数据科学的数学基础
第 2 次理论作业

陈万祺　3220102895

2025 年 10 月 1 日

Problem 2.2

Let X be a random variable that you know is in the range $[-1, 2]$ and you know has an expected value of $E[X] = 0$. Use the Markov inequality to upper-bound $Pr[X > 1.5]$. (Hint: you will need to use a change of variables.)

解：

由于马尔可夫不等式要求随机变量非负，令 $Y = X + 1$，则：

$$Y \in [0, 3], \quad E[Y] = E[X + 1] = E[X] + 1 = 1$$

注意到：

$$Pr[X > 1.5] = Pr[Y > 2.5]$$

对非负随机变量 Y 应用马尔可夫不等式：

$$Pr[Y \geq 2.5] \leq \frac{1}{2.5} = \frac{2}{5}$$

由于 $Pr[Y > 2.5] \leq Pr[Y \geq 2.5]$，因此：

$$Pr[X > 1.5] = Pr[Y > 2.5] \leq \frac{2}{5}$$

Problem 2.4

2.4 Consider a pdf f so that a random variable $X \sim f$ has expected value $E[X] = 5$ and variance $Var[X] = 100$. Now consider $n = 16$ iid random variables $X_1, X_2, \ldots, X_{16}$ drawn from f. Let $\bar{X} = \frac{1}{16}\sum_{i=1}^{16} X_i$.

1. What is $E[\bar{X}]$?
2. What is $Var[\bar{X}]$? Assume we know that X is never smaller than 0 and never larger than 20.
3. Use the Markov inequality to upper-bound $Pr[\bar{X} > 8]$.
4. Use the Chebyshev inequality to upper-bound $Pr[\bar{X} > 8]$.
5. Use the Chernoff-Hoeffding inequality to upper-bound $Pr[\bar{X} > 8]$.
6. If we increase n to 100, how will the above three bounds be affected?

解：

1. 由于 $X_1, X_2, \ldots, X_{16}$ 是独立同分布随机变量，且 $E[X] = 5$，则：

$$E[\bar{X}] = E\left[\frac{1}{16}\sum_{i=1}^{16} X_i\right] = \frac{1}{16}\sum_{i=1}^{16} E[X_i] = \frac{1}{16} \cdot 16 \cdot 5 = 5.$$

因此，$E[\bar{X}] = 5$.

2. 由于方差性质，且 $Var[X] = 100$，则：

$$Var[\bar{X}] = Var\left[\frac{1}{16}\sum_{i=1}^{16} X_i\right] = \frac{1}{16^2}\sum_{i=1}^{16} Var[X_i] = \frac{1}{256} \cdot 16 \cdot 100 = \frac{1600}{256} = 6.25.$$

因此，$Var[\bar{X}] = 6.25$.

3. 使用马尔可夫不等式。由于 $X \in [0, 20]$，故 $\bar{X} \geq 0$。马尔可夫不等式给出：

$$Pr[\bar{X} > 8] \leq Pr[\bar{X} \geq 8] \leq \frac{5}{8}.$$

4. 使用切比雪夫不等式。切比雪夫不等式给出：

$$Pr[|\bar{X} - E[\bar{X}]| \geq k] \leq \frac{Var[\bar{X}]}{k^2}.$$

这里 $E[\bar{X}] = 5$，欲求 $Pr[\bar{X} > 8] = Pr[\bar{X} - 5 > 3]$。注意到：

$$Pr[\bar{X} > 8] \leq Pr[|\bar{X} - 5| \geq 3] \leq \frac{Var[\bar{X}]}{3^2} = \frac{6.25}{9} = \frac{25}{36}.$$

5. 使用 Chernoff-Hoeffding 不等式。由于 $X_i \in [0, 20]$，则 $\bar{X}$ 是 $n = 16$ 个独立随机变量的均值，且 $E[\bar{X}] = 5$。Chernoff-Hoeffding 不等式给出：

$$Pr[\bar{X} - E[\bar{X}] \geq t] \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right),$$

其中 $a=0, b=20$，故 $b-a=20$。取 $t=8-5=3$，得：

$$Pr[\bar{X}>8]=Pr[\bar{X}-5\geq 3]\leq \exp\left(-\frac{2\cdot 16\cdot 3^2}{20^2}\right)=\exp\left(-\frac{288}{400}\right)=\exp(-0.72).$$

6. 当 n 增加到 100 时：
由于 $E[\bar{X}]=5$ 不变，故马尔可夫界仍为 $\frac{5}{8}$。
由于 $Var[\bar{X}]=\frac{100}{100}=1$，故切比雪夫界变为 $\frac{1}{3^2}=\frac{1}{9}$，比原来的更小。
由于指数变为 $\exp\left(-\frac{2\cdot 100\cdot 3^2}{20^2}\right)=\exp(-4.5)$，Chernoff-Hoeffding 界比原来更小。
因此，马尔可夫界不变，切比雪夫界和 Chernoff-Hoeffding 界均减小。

Problem 2.5

Consider a (parked) self-driving car that returns n iid estimates to the distance of a tree. We will model these n estimates as a set of n scalar random variables $X_1, X_2, \ldots, X_n$ taken iid from an unknown pdf f, which we assume models the true distance plus unbiased noise (the sensor can take many iid estimates in rapid-fire fashion). The sensor is programmed to only return values between 0 and 20 feet, and that the variance of the sensing noise is 64 feet squared. Let $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$. We want to understand as a function of n how close $\bar{X}$ is to μ, which is the true distance to the tree.

1. Use Chebyshev's inequality to determine a value n so that $Pr[|\bar{X}-\mu|\geq 1]\leq 0.5$.
2. Use Chebyshev's inequality to determine a value n so that $Pr[|\bar{X}-\mu|\geq 0.1]\leq 0.1$.
3. Use the Chernoff-Hoeffding bound to determine a value n so that $Pr[|\bar{X}-\mu|\geq 1]\leq 0.5$.
4. Use the Chernoff-Hoeffding bound to determine a value n so that $Pr[|\bar{X}-\mu|\geq 0.1]\leq 0.1$.

解：

1. 由于 X_i 独立同分布，且 $E[X_i]=\mu$，$Var[X_i]=64$，则

$$Var[\bar{X}]=\frac{64}{n}.$$

切比雪夫不等式给出：

$$Pr[|\bar{X}-\mu|\geq\varepsilon]\leq\frac{Var[\bar{X}]}{\varepsilon^2}=\frac{64}{n\varepsilon^2}.$$

令 $\varepsilon=1$，要求 $Pr[|\bar{X}-\mu|\geq 1]\leq 0.5$，即

$$\frac{64}{n}\leq 0.5 \quad\Rightarrow\quad n\geq 128.$$

因此，取 $n=128$。

2. 令 $\varepsilon = 0.1$，要求 $Pr[|\bar{X} - \mu| \geq 0.1] \leq 0.1$，即

$$\frac{64}{n \cdot (0.1)^2} \leq 0.1 \quad \Rightarrow \quad \frac{6400}{n} \leq 0.1 \quad \Rightarrow \quad n \geq 64000.$$

因此，取 $n = 64000$。

3. 由于 $X_i \in [0, 20]$，则 $b - a = 20$。Chernoff–Hoeffding 不等式给出：

$$Pr[|\bar{X} - \mu| \geq \varepsilon] \leq 2\exp\left(-\frac{2n\varepsilon^2}{(b-a)^2}\right) = 2\exp\left(-\frac{2n\varepsilon^2}{400}\right) = 2\exp\left(-\frac{n\varepsilon^2}{200}\right).$$

令 $\varepsilon = 1$，要求 $Pr[|\bar{X} - \mu| \geq 1] \leq 0.5$，即

$$2\exp\left(-\frac{n}{200}\right) \leq 0.5 \quad \Rightarrow \quad \exp\left(-\frac{n}{200}\right) \leq 0.25 \quad \Rightarrow \quad -\frac{n}{200} \leq \ln(0.25) = -\ln(4).$$

因此，

$$n \geq 200\ln(4) \approx 200 \times 1.3863 = 277.26.$$

取整数 $n = 278$。

4. 令 $\varepsilon = 0.1$，要求 $Pr[|\bar{X} - \mu| \geq 0.1] \leq 0.1$，即

$$2\exp\left(-\frac{n \cdot (0.1)^2}{200}\right) \leq 0.1 \quad \Rightarrow \quad 2\exp\left(-\frac{n}{20000}\right) \leq 0.1 \quad \Rightarrow \quad \exp\left(-\frac{n}{20000}\right) \leq 0.05.$$

因此，

$$-\frac{n}{20000} \leq \ln(0.05) = -\ln(20) \quad \Rightarrow \quad n \geq 20000\ln(20) \approx 20000 \times 2.9957 = 59914.$$

取整数 $n = 59915$。

Problem 2.6

Consider two random variables C and T describing how many coffees and teas I will buy in the coming week; clearly neither can be smaller than 0. Based on personal experience, I know the following summary statistics about my coffee and tea buying habits: $\mathbb{E}[C] = 3$ and $\mathrm{Var}[C] = 1$, $\mathbb{E}[T] = 2$ and $\mathrm{Var}[T] = 5$

1.Use Markov's inequality to upper-bound the probability that I buy 4 or more coffees, and the same for teas: $\Pr[C \geq 4]$ and $\Pr[T \geq 4]$

2. Use Chebyshev's inequality to upper-bound the probability that I buy 4 or more coffees, and the same for teas: $\Pr[C \geq 4]$ and $\Pr[T \geq 4]$

解：

1. 使用马尔可夫不等式：马尔可夫不等式指出，对于非负随机变量 X 和常数 $a > 0$，有

$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}.$$

对于咖啡 C，有 $\mathbb{E}[C] = 3$，取 $a = 4$，得

$$\Pr[C \geq 4] \leq \frac{3}{4}.$$

对于茶 T，有 $\mathbb{E}[T] = 2$，取 $a = 4$，得

$$\Pr[T \geq 4] \leq \frac{2}{4} = \frac{1}{2}.$$

因此，马尔可夫上界为 $\Pr[C \geq 4] \leq \frac{3}{4}$ 和 $\Pr[T \geq 4] \leq \frac{1}{2}$.

2. 使用切比雪夫不等式：切比雪夫不等式指出，对于随机变量 X 和常数 $k > 0$，有

$$\Pr[|X - \mathbb{E}[X]| \geq k] \leq \frac{\mathrm{Var}[X]}{k^2}.$$

对于咖啡 C，有 $\mathbb{E}[C] = 3$，$\mathrm{Var}[C] = 1$。注意到

$$\Pr[C \geq 4] = \Pr[C - 3 \geq 1] \leq \Pr[|C - 3| \geq 1] \leq \frac{1}{1^2} = 1.$$

对于茶 T，有 $\mathbb{E}[T] = 2$，$\mathrm{Var}[T] = 5$。注意到

$$\Pr[T \geq 4] = \Pr[T - 2 \geq 2] \leq \Pr[|T - 2| \geq 2] \leq \frac{5}{2^2} = \frac{5}{4}.$$

因此，切比雪夫上界为 $\Pr[C \geq 4] \leq 1$ 和 $\Pr[T \geq 4] \leq \frac{5}{4}$.

EX_Problem 1

Let X be a discrete random variable, with $X \in [-1, 1]$ and $\mathbb{E}[X] = 0$. If $t \in [0, 1]$, prove that

$$\mathbb{E}[e^{tX}] \leq 1 + t^2 \mathrm{Var}[X] \leq e^{t^2 \mathrm{Var}[X]}.$$

证明： 由于 $X \in [-1, 1]$ 且 $\mathbb{E}[X] = 0$，考虑函数 e^{tx} 在区间 $[-1, 1]$ 上的性质。

对于任意 $x \in [-1, 1]$，函数 e^{tx} 是凸函数，因此它在该区间上位于连接点 $(-1, e^{-t})$ 和 $(1, e^{t})$ 的线段下方，即：

$$e^{tx} \leq \frac{1 - x}{2} e^{-t} + \frac{1 + x}{2} e^{t} \quad \text{对所有} x \in [-1, 1]$$

对上述不等式两边取期望（关于 X），并利用 $\mathbb{E}[X] = 0$：

$$\mathbb{E}[e^{tX}] \leq \mathbb{E}\left[\frac{e^{-t} + e^{t}}{2} + X \cdot \frac{e^{t} - e^{-t}}{2}\right] = \frac{e^{-t} + e^{t}}{2} + \frac{e^{t} - e^{-t}}{2} \mathbb{E}[X] = \frac{e^{-t} + e^{t}}{2}$$

已知 $\cosh(t) = \frac{e^t + e^{-t}}{2}$，所以：

$$\mathbb{E}[e^{tX}] \leq \cosh(t)$$

利用不等式 $\cosh(t) \leq 1 + \frac{t^2}{2}$（当 $t \in [0,1]$ 时成立），以及 $\mathrm{Var}[X] = \mathbb{E}[X^2] \leq 1$（因为 $X^2 \leq 1$），可得：

$$\mathbb{E}[e^{tX}] \leq 1 + \frac{t^2}{2} \leq 1 + t^2 \mathrm{Var}[X]$$

最后，利用不等式 $1 + y \leq e^y$（对所有实数 y 成立），令 $y = t^2\mathrm{Var}[X]$，得：

$$1 + t^2\mathrm{Var}[X] \leq e^{t^2 \mathrm{Var}[X]}$$

综上，不等式链得证。

EX_Problem 2

Let $X_1, \ldots, X_n$ be discrete, identical and independent random variables with $\mathbb{E}[X_i] = 0$, $\mathrm{Var}[X_i] = \sigma^2$, and $X_i \in [-1, 1]$. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Prove that

$$\Pr[|\bar{X}| \geq \lambda\sigma] \leq 2e^{-\lambda^2 n/4}$$

for any $0 \leq \lambda \leq 2\sigma$.

证明： 令 $s = \frac{\lambda}{2\sigma}$，由于 $0 \leq \lambda \leq 2\sigma$，有 $s \in [0,1]$，满足问题 1 的条件。

对于任意 $s > 0$，根据马尔可夫不等式：

$$\Pr[\bar{X} \geq \lambda\sigma] = \Pr\left[e^{sn\bar{X}} \geq e^{sn\lambda\sigma}\right] \leq \frac{\mathbb{E}[e^{sn\bar{X}}]}{e^{sn\lambda\sigma}}$$

由于 X_i 独立同分布，且 $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$，有：

$$\mathbb{E}[e^{sn\bar{X}}] = \mathbb{E}[e^{s\sum_{i=1}^n X_i}] = \prod_{i=1}^n \mathbb{E}[e^{sX_i}] = (\mathbb{E}[e^{sX_1}])^n$$

应用问题 1 的结论，有 $\mathbb{E}[e^{sX_1}] \leq e^{s^2\mathrm{Var}[X_1]} = e^{s^2\sigma^2}$，因此：

$$\Pr[\bar{X} \geq \lambda\sigma] \leq \frac{(e^{s^2\sigma^2})^n}{e^{sn\lambda\sigma}} = e^{ns^2\sigma^2 - sn\lambda\sigma}$$

令 $s = \frac{\lambda}{2\sigma}$，代入上式：

$$\Pr[\bar{X} \geq \lambda\sigma] \leq e^{n\cdot(\frac{\lambda}{2\sigma})^2\sigma^2 - \frac{\lambda}{2\sigma}\cdot n\lambda\sigma} = e^{n\cdot\frac{\lambda^2}{4} - n\cdot\frac{\lambda^2}{2}} = e^{-\frac{\lambda^2 n}{4}}$$

同理，对于 $\Pr[\bar{X} \leq -\lambda\sigma]$，可以得到相同的上界。

因此，结合两侧：

$$\Pr[|\bar{X}| \geq \lambda\sigma] = \Pr[\bar{X} \geq \lambda\sigma] + \Pr[\bar{X} \leq -\lambda\sigma] \leq 2e^{-\frac{\lambda^2 n}{4}}$$