathbf{E}[X^2]$ for the following cases.

(1) $X \sim N(\mu, \sigma^2)$.

(2) $X \sim \text{Gamma}(\nu, \lambda)$. Note the answers for the particular case of $\text{Exp}(\lambda)$.

(3) $X \sim \text{Beta}(p, q)$. Note the answers for the particular case of $\text{Unif}[0, 1]$.

Problem 65. What is the mode of the (a) Pois(λ) distribution? (b) Hpergeometric(M, W, K) distribution? (Mode means the point(s) where the pmf (or pdf) attains its maximal value).

Problem 66. (1) Let $f(k) = \frac{1}{k(k+1)}$ for integer $k \geq 1$. Show that f is a pmf and find the corresponding CDF.

(2) Let $\alpha > 0$ and set $F(x) = 1 - \frac{1}{x^\alpha}$ for $x \geq 1$ and $F(x) = 0$ for $x < 1$. Show that F is a CDF and find the corresponding density function. (This is known as the *Pareto* distribution).

Problem 67. Give explicit description of how you would simulate random variables from the following distributions.

(1) The standard Cauchy distribution with density $f(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$.

(2) The Beta$(1/2, 1/2)$ distribution with density $\frac{1}{\pi\sqrt{x(1-x)}}$.

(3) (Do not need to submit this) Draw 100 random numbers from either of these densities (on MATLAB or any other program that gives uniform random numbers) using the above procedure and draw the histograms. Compare the histograms to the plot of the densities.

Problem 68. Let X be a random variable with distribution function F. Let $a > 0$ and $b \in \mathbb{R}$ and define $Y = aX + b$.

(1) What is the CDF of Y?

(2) If X has a density f, find the density of Y.

Problem 69.   (1) Let $X \sim \text{Exp}(\lambda)$. Fix $s, t > 0$ and compute the conditional probability of the event $X > t + s$ given that $X > s$.

(2) Let ν be a positive integer. Show that the CDF of Gamma(ν, λ) distribution is given by

$$F(x) = 1 - e^{-\lambda x} \sum_{k=0}^{\nu-1} \frac{\lambda^k}{k!} x^k.$$

Problem 70. Let $U \sim \text{Uniform}[0, 1]$. Find the density and distribution functions of (a) U^p (where $p > 0$), (b) $U/(1 - U)$, (c) $\log(1/U)$, (d) $\frac{2}{\pi}\arcsin(U)$.

Problem 71. Let $X \sim N(0, 1)$. Find the density of (a) $aX + b$ (where $a, b \in \mathbb{R}$), (b) X^2 , (c) X^3, (d) e^X.

Problem 72. In a game, there are three closed boxes, in exactly one of which there is a prize. The player is asked to pick one of the three boxes. The organizer (who knows where the prize is), opens one of the other two boxes and shows that it is empty. Now the player has two choices, can stick to her first choice or to switch to the other closed closed. What should she do? This is known as the *onty hall paradox*. The word "paradox" is used to convey the strong feeling that many have that the probabilities of the two boxes are $1/2$ and $1/2$, since there is always one empty box out of the other two and it gives no information.

To make the problem well-defined, one has to specify how the organizer chooses the empty box. If the player's first choice is empty, then exactly one of the other two boxes is empty and the organizer has no choice but to show that. If the player's first choice is correct, *assume* that the organizer (secretly) tosses a fair coin to choose which of the other two boxes to show. With this specification, show that the probability that the prize is in the original choice is $1/3$ (in other words, if you switch, the chances are higher, that is $2/3$).

Problem 73. If F is a CDF, show that it can have at most countably many discontinuity points.

Problem 74. (*) Let A, B be two events in a common probability space. Write the joint distributions (joint pmf) of the following random variables.

(1) $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$.

(2) $X = \mathbf{1}_{A \cap B}$ and $Y = \mathbf{1}_{A \cup B}$.

Problem 75. Let $a > 0, b > 0$ and $ab > c^2$. Let (X, Y) have the bivariate normal distribution with density
$$f(x, y) = \frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}\left[a(x-\mu)^2 + b(y-\nu)^2 + 2c(x-\mu)(y-\nu)\right]}.$$
Show that the marginal distributions are one-dimensional normal and find the parameters. For what values of the parameters are X and Y independent?

Problem 76. (*) Fix $r > 0$. Let (X, Y) be a random vector with density
$$f(x, y) = \begin{cases} \frac{1}{\pi r^2} & \text{if } x^2 + y^2 \leq r^2, \\ 0 & \text{otherwise.} \end{cases}$$
This models the experiment of drawing a point at random from a disk of radius r centered at $(0, 0)$.

(1) Find the marginal densities of X and Y (i.e., find the density of X and find the density of Y separately).

(2) Can you solve the same problem if the point is drawn uniformly from the ellipse $\{(x, y) : \frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1\}$?

Problem 77. (*) Let $\mathbf{X} = (X_1, \ldots, X_{n-1})$ be a Multinomial random variable with parameters $r, n, p_1, \ldots, p_n$ where r, n are positive integers and p_i are non-negative numbers that sum to 1. This means that $\mathbf{X}$ has pmf
$$f(k_1, \ldots, k_{n-1}) = \frac{n!}{k_1! k_2! \ldots k_{n-1}!(r - k_1 - \ldots - k_{n-1})!} p_1^{k_1} \ldots p_{n-1}^{k_{n-1}} p_n^{r - k_1 - \ldots - k_{n-1}}$$
if $k_i \geq 0$ are integers that add to at most r.

(1) Let $m \leq n$. Show that the distribution of $(X_1, \ldots, X_{m-1})$ is Multinomial with parameters $r, m, \tilde{p}_1, \ldots, \tilde{p}_m$ where $\tilde{p}_i = p_i$ for $i \leq m-1$ and $\tilde{p}_m = p_m + \ldots + p_n$.

(2) The distribution of X_k is $\text{Bin}(r, p_k)$.

(3) (<u>Do not need to submit this</u>) Let $k_1 < k_2 < \ldots < k_m = n$. Define $Y_1 = X_1 + \ldots + X_{k_1 - 1}$, $Y_2 = X_{k_1} + \ldots + X_{k_2 - 1}, \ldots Y_m = X_{k_{m-1}} + \ldots + X_{k_m - 1}$. What is the distribution of $(Y_1, \ldots, Y_m)$?

[**Note** Remember the balls-in-bins interpretation of Multinomial. Based on it, try to guess the answers before you start calculating anything!].

Problem 78. Let r balls be placed in m bins at random. Let X_k be the number of balls in the kth bin. Recall that $(X_1, \ldots, X_m)$ has a multinomial distribution. Find the joint distribution of (X_1, X_2) and the marginal distribution of X_1 and of X_2.

Problem 79. (*) [*Submit only parts (1) and (2)*]

(1) Let X and Y be independent integer-valued random variables with pmf f and g respectively. That is, $\mathbf{P}\{X = k\} = f(k)$ and $\mathbf{P}\{Y = k\}g(k)$ for every $k \in \mathbb{Z}$. Then, show that $X + Y$ has the pmf h given by $h(k) = \sum_{n \in \mathbb{Z}} f(n)g(k - n)$ for each $k \in \mathbb{Z}$.

(2) Let $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\mu)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Pois}(\lambda + \mu)$.

(3) Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Bin}(n + m, p)$.

(4) Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$ and assume that X and Y are independent. Show that $X + Y$ has negative binomial distribution and find the parameters.

Problem 80. (*) [*Submit only parts (1) and (2)*]

(1) Let X and Y be independent random variables with densities $f(x)$ and $g(y)$ respectively. Use the change of variable formula to show that $X + Y$ has the density $h(u)$ given by $h(u) = \int_{-\infty}^{\infty} f(s)g(u - s)ds$.

(2) Let X, Y be independent $\text{Unif}[-1, 1]$ random variables. Find the density of $X + Y$.

(3) Let $X \sim \text{Gamma}(\mu, \lambda)$ and $Y \sim \text{Gamma}(\nu, \lambda)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Gamma}(\mu + \nu, \lambda)$.

(4) Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and assume that X and Y are independent. Show that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Problem 81. Find examples of discrete probability spaces and events A, B, C so that the following happen.

(1) The events A, B, C are pairwise independent but not mutually independent.

(2) $\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A)\mathbf{P}(B)\mathbf{P}(C)$ but A, B, C are not independent.

Problem 82. (*) In each of the following cases, X and Y are independent random variables with the given distributions. You are asked to find the distribution of $X + Y$ using the convolution formula (when you encounter a named distribution, do identify it!).

(1) $X \sim \text{Gamma}(\nu, \lambda)$ and $Y \sim \text{Gamma}(\nu', \lambda)$.

(2) $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$.

(3) $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\lambda')$.

(4) $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$.

[**Note:** Submit 2, 3, 4 only]

Problem 83. Use the change of variable formula to solve the following problems.

(1) Let $X \sim \text{Pois}(\lambda)$ and $Y \sim \text{Pois}(\lambda')$ be independent. Let $Z = X + Y$. Show that the conditional distribution of X given $Z = m$ is $\text{Bin}(m, \frac{\lambda}{\lambda+\lambda'})$.

(2) Let $X \sim \text{Gamma}(\nu, \lambda)$ and $Y \sim \text{Gamma}(\nu', \lambda)$ be independent. Show that $X/(X + Y)$ has a Beta distribution and find its parameters.

(3) Let $X \sim N(0, 1)$ and $Y \sim N(0, 1)$ be independent. Show that X/Y has Cauchy distribution.

Problem 84. (*) Let (X, Y) be a bivariate normal with density

$$\frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}(a(x-\mu)^2 + b(y-\nu)^2 + 2c(x-\mu)(y-\nu))}$$

where $a, b, ab - c^2$ are all positive and μ, ν are any real numbers.

(1) Show that $\mathbf{E}[X] = \mu$, $\mathbf{E}[Y] = \nu$, $\text{Var}(X) = \sigma_{1,1}$, $\text{Var}(Y) \sim \sigma_{2,2}$ and $\text{Cov}(X, Y) = \sigma_{1,2}$ where the matrix Σ (called the covariance matrix of (X, Y)) is defined as

$$\Sigma = \begin{bmatrix} \sigma_{1,1} & \sigma_{1,2} \\ \sigma_{2,1} & \sigma_{2,2} \end{bmatrix} := \begin{bmatrix} a & c \\ c & b \end{bmatrix}^{-1}.$$

(2) Find the conditional density of Y given X. When are X and Y independent? ("When" means under what conditions on the parameters a, b, c, μ, ν or in terms of $\sigma_{1,1}, \sigma_{2,2}, \sigma_{1,2}, \mu, \nu$?).

Problem 85. In these problems, use change of variable formula in one dimension to show that the families of distributions we have defined

(1) If $X \sim \text{Exp}(\lambda)$, show that $\lambda X \sim \text{Exp}(1)$. More generally, if $X \sim \text{Gamma}(\nu, \lambda)$, show that $\lambda X \sim \text{Gamma}(\nu, 1)$.

(2) If $X \sim N(\mu, \sigma^2)$, show that $\frac{X-\mu}{\sigma} \sim N(0,1)$.

Problem 86. Let X_1, X_2, X_3 be independent random variables, each having $\mathrm{Ber}_{\pm}(1/2)$ distribution. This means $\mathbf{P}(X_1 = 1) = \mathbf{P}(X_1 = -1) = \frac{1}{2}$.

(1) Let $Y_1 = X_2 X_3$, $Y_2 = X_1 X_3$ and $Y_3 = X_1 X_2$. Show that Y_1, Y_2, Y_3 are pairwise independent (i.e., any two of them are independent) but are not independent.

(2) Can you find three events A, B, C in some probability space such that they are pair-wise independent but not independent?

Problem 87. (1) Let $A_1, \ldots, A_n$ be independent with $\mathbf{P}(A_i) = p$ for all i. Find the probability that (a) none of the events $A_1, \ldots, A_n$ occur, (b) all of the events $A_1, \ldots, A_n$ occur.

(2) Let X_1, X_2 be independent random variables, both having $\mathrm{Exp}(\lambda)$ distribution. Let $Z = \min\{X_1, X_2\}$. Show that $Z \sim \mathrm{Exp}(2\lambda)$. What if we take the minimum of n independent exponential random variables?

Problem 88. Let A, B be two events in a common probability space. Write the joint distributions (joint pmf) of the following random variables.

(1) $X = \mathbf{1}_A$ and $Y = \mathbf{1}_B$.

(2) $X = \mathbf{1}_{A \cap B}$ and $Y = \mathbf{1}_{A \cup B}$.

Problem 89. (1) Let $X \sim \mathrm{Exp}(\lambda)$. For any $t, s > 0$, show that $\mathbf{P}\{X > t + s \mid X > t\} = \mathbf{P}\{X > s\}$. (This is called the *memoryless property* of the exponential distribution).

(2) Show that if a non-negative random variable Y has memoryless property (i.e., $\mathbf{P}\{Y > t + s \mid Y > t\} = \mathbf{P}\{Y > s\}$ for all $s, t > 0$), then Y must have exponential distribution.

Problem 90. (1) Let X and Y be independent integer-valued random variables with pmf f and g respectively. That is, $\mathbf{P}\{X = k\} = f(k)$ and $\mathbf{P}\{Y = k\}g(k)$ for every $k \in \mathbb{Z}$. Then, show that $X + Y$ has the pmf h given by $h(k) = \sum_{n \in \mathbb{Z}} f(n)g(k-n)$ for each $k \in \mathbb{Z}$.

(2) Let $X \sim \mathrm{Pois}(\lambda)$ and $Y \sim \mathrm{Pois}(\mu)$ and assume that X and Y are independent. Show that $X + Y \sim \mathrm{Pois}(\lambda + \mu)$.

Problem 91. Continuation of the previous problem.

(1) Let $X \sim \mathrm{Bin}(n, p)$ and $Y \sim \mathrm{Bin}(m, p)$ and assume that X and Y are independent. Show that $X + Y \sim \mathrm{Bin}(n + m, p)$.

(2) Let $X \sim \text{Geo}(p)$ and $Y \sim \text{Geo}(p)$ and assume that X and Y are independent. Show that $X + Y$ has negative binomial distribution and find the parameters.

Problem 92. (1) Let X and Y be independent random variables with densities $f(x)$ and $g(y)$ respectively. Use the change of variable formula to show that $X + Y$ has the density $h(u)$ given by $h(u) = \int_{-\infty}^{\infty} f(s)g(u - s)ds$.

(2) Let X, Y be independent $\text{Unif}[-1, 1]$ random variables. Find the density of $X + Y$.

Problem 93. Continuation of the previous problem.

(1) Let $X \sim \text{Gamma}(\mu, \lambda)$ and $Y \sim \text{Gamma}(\nu, \lambda)$ and assume that X and Y are independent. Show that $X + Y \sim \text{Gamma}(\mu + \nu, \lambda)$.

(2) Let $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ and assume that X and Y are independent. Show that $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.

Problem 94. Let (X, Y) have the bivariate normal distribution with density

$$f(x, y) = \frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}[ax^2 + by^2 + 2cxy]}.$$

Assume that $a > 0, c > 0, ab - c^2 > 0$ so that this is a valid density.

(1) Show that the marginal distributions are one-dimensional normal and find the parameters.

(2) For what values of the parameters are X and Y independent?

Problem 95. A few more exercises in change of variable formula.

(1) If X, Y are independent $N(0, 1)$ random variables, show that X/Y has the Cauchy distribution (with density $\frac{1}{\pi(1+x^2)}$).

(2) If $X \sim \text{Gamma}(\alpha, 1)$, $Y \sim \text{Gamma}(\beta, 1)$ are independent, then show that $X + Y$ and $X/(X + Y)$ are independent, $X + Y \sim \text{Gamma}(\alpha + \beta, 1)$ and $X/(X + Y) \sim \text{Beta}(\alpha, \beta)$.

(3) If X, Y are independent $N(0, 1)$ random variables, show that $X^2 + Y^2$ has $\text{Exp}(1/2)$ distribution.

Problem 96. Find the means and variances of X in each of the following cases.

$$(a) \ X \sim \text{Bin}(n, p). \qquad (b) \ X \sim \text{Pois}(\lambda). \qquad (c) X \sim \text{Geo}(p).$$

Problem 97. Find the means and variances of X in each of the following cases.

$$(a) \ X \sim N(\mu, \sigma^2). \qquad (b) \ X \sim \text{Gamma}(\nu, \lambda). \qquad (c) \ X \sim \text{Beta}(\nu_1, \nu_2). \qquad (d) \ X \sim \text{Unif}[a, b].$$

Problem 98.   (1) Let $\xi \sim \text{Exp}(\lambda)$. For any $t, s \geq 0$, show that $\mathbf{P}\{\xi > t + s \mid \xi > t\} = \mathbf{P}\{\xi > s\}$.
   (This is called the *memoryless property* of the exponential distribution).

   (2) Show that if a non-negative random variable ξ has memoryless property (i.e., $\mathbf{P}\{\xi > t + s \mid \xi > t\} = \mathbf{P}\{\xi > s\}$), then ξ must have exponential distribution.

Problem 99. Let X be a non-negative random variable with CDF $F(t)$.

   (1) Show that $\mathbf{E}[X] = \int\limits_0^\infty (1 - F(t))dt$ and more generally $\mathbf{E}[X^p] = \int\limits_0^\infty pt^{p-1}(1 - F(t))dt$.

   (2) If X is non-negative integer valued, then $\mathbf{E}[X] = \sum_{k=1}^\infty \mathbf{P}\{X \geq k\}$.

Problem 100. (*) Find all possible joint distributions of (X, Y) such that $X \sim \text{Ber}(1/2)$ and $Y \sim \text{Ber}(1/2)$. Find the correlation for each such joint distribution.

Problem 101. Let (X, Y) have the bivariate normal distribution with density

$$f(x, y) = \frac{\sqrt{ab - c^2}}{2\pi} e^{-\frac{1}{2}\left[a(x-\mu)^2 + b(y-\nu)^2 + 2c(x-\mu)(y-\nu)\right]}.$$

   (1) Find the marginal distributions of X and of Y.

   (2) Find means and variances of X and Y and the covariance and correlation of X with Y.
   Under what conditions on the parameters are X and Y independent?

[**Note:** It is very useful to introduce the matrix $\Sigma = \begin{bmatrix} a & c \\ c & b \end{bmatrix}^{-1} = \begin{bmatrix} \frac{b}{\Delta} & -\frac{c}{\Delta} \\ -\frac{c}{\Delta} & \frac{a}{\Delta} \end{bmatrix}$ which is called the *covariance matrix* of (X, Y). The answers can be written in terms of the entries of Σ.]

Problem 102. Let r balls be placed in m bins at random. Let X_k be the number of balls in the kth bin. Recall that $(X_1, \ldots, X_m)$ has a multinomial distribution.

   (1) Find the joint distribution of (X_1, X_2) and the marginal distribution of X_1 and of X_2.

   (2) Find the means, variances, covariance and correlation of X_1 and X_2.

   (3) Let Y be the number of empty bins. Find the mean and variance of Y. [**Hint:** Write Y as $\mathbf{1}_{A_1} + \ldots + \mathbf{1}_{A_m}$ where A_k is the event that the kth bin is empty].

Problem 103. (*) A box contains N coupons where the number w_k is written on the kth coupon. Let $\mu = \frac{1}{N} \sum_{k=1}^N w_k$ be the "population mean" and let $\sigma^2 = \frac{1}{n} \sum_{i=1}^n (w_i - \mu)^2$ be the "population variance". A sample of size m is drawn from the population, the values seen are $X_1, \ldots, X_m$. The sample mean $\bar{X}_m = (X_1 + \ldots + X_m)/m$ is formed. Find the mean and variance of $\bar{X}_m$ in the following two cases.

(1) The samples are drawn with replacement (i.e., draw a coupon, note the number, put the coupon back in the box, and draw again...).

(2) The samples are drawn without replacement.

Problem 104. Find the expectation and variance for a random variable with the following distributions.

(1) (a) $\text{Bin}(n, p)$, (b) $\text{Geo}(p)$, (c) $\text{Pois}(\lambda)$, (d) $\text{Hypergeo}(N_1, N_2, m)$.

(2) (a) $N(\mu, \sigma^2)$, (b) $\text{Gamma}(\nu, \lambda)$, (c) $\text{Beta}(p, q)$.

[**Note:** Although the computations are easy, the answers you get are worth remembering as they occur in various situations.]

Problem 105. (*) Place r balls in n bins uniformly at random. Let X_k be the number of balls in the k th bin. Find $\mathbf{E}[X_k]$, $\text{Var}(X_k)$ and $\text{Cov}(X_k, X_\ell)$ for $1 \le k, \ell \le n$. [**Hint:** First do the case when $r = 1$. Then think how to use that to get the general case].

Problem 106. Suppose X, Y, Z are i.i.d. random variables with each having marginal density $f(t)$.

(1) Find $\mathbf{E}\left[\frac{X}{X+Y+Z}\right]$ (assume that it exists).

(2) Find $\mathbf{P}(X < Y > Z)$.

Problem 107. Consider the following integrals

$$\int_0^1 \frac{4}{1+x^2} dx = \pi, \qquad \int_0^1 \frac{1}{\sqrt{x(1-x)}} dx = \pi.$$

In either case, use Monte-Carlo integration with 100, 1000 and 10000 samples from uniform distribution to find approximations of π. Compare the approximations to the true value $3.1416\ldots$.

Problem 108. Recall the *coupon collector problem*. A box contains n coupons labelled $1, 2, \ldots, n$. Coupons are drawn at random from the box, repeatedly and with replacement. Let T_n be the number of draws needed till each of the coupons has appeared at least once.

(1) Show that $\mathbf{E}[T_n] \sim n \log n$ (this just means $\frac{1}{n \log n} \mathbf{E}[T_n] \to 1$).

(2) Show that $\text{Var}(T_n) \le 2n^2$.

(3) Show that $\mathbf{P}\left(\left|\frac{T_n}{n \log n} - 1\right| > \delta\right) \to 0$ for any $\delta > 0$.

[*Hint:* Consider the number of draws needed to get the first new coupon, the further number of draws needed to get the next coupon and so on].

Problem 109. (**) Recall the coupon collector problem where coupons are drawn repeatedly (with replacement) from a box containing coupons labelled $1, 2, \ldots, N$. Let T_N be the number of draws made till all the coupons are seen.

(1) Find $\mathbf{E}[T_N]$ and $\mathrm{Var}(T_N)$.

(2) Use Chebyshev's inequality to show that for any $\delta > 0$, as $N \to \infty$ we have

$$\mathbf{P}\{(1 - \delta)N \log N \le T_N \le (1 + \delta)N \log N\} \to 1.$$

Problem 110. (*) Recall the problem of a psychic guessing cards. Consider a shuffled deck of n cards and a psychic is supposed to guess the order of cards. Let M_n be the number of correct guesses.

(1) Assuming random guessing by the psychic, show that $\mathbf{E}[M_n] = 1$ and $\mathrm{Var}(M_n) = 1$. [**Hint** Write M_n as $X_1 + \ldots + X_n$ where X_k is the indicator of the event that the kth card is guessed correctly].

(2) Consider a variant of the game where the cards are dealt one by one and before each card is dealt, the psychic guesses what card it is going to be. In this case find $\mathbf{E}[M_n]$ and $\mathrm{Var}(M_n)$.

Problem 111. (**) Let X be a random variable. Let $f(a) = \mathbf{E}[|X - a|]$ (makes sense if the first moment exists) and $g(a) = \mathbf{E}[(X - a)^2]$ (makes sense if the second moment exists).

(1) Show that f is minimized uniquely at $a = \mathbf{E}[X]$.

(2) Show that the minimizers of g are precisely the medians of X (recall that a number b is a median of X if $\mathbf{P}\{X \ge t\} \ge \frac{1}{2}$ and $\mathbf{P}\{X \le t\} \ge \frac{1}{2}$).

Problem 112. (**) Let X be a non-negative random variable. Read the discussion following the problem to understand the significance of this problem.

(1) Suppose X_n takes the values n^2 and 0 with probabilities $1/n$ and $1 - (1/n)$, respectively. Compare $\mathbf{P}\{X_n > 0\}$ and $\mathbf{E}[X_n]$ for large n.

(2) Show the *second moment inequality* (aka *Paley-Zygmund inequality*): $\mathbf{P}\{X > 0\} \ge (\mathbf{E}[X])^2/\mathbf{E}[X^2]$.

[**Discussion:** Markov's inequality tells us that that the tail probability $\mathbf{P}\{X \ge t\}$ can be bounded from above using $\mathbf{E}[X]$. In particular, $\mathbf{P}\{X \ge r\mathbf{E}[X]\} \le \frac{1}{r}$. A natural question is whether there is a lower bound for the tail probability in terms of the expected value. In other words, if the mean is large, must the random variable be large with significant probability?

The first part shows that the answer is 'No' in general. The second part shows that the answer is 'Yes', provided we have control on the second moment $\mathbf{E}[X^2]$ from above. Notice why the inequality does not give any useful bound in the first part of the problem (what happens to the second moment of X_n?)]

Hoeffding's inequality: Using Chebyshev's inequality we got a bound of σ^2/nt^2 for the probability that the sample mean deviates from the population mean by more than t. This is very general. If we make more assumptions about our random variable, we can give better bounds. The following exercise is to illustrate this.

Problem 113. (Optional!) Let $X_1, \ldots, X_n$ be i.i.d. $\mathrm{Ber}_{\pm}(1/2)$. That is, $\mathbf{P}\{X_k = +1\} = \mathbf{P}\{X_k = -1\} = \frac{1}{2}$. Let $\bar{X}_n = \frac{1}{n}(X_1 + \ldots + X_n)$. Show that

$$\mathbf{P}\{|\bar{X}_n| > t\} \le 2e^{-nt^2/2}$$

by following these steps.

(1) Show that $\mathbf{P}\{\bar{X}_n > t\} \le e^{-\theta t}\left(\frac{e^{\theta/n} + e^{-\theta/n}}{2}\right)^n$ for any $\theta > 0$.

(2) Prove the inequality $e^x + e^{-x} \le 2e^{x^2/2}$ for any $x > 0$.

(3) Use the first two parts to show that $\mathbf{P}\{\bar{X}_n > t\} \le e^{-nt^2/2}$ (you must make an appropriate choice of θ depending on t).

(4) Now consider $|\bar{X}_n|$ and break $\mathbf{P}\{|\bar{X}_n| > t\}$ into two summands to get the desired inequality.

[**Note:** Here $\mu = 0$ and $\sigma^2 = 1$, and hence Chebyshev's inequality only gives the bound $\mathbf{P}\{|\bar{X}_n| > t\} \le \frac{1}{nt^2}$. Do you see that Hoeffding's inequality is better?]

Problem 114. Find the expectation and variance for a random variable with the following distributions. (a) $\mathrm{Bin}(n, p)$, (b) $\mathrm{Geo}(p)$, (c) $\mathrm{Pois}(\lambda)$, (d) $\mathrm{Hypergeo}(N_1, N_2, m)$. [**Note:** Although the computations are easy, the answers you get are worth remembering as they occur in various situations.]

Problem 115. Find the expectation and variance for a random variable with the following distributions. (a) $N(\mu, \sigma^2)$, (b) $\mathrm{Gamma}(\nu, \lambda)$, (c) $\mathrm{Beta}(p, q)$. [**Note:** Although the computations are easy, the answers you get are worth remembering as they occur in various situations.]

Problem 116. Place r balls in n bins uniformly at random. Let X_k be the number of balls in the k th bin. Find $\mathbf{E}[X_k]$, $\mathrm{Var}(X_k)$ and $\mathrm{Cov}(X_k, X_\ell)$ for $1 \le k, \ell \le n$. [**Hint:** First do the case when $r = 1$. Then think how to use that to get the general case].

Problem 117. Let X be a non-negative random variable with CDF $F(t)$.

(1) Show that $\mathbf{E}[X] = \int_0^\infty (1 - F(t))dt$ and more generally $\mathbf{E}[X^p] = \int_0^\infty pt^{p-1}(1 - F(t))dt$. [**Hint:** In showing this, you may assume that X has a density if you like, but it is not necessary for the above formulas to hold true]

(2) If X is non-negative integer valued, then $\mathbf{E}[X] = \sum_{k=1}^{\infty} \mathbf{P}\{X \geq k\}$.

Problem 118. A deck consists of cards labelled $1, 2, \ldots, N$. The deck is shuffled well. Let X be the label on the first card and let Y be the label on the second card. Find the means and variances of X and Y and the covariance of X and Y.

Problem 119. A box contains N coupons labelled $1, 2, \ldots, N$. A sample of size m is drawn from the population and the sample average $\bar{X}_m$ is computed. Find the mean and standard deviation of $\bar{X}_m$ in both the following cases.

(1) The m coupons are drawn with replacement.

(2) The m coupons are drawn without replacement (in this case, assume $m \leq N$).

Problem 120. Let $X \sim N(0, 1)$. Although it is not possible to get an exact expression for the CDF of X, show that for any $t > 0$,

$$\mathbf{P}\{X \geq t\} \leq \frac{1}{\sqrt{2\pi}} \frac{e^{-t^2/2}}{t}$$

which shows that the tail of the CDF decays rapidly. [**Hint:** Use the idea used in the proof of Markov's inequality]

The following problem is for the more mathematically minded students. You may safely skip this.

Problem 121. *The coupon collector problem.* A box contains n coupons labelled $1, 2, \ldots, n$. Coupons are drawn at random from the box, repeatedly and with replacement. Let T_n be the number of draws needed till each of the coupons has appeared at least once.

(1) Show that $\mathbf{E}[T_n] \sim n \log n$ (this just means $\frac{1}{n \log n} \mathbf{E}[T_n] \to 1$).

(2) Show that $\mathrm{Var}(T_n) \leq 2n^2$.

(3) Show that $\mathbf{P}\left(\left| \frac{T_n}{n \log n} - 1 \right| > \delta \right) \to 0$ for any $\delta > 0$.

[*Hint:* Consider the number of draws needed to get the first new coupon, the further number of draws needed to get the next coupon and so on].

Problem 122. Let $X_1, X_2, \ldots$ be i.i.d. Uniform$[1, 2]$ distribution. Let $S = X_1 + \ldots + X_{100}$. Give approximate quantiles at levels $0.01, 0.25, 0.5, 0.75, 0.99$ for S. Use CLT and normal distribution tables.

Problem 123. Let $X_1, \ldots, X_n$ be i.i.d. samples from a parametric family of discrete distributions. In each of the following cases, find the MLE for the unknown parameter(s) and find the bias.

(1) X_i are i.i.d. $\text{Ber}(p)$ where p is unknown.

(2) X_i are i.i.d. $N(\mu, \sigma^2)$ where μ, σ^2 are unknown.

Problem 124. Let $X_1, \ldots, X_n$ be i.i.d. samples from a parametric family of discrete distributions. In each of the following cases, find the MLE for the unknown parameter(s) and calculate the bias.

(1) X_i are i.i.d. $\text{Geo}(p)$ where p is unknown.

(2) X_i are i.i.d. $\text{Unif}[a, b]$ where a, b are unknown.

Problem 125. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. samples from a bivariate distribution. Let $\tau = \text{Cov}(X_1, Y_1)$. Let $r_n = \frac{1}{n} \sum_{k=1}^{n} (X_k - \bar{X}_n)(Y_k - \bar{X}_n)$ be the sample covariance.

(1) Show that r_n is a biased estimate for τ and find the bias.

(2) Modify the estimate r_n to get an unbiased estimate of τ.

[**Remark:** It is often convenient, here and elsewhere, to realise that $\tau = \mathbf{E}[X_1 Y_1] - \mathbf{E}[X_1]\mathbf{E}[Y_1]$ and $r_n = \left(\frac{1}{n} \sum_{k=1}^{n} X_k Y_k \right) - \bar{X}_n \bar{Y}_n$.]

Problem 126. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables from a distribution F. Let M_n be a median of $X_1, \ldots, X_n$. Assume that the distribution F has a unique median, that is there is a unique number m such that $F(m) = \frac{1}{2}$. For any $\delta > 0$ show that $\mathbf{P}\{|M_n - m| \geq \delta\} \to 0$ as $n \to \infty$.

[**Remark:** The above statement justifies using the sample median to estimate the population median, in the sense that at least for large sample sizes, the two are close. Similar justification for using sample mean to estimate expected value came from the law of large numbers]

The following problem is only for those mathematically minded.

Problem 127. Let $X_1, X_2, \ldots$ be i.i.d. $\text{Pois}(\lambda)$ random variables. Work out the exact distribution of $X_1 + \ldots + X_n$ and use it show the central limit theorem in this case. That is, show that for any $a < b$,

$$\mathbf{P}\left\{a \leq \frac{\sqrt{n}(\bar{X}_n - \lambda)}{\sqrt{\lambda}} \leq b\right\} \longrightarrow \mathbf{P}\{a \leq Z \leq b\}$$

where $Z \sim N(0, 1)$.

[**Remark:** This is analogous to the two cases of CLT that we showed in class, for exponential and for Bernoulli random variables].

The following problem shows that in certain situations, sums of random variables are approximately Poisson distributed. This gives a hint as to why Poisson distribution arises in many contexts. The question may be ignored safely from the exam point of view.

23

Problem 128. Let $X_{n,1}, X_{n,2}, \ldots X_{n,n}$ be i.i.d. Ber(p_n) random variables. Let $S_n = X_{n,1} + \ldots + X_{n,n}$. If $np_n \to \lambda$ (a finite positive number), show that S_n has approximately Pois(λ) distribution in the sense that for any $k \in \mathbb{N}$,

$$\mathbf{P}\{S_n = k\} \to e^{-\lambda}\frac{\lambda^k}{k!}.$$

[**Remark:** In contrast, if $np_n \to \infty$, deduce from CLT that S_n has approximately a normal distribution, i.e.,

$$\mathbf{P}\left\{a \leq \frac{S_n - \mathbf{E}[S_n]}{\sqrt{\mathrm{Var}(S_n)}} \leq b\right\} \to \mathbf{P}\{a \leq Z \leq b\}$$

for any $a < b$.]