信息学中的概率统计：作业一

截止日期：2024 年 9 月 27 日（周五）下课前。**如无特殊情况，请不要提交电子版！**

第一题

对于 n 个事件 $A_1, A_2, \ldots, A_n$，从概率的公理化定义和条件概率的定义出发证明下述结论。

(1) 一般加法公式：

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i) - \sum_{1\le i<j\le n} P(A_iA_j) + \sum_{1\le i<j<k\le n} P(A_iA_jA_k) + \ldots + (-1)^{n-1} P(A_1A_2\ldots A_n)。$$

(2) 一般 Union Bound：

$$P\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} P(A_i)。$$

(3) 一般乘法公式：若 $P(A_1A_2\ldots A_n) > 0$，有

$$P(A_1A_2\ldots A_n) = P(A_1)\cdot P(A_2 \mid A_1)\cdot P(A_3 \mid A_1A_2)\cdot \ldots \cdot P(A_n \mid A_1A_2\ldots A_{n-1})。$$

第二题

对于三个事件 A，B 和 C，若 $P(C) > 0$，我们称事件 A 和 B 在事件 C 发生时是条件独立的，当且仅当

$$P(AB \mid C) = P(A \mid C)P(B \mid C)。$$

对于下述命题，从概率的公理化定义和条件概率的定义出发给出证明，或给出反例。

(1) 事件 A 和 B 在事件 C 发生时是条件独立的，且有 $0 < P(C) < 1$，则事件 A 和 B 在事件 $\overline{C}$ 发生时条件独立。这里，事件 $\overline{C}$ 是事件 C 的对立事件。

(2) 事件 A 和 B 相互独立，则对于任意事件 C，若 $P(C) > 0$，事件 A 和 B 在事件 C 发生时是条件独立的。

(3) 事件 A 和 B 相互独立，则事件 A 和事件 $\overline{B}$ 相互独立。这里，事件 $\overline{B}$ 是事件 B 的对立事件。

第三题

在课上，我们考虑了如下球与桶模型：有 $n \ge 1$ 个球，每个球都等可能被放到 $m \ge 1$ 个桶中的任一个。用 $P_{n,m}$ 表示每个桶中至多有一个球的概率。在课上，我们已经证明了，

$$P_{n,m} \le e^{-\frac{n(n-1)}{2m}}。$$

现在，请证明

$$P_{n,m} \geq e^{-\frac{n(n-1)}{2m}} \cdot \left(1 - \frac{8n^3}{m^2}\right)。$$

提示：证明对于任意 $0 \leq x \leq 1/2$,

$$\ln(1-x) \geq -x - x^2。$$

第四题

将一枚骰子投掷 $n \geq 1$ 次，求在 n 次投掷中，六个数字均出现过至少一次的概率。

第五题

某路由器有 A 和 B 两种运行模式。路由器每天有等概率以 A 模式或者 B 模式运行，且每天的运行模式均独立。当以 A 模式运行时，有 90% 的概率网络堵塞，有 10% 的概率网络正常。当以 B 模式运行时，有 10% 的概率网络堵塞，有 90% 的概率网络正常。若某两天观测到网络堵塞，求这两天路由器均以 A 模式运行的概率。

第六题

对于自然数 n，m 和 k，满足 $m \geq 2k$。有 $2n$ 个 $\{1, 2, \ldots, m\}$ 的子集 $A_1, B_1, A_2, B_2, \ldots, A_n, B_n \subseteq \{1, 2, \ldots, m\}$，满足

- 对于任意 $1 \leq i \leq n$，有 $|A_i| = |B_i| = k$；
- 对于任意 $1 \leq i \leq n$，有 $A_i \cap B_i = \emptyset$；
- 对于任意 $1 \leq i, j \leq n$，若 $i \neq j$，有 $A_i \cap B_j \neq \emptyset$。

(1) 考虑一个 $\{1, 2, \ldots, m\}$ 的随机排列，每一种排列均等概率出现。对于任意 $1 \leq i \leq n$，事件 U_i 表示在随机排列中，集合 A_i 中的元素均排在 B_i 前面。证明

$$P(U_i) = \frac{1}{\binom{2k}{k}}。$$

(2) 证明 $n \leq \binom{2k}{k}$。提示：考虑事件 $\bigcup_{i=1}^{n} U_i$ 的概率。

(3) 对于 $n = \binom{2k}{k}$，构造满足条件的 $A_1, B_1, A_2, B_2, \ldots, A_n, B_n \subseteq \{1, 2, \ldots, m\}$。这里 m 可取任意自然数。

信息学中的概率统计：作业二

截止日期：2024 年 10 月 18 日（周五）下课前。**如无特殊情况，请不要提交电子版！**

第一题

对于任意 $a \geq 1$，构造**非负**离散随机变量 X，使得 $P(X \geq a \cdot E(X)) = 1/a$。

第二题

在课上，我们介绍了 n 重伯努利试验。如果某个随机试验只有两个可能的结果 A 和 $\overline{A}$，且 $P(A) = p$，将试验独立地重复进行 n 次，令 X 表示结果 A 的发生次数。在课上，我们利用二项式系数的性质证明了 $E(X) = np$。在本题中，我们将用另一种方法计算 $E(X)$ 和 $E(X^2)$。

(1) 对于任意 $t \in \mathbb{R}$，计算 $E\left(e^{Xt}\right)$。

(2) 对于任意 $t \in \mathbb{R}$，证明

$$E\left(e^{Xt}\right) = \sum_{i=0}^{\infty} \frac{t^i}{i!} \cdot E(X^i)。$$

提示：对于固定的 $0 \leq k \leq n$，考虑对 e^{kt} 应用泰勒公式。

(3) 利用上一问中的结论，计算 $E(X)$ 和 $E(X^2)$。提示：令 $f(t) = E\left(e^{Xt}\right)$。如何利用上一问中的结论，通过 $f(t)$ 求得 $E(X)$ 和 $E(X^2)$?

第三题

在课上，我们考虑了如下球与桶模型：有 n 个球，每个球都等可能被放到 m 个桶中的任一个。在本题中，我们考虑 $m = n$ 的情况，并假设 $n = m \geq 2$。

(1) 随机变量 X_i 表示第 i 个桶中球的数量。对于任意 $i \in \{1, 2, \ldots, n\}$，证明 $E(X_i) = 1$。

(2) 对于任意 $i \in \{1, 2, \ldots, n\}$ 和任意 $1 \leq k \leq n$，证明 $P(X_i = k) \leq \frac{1}{k!}$。

(3) 定义随机变量 $Y = \max\{X_1, X_2, \ldots, X_n\}$，证明 $P(Y \geq 4 \log_2 n) \leq 1/n$。提示：考虑使用 Union Bound。

(4) 证明 $E(Y) \leq 5 \log_2 n$。

第四题

给定离散随机变量 X，证明对于任意实数 c，$E((X - c)^2) \geq \mathrm{Var}(X)$。

第五题

给定离散随机变量 X, 假设其期望 $E(X)$ 和标准差 $\sigma(X)$ 均存在。对于任意实数 m, 若满足 $P(X \geq m) \geq 1/2$ 且 $P(X \leq m) \geq 1/2$，证明 $|E(X) - m| \leq \sqrt{2}\sigma$。

第六题

令 $X \sim \pi(\lambda)$，也即随机变量 X 服从参数为 $\lambda > 0$ 的泊松分布。

(1) 对于任意实数 t，计算 $E(e^{tX})$。

(2) 证明对于任意实数 $x > 0$,

$$P((x/\lambda)^X \geq (x/\lambda)^x) \leq \frac{e^{-\lambda}(e\lambda)^x}{x^x}。$$

(3) 证明对于任意 $x > \lambda$,

$$P(X \geq x) \leq \frac{e^{-\lambda}(e\lambda)^x}{x^x},$$

且对于任意 $0 < x < \lambda$,

$$P(X \leq x) \leq \frac{e^{-\lambda}(e\lambda)^x}{x^x}。$$

(4) 证明

$$P(|X - \lambda| \geq 0.2\lambda) \leq 2 \cdot e^{-0.01\lambda}。$$

信息学中的概率统计：作业三

截止日期：2024 年 11 月 1 日（周五）下课前。**如无特殊情况，请不要提交电子版！**

第一题

(1) X 为离散随机变量，且 X 仅取非负整数值。证明 $E(X)=\sum_{x=0}^{+\infty}P(X>x)$。

(2) X 为连续随机变量，且 X 仅取非负实数值。证明 $E(X)=\int_0^{+\infty}P(X>x)dx$。

第二题

在 Unix 操作系统中，用随机变量 X 表示一个随机的任务所需的内存。历史数据表明，对于任意实数 $x\geq 1$，$P(X>x)=1/x^{\alpha}$。这里 $\alpha\in(0,2)$ 是固定常数。

(1) 计算随机变量 X 的概率分布函数和概率密度函数。

(2) 计算 $E(X)$ 和 $E(X^2)$

第三题

(1) 对于任意实数 $x>0$，证明

$$\int_x^{+\infty}\frac{t}{x}e^{-t^2/2}dt=\frac{e^{-x^2/2}}{x}。$$

(2) 令 $X\sim N(0,1)$，也即连续随机变量 X 服从标准高斯分布，证明对于任意实数 $x>0$,

$$P(X\geq x)\leq\frac{e^{-x^2/2}}{x\sqrt{2\pi}}。$$

(3) 令 $Y\sim N(\mu,\sigma)$，证明对于任意实数 $k>0$,

$$P(|Y-\mu|\leq k\sigma)\geq 1-\frac{e^{-k^2/2}}{k}\cdot\sqrt{\frac{2}{\pi}}。$$

第四题

随机变量 X 的分布函数 $F(x)$ 为严格单调增的连续函数，其反函数存在。证明 $Y=F(X)$ 服从 $(0,1)$ 上的均匀分布 $U(0,1)$。

第五题

对于实数参数 μ 和 $b > 0$，已知连续随机变量 X 的概率密度函数 $f(x)$ 满足对于任意实数 x，

$$f(x) = c \cdot e^{-|x-\mu|/b}。$$

这里 c 为某个与参数 μ 和 b 有关的常数。

(1) 计算常数 c 以及 X 的分布函数

(2) 计算 $E(X)$ 和 $\mathrm{Var}(X)$

第六题

(1) 若 $X \sim N(0,1)$，对于任意实数 t，计算 $E(e^{tX^2})$

(2) 对于正整数 n，若 $Y_n \sim \chi^2(n)$，也即 $Y_n \sim \Gamma(n/2, 1/2)$。对于任意实数 t，计算 $E(e^{tY_n})$

(3) 若 $X \sim N(0,1)$，计算 $Y = X^2$ 的概率密度函数

信息学中的概率统计：作业四

截止日期：2024 年 11 月 15 日（周五）下课前。**如无特殊情况，请不要提交电子版！**

第一题

一个盒子中有 n 个小球，编号分别为 $1, 2, \ldots, n$。从盒子中取出 $k \leq n$ 个小球，每次等概率从盒子中剩余的小球中取出一个，且每次取完后均不放回。也即，第一次取小球时，每个小球被取出的概率均为 $\frac{1}{n}$。第二次取小球时，剩余的 $n-1$ 个小球各自被取出的概率均为 $\frac{1}{n-1}$。以此类推，直至一共取出 k 个小球。

定义随机变量 $X_1, X_2, \ldots, X_k$，其中 X_i（$1 \leq i \leq k$）表示第 i 次取出小球的编号。

(1) 对于 $1 \leq i < j \leq k$，判断 X_i 是否与 X_j 相互独立。

(2) 计算 $X_1, X_2, \ldots, X_k$ 的联合分布列。

(3) 对于 $1 \leq i \leq k$，计算 X_i 的边际分布列。

(4) 对于任意 $1 \leq i < j \leq k$ 和 $1 \leq a_i, a_j \leq n$，计算 $P(X_i = a_i \cap X_j = a_j)$。

(5) 利用恒等式 $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$，对于 $1 \leq i \leq k$，计算 $E(X_i)$ 和 $\mathrm{Var}(X_i)$。

(6) 对于 $1 \leq i < j \leq k$，计算 $\mathrm{Cov}(X_i, X_j)$。

第二题

将 n 个编号为 $1, 2, \ldots, n$ 的小球随机打乱，每一种排列等概率出现。用 $\pi_1, \pi_2, \ldots, \pi_n$ 表示随机打乱后每个位置上小球的编号，也即 π_i 表示随机打乱后，位置为 i 的小球的原始编号。对于 $1 < i < n$，我们称 i 是一个*局部极大值*，当且仅当 $\pi_i > \pi_{i-1}$ 且 $\pi_i > \pi_{i+1}$。令随机变量 X 表示所有 $1 < i < n$ 中局部极大值的总数量。计算 $E(X)$。

第三题

令随机变量 $X \sim G(p)$，也即随机变量 X 服从参数为 p 的几何分布。证明

$$E(X^2) = p + E((X+1)^2)(1-p),$$

$$E(X^3) = p + E((X+1)^3)(1-p),$$

并计算 $E(X^2)$ 和 $E(X^3)$。

第四题

令 $X_1, X_2, \ldots$ 为一列同分布的离散随机变量。离散随机变量 N 取正整数值，且 $N, X_1, X_2, \ldots,$ 相互独立。在课上，我们证明了 $E\left(\sum_{i=1}^{N} X_i\right) = E(N) \cdot E(X_1)$。

(1) 给出例子使得

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) \neq E(N) \cdot \mathrm{Var}(X_1)。$$

(2) 证明

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = E(N) \cdot \mathrm{Var}(X_1) + \mathrm{Var}(N)(E(X_1))^2。$$

第五题

(1) 对于正整数 r 和实数 $0 < p < 1$，给定 $X \sim NB(1, p)$，$Y \sim NB(r, p)$，也即随机变量 X 服从参数为 $1, p$ 的负二项分布，随机变量 Y 服从参数为 r, p 的负二项分布。若 X 和 Y 相互独立，证明 $X + Y \sim NB(r + 1, p)$，也即 $X + Y$ 服从参数为 $r + 1, p$ 的负二项分布。提示：使用恒等式

$$\binom{n}{m} = \binom{n-1}{m-1} + \binom{n-1}{m}。$$

(2) 对于正整数 r，$X_1, X_2, \ldots, X_r$ 为独立同分布的随机变量，且均服从 $G(p)$，也即参数为 p 的几何分布。证明 $X_1 + X_2 + \cdots + X_r \sim NB(r, p)$。

信息学中的概率统计：作业五

截止日期：2024 年 11 月 29 日（周五）下课前。**如无特殊情况，请不要提交电子版！**

第一题

给定二维随机变量 X, Y。证明 $\mathrm{Corr}(X, Y) = \pm 1$ 当且仅当存在实数 $a \neq 0, b$，使得 $P(Y = aX + b) = 1$。

提示：利用结论（无需证明），若随机变量 Z 满足 $\mathrm{Var}(Z) = 0$，则 $P(Z = E(Z)) = 1$。

第二题

对于 $\sigma_1 > 0, \sigma_2 > 0, -1 < \rho < 1$，二维随机变量 $U, V \sim N(0, 0, \sigma_1^2, \sigma_2^2, \rho)$。本题中，我们将计算 $E(\mathrm{ReLU}(U) \cdot \mathrm{ReLU}(V))$。这里，$\mathrm{ReLU}(x) = \max\{x, 0\}$。

设二维随机变量 $(X, Y) \sim N(0, 0, 1, 1, \rho)$，令二维随机变量 (R, Θ) 满足 $R \geq 0, \Theta \in [0, 2\pi]$，且

$$\begin{cases} X = R \cdot (\sqrt{1-\rho^2} \cdot \cos\Theta + \rho \cdot \sin\Theta) = R \cdot \sin(\arccos\rho + \Theta) \\ Y = R\sin\Theta \end{cases} 。$$

(1) 令 $x = r \cdot (\sqrt{1-\rho^2} \cdot \cos\theta + \rho \cdot \sin\theta)$, $y = r\sin\theta$。验证 $x^2 + y^2 - 2\rho xy = r^2(1-\rho^2)$。

(2) 计算 R, Θ 的联合密度函数，R 和 Θ 的各自的边际密度函数，并判断 R 和 Θ 的独立性。

(3) 计算 $E(\mathrm{ReLU}(X) \cdot \mathrm{ReLU}(Y))$。提示：利用结论（无需证明）

$$\int_0^{+\infty} x^3 e^{-x^2/2} dx = 2,$$

$$\int_0^{\pi - \arccos\rho} (\rho \cdot \sin^2\theta + \sqrt{1-\rho^2}\sin\theta\cos\theta) d\theta = \frac{1}{2}\left(\rho(\pi - \arccos\rho) + \sqrt{1-\rho^2}\right) 。$$

(4) 验证 $(\sigma_1 X, \sigma_2 Y)$ 与 (U, V) 服从相同的分布。

(5) 计算 $E(\mathrm{ReLU}(U) \cdot \mathrm{ReLU}(V))$。

第三题

在课上我们考虑了如下矩阵 $A \in \mathbb{R}^{n \times n}$：对于任意 $1 \leq i, j \leq n$，$A_{i,j} \sim N(0, 1)$，且不同元素相互独立。计算 $E(\mathrm{trace}(A^3))$ 和 $E(\mathrm{trace}(A^4))$。提示：首先考虑 $n = 1$ 的情况，并参考作业三第六题。

第四题

(1) 令 $X_1, X_2, \ldots, X_n$ 为独立同分布随机变量，且 $X_i \sim N(0,1)$。令 $Y = \sum_{i=1}^{n} X_i^2$。对于任意实数 $t \in [0, 1/4)$，证明

$$E(e^{t(Y-n)}) \leq e^{2t^2 n}。$$

提示：首先考虑 $n = 1$ 的情况，并参考作业三第六题，以及作业一第三题的提示。

(2) 对于任意 $0 \leq \Delta < 1$，证明

$$P(Y \geq (1+\Delta)n) \leq e^{-n\Delta^2/8}。$$

提示：根据 $0 \leq \Delta < 1$，选择合适的 t 使得 $t \in [0, 1/4)$，并使用马尔可夫不等式。

(3) 对于任意 $0 \leq \Delta < 1$，证明

$$P(Y \leq (1-\Delta)n) \leq e^{-n\Delta^2/8}。$$

信息学中的概率统计：作业六

截止日期：2024 年 12 月 13 日（周五）下课前。**如无特殊情况，请不要提交电子版！**
注意：本次作业第六题为附加题，正确解决该题目本次作业可以得到额外 30% 的分数。

第一题

令 $X \sim \mathrm{Exp}(\lambda)$，$\lambda > 0$。本题中，我们将对 $a > 1$ 给出 $P(X \geq a/\lambda)$ 的上界。

(1) 使用马尔可夫不等式，给出 $P(X \geq a/\lambda)$ 的上界。

(2) 使用切比雪夫不等式，证明 $P(X \geq a/\lambda) \leq \frac{1}{(a-1)^2}$。

(3) 使用 Chernoff Bound，证明 $P(X \geq a/\lambda) \leq a \cdot e^{-a+1}$。

(4) 计算 $P(X \geq a/\lambda)$ 的准确值。

第二题

在课上，我们介绍了随机变量的收敛性。设 $\{X_n\}$ 为一列随机变量，X 为另一随机变量。如果对于任意 $\epsilon > 0$，有

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1,$$

则称 $\{X_n\}$ 依概率收敛于 X，写作 $X_n \xrightarrow{P} X$。在本题中，我们将介绍随机变量的另一种收敛性。

设 $\{X_n\}$ 为一列随机变量，X 为另一随机变量。如果 $P(\lim_{n\to\infty} X_n \to X) = 1$，也即对于任意 $\epsilon > 0$，

$$\lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} |X_m - X| \geq \epsilon\right) = 0,$$

则称 $\{X_n\}$ 几乎必然收敛于 X，写作 $X_n \xrightarrow{a.s.} X$。

(1) 令 $\{X_n\}$ 为一列相互独立的随机变量，且 $X_n \sim B(1, 1/n)$。证明 $\{X_n\}$ 依概率收敛于 0，但 $\{X_n\}$ 不几乎必然收敛于 0。

(2) 令 $\{X_n\}$ 为一列独立同分布的随机变量，$X_n \sim B(1, p)$。令 $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_n$。证明 $Y_n \xrightarrow{a.s.} p$。

第三题

某个不使用随机性的计算机程序 A，为了输出正确结果，该程序需要对另一计算机程序 B 进行 T 次调用，每次调用使用可能不同的输入，且每次调用使用的输入可能依赖于之前对程序 B 的调用返回的结果。程序 A 使用对程序 B 的 T 次调用返回的结果以输出最终结果 θ。具体来说，假设对程序 B 进行 T 次调用返回的结果为 $\omega_1, \omega_2, \ldots, \omega_T$，在正确得到 $\omega_1, \omega_2, \ldots, \omega_T$ 的前提下，程序 A 总是能输出正确的结果 θ。

现有计算机程序 B'。在同样的输入下，程序 B' 以 2/3 的概率返回与程序 B 相同的结果，以 1/3 的概率返回不同的结果。现在，在没有程序 B，仅有程序 A 和程序 B' 的情况下，设计一个方案，以 $1 - \delta$ 的概率得到正确结果 θ。该方案对程序 A 和程序 B' 的调用次数应与 T 和 $\log(1/\delta)$ 为多项式关系。

第四题

在课上，我们用 Chernoff bound 证明了下述不等式：若 $X \sim B(n,p)$，则

$$P(X \geq E(X) + n\epsilon) \leq e^{-2n\epsilon^2},$$

$$P(X \leq E(X) - n\epsilon) \leq e^{-2n\epsilon^2}。$$

在本题中，我们将对二项分布证明另一版本的 Chernoff bound。

(1) 证明 $M_X(t) \leq e^{(e^t-1)\cdot E(X)}$。提示：使用不等式 $1+x \leq e^x$。

(2) 证明对于任意 $\epsilon > 0$，

$$P(X \geq (1+\epsilon)E(X)) \leq \left(\frac{e^\epsilon}{(1+\epsilon)^{1+\epsilon}}\right)^{E(X)};$$

对于任意 $0 < \epsilon < 1$，

$$P(X \leq (1-\epsilon)E(X)) \leq \left(\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right)^{E(X)}。$$

提示：参考作业二第六题。

(3) 利用 (2) 中的结论，重新证明作业二第二题 (3)。也即，有 n 个球，每个球都等可能被放到 $m = n$ 个桶中的任一个。令 X_i 表示第 i 个桶中球的数量，$Y = \max\{X_1, X_2, \ldots, X_n\}$。证明 $P(Y \geq 4\log_2 n) \leq 1/n$。

第五题

在课上，我们证明了下述结论：对于任意向量 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$，令 $A \in \mathbb{R}^{k\times d}$ 为随机矩阵，A 的不同元素独立同分布且均服从 $N(0,1)$，$k = O(\log n/\epsilon^2)$，则以至少 $1/2$ 的概率，对于任意 $1 \leq i, j \leq n$，

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \left\|\frac{1}{\sqrt{k}}A(x_i - x_j)\right\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2,$$

也即令 $F(x) = \frac{1}{\sqrt{k}}Ax$ 为一随机线性变换，则以至少 $1/2$ 的概率，$F(x)$ 保持了每一对 x_i 和 x_j 之间的距离。

证明该结论的核心工具是下述引理：对于任意 $x \in \mathbb{R}^d$，

$$P\left((1-\epsilon)\|x\|^2 \leq \left\|\frac{1}{\sqrt{k}}Ax\right\|^2 \leq (1+\epsilon)\|x\|^2\right) \geq 1 - 2e^{-k\epsilon^2/8}。 \tag{1}$$

为了证明原结论，对所有可能的 $x = x_i - x_j$ 使用上述结论，并使用 Union bound。

在本题中，我们将证明随机线性变换 $F(x) = \frac{1}{\sqrt{k}}Ax$ 不仅可以保持每一对 x_i 和 x_j 之间的距离，还可以保持每一对 x_i 和 x_j 之间的点积。在本题中，对于向量 $a, b \in \mathbb{R}^d$，$\langle a, b\rangle = a^\top b$ 为向量 a 与 b 的点积。

(1) 考虑向量 $y_1, y_2, \ldots, y_n \in \mathbb{R}^d$，对于全部 $1 \leq i \leq n$，满足 $\|y_i\| = 1$。令 $A \in \mathbb{R}^{k\times d}$ 为随机矩阵，A 的不同元素独立同分布且均服从 $N(0,1)$，$k = O(\log n/\epsilon^2)$ 。证明以至少 $1/2$ 的概率，下述事件同时成立：

- 对于任意 $1 \leq i \leq n$，$(1-\epsilon/4)\|y_i\|^2 \leq \left\|\frac{1}{\sqrt{k}}Ay_i\right\|^2 \leq (1+\epsilon/4)\|y_i\|^2$；
- 对于任意 $1 \leq i, j \leq n$ 且 $i \neq j$，$(1-\epsilon/4)\|y_i + y_j\|^2 \leq \left\|\frac{1}{\sqrt{k}}A(y_i + y_j)\right\|^2 \leq (1+\epsilon/4)\|y_i + y_j\|^2$。

(2) 在（1）中结论的基础上，证明以至少 $1/2$ 的概率，对于任意 $1 \leq i, j \leq n$，

$$\left|\left\langle \frac{1}{\sqrt{k}}Ay_i, \frac{1}{\sqrt{k}}Ay_j \right\rangle - \langle y_i, y_j\rangle\right| \leq \epsilon。$$

(3) 考虑向量 $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$。注意 x_i 不一定满足 $\|x_i\| = 1$。证明以至少 1/2 的概率，对于任意 $1 \le i, j \le n$，

$$\left| \left\langle \frac{1}{\sqrt{k}} A x_i, \frac{1}{\sqrt{k}} A x_j \right\rangle - \langle x_i, x_j \rangle \right| \le \epsilon \|x_i\| \|x_j\|。$$

第六题

在课上，我们证明了对于任意 $S_1, S_2, \ldots, S_m \subseteq \{1, 2, \ldots, n\}$，存在 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，使得对于任意 $1 \le i \le m$，

$$\mathrm{disc}_\chi(S_i) = \left| \sum_{j \in S_i} \chi(j) \right| \le O(\sqrt{n \log m})。$$

在本题中，我们将证明存在 $S_1, S_2, \ldots, S_n \subseteq \{1, 2, \ldots, n\}$，对于任意 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，存在 $1 \le i \le n$ 使得

$$\mathrm{disc}_\chi(S_i) = \left| \sum_{j \in S_i} \chi(j) \right| \ge \Omega(\sqrt{n}),$$

也即课上给出的上界 $O(\sqrt{n \log m})$ 几乎是最优的。

(1) 证明下述反集中不等式：$X \sim B(n, 1/2)$，存在常数 $c_1, c_2 > 0$，使得

$$P(X \ge n/2 + c_1 \cdot \sqrt{n}) \ge c_2。$$

提示：该不等式有多种证明方法。一种可能的思路是首先使用定量化的中心极限定理（课上提到的 Berry–Esseen 定理）建立二项分布与标准正态分布的联系，之后对标准正态分布证明反集中不等式。

(2) 令 S 为 $\{1, 2, \ldots, n\}$ 的子集，对于每个 $j \in \{1, 2, \ldots, n\}$，$P(j \in S) = 1/2$，且不同 j 是否被包含在 S 中相互独立。利用（1）中的结论，证明存在常数 $c_3, c_4 > 0$，对于任意 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，

$$P\left(\left| \sum_{j \in S} \chi(j) \right| \ge c_3 \sqrt{n} \right) \ge c_4。$$

(3) 证明存在 $m = O(n)$（也即对于某个常数 C，$m \le Cn$）个集合 $S_1, S_2, \ldots, S_m \subseteq \{1, 2, \ldots, n\}$ 和常数 $c > 0$，对于任意 $\chi : \{1, 2, \ldots, n\} \to \{-1, +1\}$，存在 $1 \le i \le m$ 使得

$$\left| \sum_{j \in S_i} \chi(j) \right| \ge c\sqrt{n}。$$

提示：考虑使用概率证法，将 $S_1, S_2, \ldots, S_m$ 取为 $\{1, 2, \ldots, n\}$ 独立同分布的随机子集，并扩展（2）中的分析。

(4) 证明当 $m = n$ 时，（3）中的结论同样成立。

信息学中的概率统计：作业七

截止日期：2024 年 12 月 27 日（周五）下课前。**如无特殊情况，请不要提交电子版！**
注意：本次作业第五题第二问为附加题，正确解决该问可以得到额外 15% 的分数。

第一题

给定未知参数 θ 的估计量 $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$，证明

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}(\hat{\theta}))^2 = \mathrm{Var}(\hat{\theta}) + (\theta - E(\hat{\theta}))^2$$

第二题

令总体 X 服从概率密度函数如下的连续分布，其中 $\theta > 0$ 为未知参数，

$$f(x) = \begin{cases} \frac{\theta}{x^2} & x \geq \theta \\ 0 & x < \theta \end{cases}。$$

给定简单随机样本 $X_1, X_2, \ldots, X_n$，给出 θ 的最大似然估计量。

第三题

令总体 $X \sim \pi(\lambda)$，也即参数为 λ 的泊松分布，λ 为未知参数。给定简单随机样本 $X_1, X_2, \ldots, X_n$，本题中，我们将考虑 $p = e^{-\lambda}$ 的两个不同的估计量。

(1) 考虑 p 的矩法估计量 $\hat{p}_1 = e^{-\overline{X}}$。这里，$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ 为样本均值。判断 $\hat{p}_1$ 是否为 $p = e^{-\lambda}$ 的最大似然估计（简要说明原因，无需严格证明），判断 $\hat{p}_1$ 是否为无偏估计量，渐进无偏估计量，一致估计量，并计算 $\hat{p}_1$ 的均方误差。提示：参考作业二第六题。

(2) 令 $\hat{p}_2 = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i=0}$。这里

$$1_{X_i=0} = \begin{cases} 1 & X_i = 0 \\ 0 & X_i > 0 \end{cases}。$$

判断 $\hat{p}_2$ 是否为无偏估计量，渐进无偏估计量，一致估计量，并计算 $\hat{p}_2$ 的均方误差。

第四题

给定样本 $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$，$Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$，满足 $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ 相互独立。

(1) 令 $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$，$\overline{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$。给出 $X - Y$ 服从的分布。

(2) 假定 σ_1^2 和 σ_2^2 均已知，利用上一问中的结果构造枢轴量并给出 $\mu_1 - \mu_2$ 的置信水平为 $1 - \alpha$ 置信区间。最终结果应依赖于 $\Phi^{-1}(1 - \alpha/2)$，其中 $\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$ 为标准正态分布的分布函数。

(3) 同样假定 σ_1^2 和 σ_2^2 均已知，利用 Chernoff bound，给出 $\mu_1 - \mu_2$ 的置信水平为 $1-\alpha$ 置信区间。最终结果不应依赖于标准正态分布的分布函数。

第五题

在课上，我们考虑了下述模型：给定 n 台游戏机，第 i 台游戏机的中奖概率为 $0 \le p_i \le 1$，且 p_i 均为未知参数。在第 t 轮中，选择一台游戏机 $1 \le i \le n$，并观测到结果 $X_t \sim B(1, p_i)$。这里 $X_1, X_2, \ldots$ 相互独立。

在课上，我们考虑了下述均匀采样策略：对每台游戏机进行 N 次观测，并返回样本均值最大的游戏机。若取 $N = O(\ln n/\epsilon^2))$，则有 $P(p_o \ge \max p_i - \epsilon) \ge 2/3$，这里 $1 \le o \le n$ 为策略返回的选择。

本题中，我们考虑 $n = 2$ 的情况，也即给定两台游戏机，中奖概率分别为 p_1 和 p_2，且 p_1 和 p_2 均为未知参数。令 $\Delta = |p_1 - p_2|$。

(1) 若 Δ 为已知参数且 $\Delta > 0$, 证明采用均匀采样策略并令 $N = O(1/\Delta^2)$, 则有 $P(p_o = \max\{p_1, p_2\}) \ge 2/3$, 这里 $o = 1$ 或 $o = 2$ 为策略返回的选择。

(2) 若 Δ 为未知参数且 $\Delta > 0$，设计策略，使得以至少 $2/3$ 的概率，下述事件同时成立：

- $p_o = \max\{p_1, p_2\}$，这里 $o = 1$ 或 $o = 2$ 为策略返回的选择；
- 策略的总观测次数与 $1/\Delta$ 为多项式关系。

本问为附加问，正确解决该问可以得到额外 15% 的分数。

信息学中的概率统计：作业八

截止日期：2024 年 1 月 3 日（周五）期末考试前。

第一题

给定 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，其中 $y_i = \alpha + \beta x_i + \epsilon_i$，$\epsilon_i$ 相互独立，且 ϵ_i 服从拉普拉斯分布，其概率密度函数（参考作业三第五题）满足对于任意实数 $x \in \mathbb{R}$，

$$f(x) = \frac{1}{2b} e^{-|x|/b},$$

这里 α, β 和 $b > 0$ 为未知参数。证明 α 和 β 的最大似然估计量为

$$\mathrm{argmin}_{\alpha,\beta} \sum_{i=1}^{n} |y_i - (\alpha + \beta x_i)| \text{。}$$

第二题

给定 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，令 $\hat{\alpha}$ 和 $\hat{\beta}$ 为最小二乘估计量，$\hat{y}_i = \hat{\beta} x_i + \hat{\alpha}$ 为 x_i 的预测值，$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$，$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$。证明

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\overline{y} - \hat{y}_i)^2 \text{。}$$

提示：利用正规方程，并证明

$$\hat{y}_i = \overline{y} + \hat{\beta}(x_i - \overline{x}) \text{。}$$

第三题

给定 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$，其中 $y_i = \alpha + \beta x_i + \epsilon_i$，$\epsilon_i$ 相互独立且 $\epsilon_i \sim N(0, \sigma^2)$。沿用第二题中的记号，并令 $s^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2$，$s_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$。

(1) 令

$$q_1 = [1/\sqrt{n}, 1/\sqrt{n}, \ldots, 1/\sqrt{n}]^T \in \mathbb{R}^n,$$

$$q_2 = \left[\frac{x_1 - \overline{x}}{\sqrt{s_{xx}}}, \frac{x_2 - \overline{x}}{\sqrt{s_{xx}}}, \ldots, \frac{x_n - \overline{x}}{\sqrt{s_{xx}}}\right]^T \in \mathbb{R}^n \text{。}$$

证明存在 $q_3, q_4, \ldots, q_n \in \mathbb{R}^n$，使得 $q_1, q_2, q_3, q_4, \ldots, q_n$ 为 $\mathbb{R}^n$ 中的一组标准正交基。

(2) 将 y 视作 $\mathbb{R}^n$ 中的向量。对于 $1 \le i \le n$，令 $z_i = q_i^T y$，也即 $z = Qy \in \mathbb{R}^n$，其中 $Q \in \mathbb{R}^{n \times n}$ 的第 i 行为 $q_i \in \mathbb{R}^n$。给出 n 维随机变量 z 服从的分布。提示：计算随机向量 y 的数学期望，并验证其与 $q_3, q_4, \ldots, q_n$ 的正交性。

(3) 证明 $z_1 = \sqrt{n}\overline{y}$，$z_2 = \sqrt{s_{xx}}\hat{\beta}$。

(4) 利用第二题中提示和结论，证明 $\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2=z_2^2$ 及 $(n-2)s^2=\sum(y_i-\hat{y}_i)^2=\sum_{i=3}^{n}z_i^2$。

(5) 给出 $(n-2)s^2/\sigma^2$ 服从的分布，并证明 s^2 与 $\hat{\alpha}$ 和 $\hat{\beta}$ 均相互独立。

(6) 当 $\beta=0$，给出统计量 $t=\frac{\hat{\beta}}{\sqrt{s^2}/\sqrt{s_{xx}}}$ 服从的分布。

(7) 若 σ^2 未知，考虑假设检验问题，原假设 $H_0:\beta=0$，备择假设 $H_1:\beta\neq 0$。拒绝域为

$$W=\{((x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n))\mid |t|\geq c\},$$

其中 c 为待定常数。若显著性水平为 α，给出 c 的取值。