数据科学的数学基础
第 3 次理论作业

陈万祺　3220102895

2025 年 10 月 18 日

Problem 2.7

The average score on a test is 82 with a standard deviation of 4 percentage points. All tests have scores between 0 and 100.

1. Using Chebyshev's inequality, what percentage of the tests have a grade of at least 70 and at most 94?
2. Using Markov's inequality, what is the highest percentage of tests which could have a score less than 60?

解：

1. 均值 $\mu = 82$，标准差 $\sigma = 4$。区间 $[70, 94]$ 对称于均值，偏差为 $|82 - 70| = 12$，因此 $k\sigma = 12$，$k = 12/4 = 3$。由切比雪夫不等式

$$P(|X - \mu| \le k\sigma) \ge 1 - \frac{1}{k^2}$$

所以

$$P(70 \le X \le 94) \ge 1 - \frac{1}{3^2} = \frac{8}{9} \approx 88.89\%$$

因此，至少约 88.89% 的测试分数在 70 和 94 之间。

2. 定义 $Y = 100 - X$，则 Y 非负，且 $E[Y] = 100 - 82 = 18$。$P(X < 60) = P(Y > 40)$。由马尔可夫不等式

$$P(Y \ge a) \le \frac{E[Y]}{a}$$

因此

$$P(Y > 40) \le \frac{18}{40} = 0.45$$

所以分数小于 60 的最高百分比为 45%。

Problem 2.8

Consider a random variable X with expected values $E[X] = 7$ and variance $\mathrm{Var}[X] = 2$. We would like to upper-bound the probability $\Pr[X < 5]$.

1. Which bound can and cannot be used with what we know about X (Markov, Chebyshev, or Chernoff-Hoeffding), and why?
2. Using that bound, calculate an upper bound for $\Pr[X < 5]$.
3. Describe a probability distribution for X where the other two bounds are definitely not applicable.

解： 1. 边界使用情况：马尔可夫不等式：无法使用，因为马尔可夫不等式要求随机变量非负，而 X 可能取负值，且无法直接用于 $\Pr[X < 5]$ 的上界。

切比雪夫不等式：可以使用，因为我们知道方差，且切比雪夫不等式适用于任何具有有限方差的随机变量。

切尔诺夫-霍夫丁不等式：无法使用，因为该不等式通常适用于有界独立随机变量的和或平均，而这里只有一个随机变量，且不知道其分布或有界性。

2. 由切比雪夫不等式：

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

这里 $\mu = 7$，且

$$\Pr[X < 5] = \Pr[X - 7 < -2] \leq \Pr[|X - 7| \geq 2]$$

因为事件 $X - 7 < -2$ 是 $|X - 7| \geq 2$ 的子集。因此

$$\Pr[X < 5] \leq \Pr[|X - 7| \geq 2] \leq \frac{2}{2^2} = 0.5$$

因此，$\Pr[X < 5]$ 的上界为 0.5。

3. 描述一个概率分布，其中其他两个边界肯定不适用：考虑 X 服从正态分布 $N(7, 2)$，即均值 7，方差 2。马尔可夫不等式不适用，因为 X 可能取负值（例如，正态分布有负值概率），且马尔可夫不等式要求非负。切尔诺夫-霍夫丁不等式不适用，因为该不等式要求随机变量有界，而正态分布无界。

Problem 2.9

Consider n iid random variables $X_1, X_2, \ldots, X_n$ with expected value $E[X_i] = 20$ and variance $\mathrm{Var}[X_i] = 2$. Assume we also know that each X_i must satisfy $15 \leq X_i \leq 22$. We now want to analyze the random variable of their average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Assume first that $n = 20$ (the number of random variables).

1. Use the Chebyshev inequality to upper-bound $\Pr[\bar{X} > 21]$.
2. Use the Chernoff-Hoeffding inequality to upper-bound $\Pr[\bar{X} > 21]$. Now assume first that $n = 200$ (the number of random variables).
3. Use the Chebyshev inequality to upper-bound $\Pr[\bar{X} > 21]$.
4. Use the Chernoff-Hoeffding inequality to upper-bound $\Pr[\bar{X} > 21]$.

解： 1. 对于平均 $\bar{X}$，有 $E[\bar{X}] = 20$，$\mathrm{Var}[\bar{X}] = \frac{\mathrm{Var}[X_i]}{n} = \frac{2}{n}$。由于 $X_i \in [15, 22]$，所以 $\bar{X} \in [15, 22]$。令 $\mu = 20$，$t = 1$（因为 $\bar{X} > 21$ 等价于 $\bar{X} - \mu > 1$）。

当 $n = 20$：1. 切比雪夫不等式：

$$\Pr[\bar{X} > 21] = \Pr[\bar{X} - 20 > 1] \leq \Pr[|\bar{X} - 20| \geq 1] \leq \frac{\mathrm{Var}[\bar{X}]}{1^2} = 0.1$$

所以上界为 0.1。

2. 切尔诺夫-霍夫丁不等式：由于 $X_i \in [15, 22]$，范围 $b - a = 22 - 15 = 7$。切尔诺夫-霍夫丁给出 $\Pr[\bar{X} - \mu \geq t] \leq \exp\left(-\frac{2nt^2}{(b-a)^2}\right)$。所以 $\Pr[\bar{X} > 21] \leq \exp\left(-\frac{2\times 20\times 1^2}{7^2}\right) = \exp\left(-\frac{40}{49}\right) \approx \exp(-0.8163) \approx 0.441$。所以上界约为 0.441。

当 $n = 200$：

3. 切比雪夫不等式：

$$\Pr[\bar{X} > 21] \leq \frac{\mathrm{Var}[\bar{X}]}{1^2} = 0.01$$

所以上界为 0.01。

4. 切尔诺夫-霍夫丁不等式：$\Pr[\bar{X} > 21] \leq \exp\left(-\frac{2\times 200\times 1^2}{7^2}\right) = \exp\left(-\frac{400}{49}\right) \approx \exp(-8.163) \approx 0.00028$。所以上界约为 0.00028。

EXProblem 1

Given the error tolerance $\varepsilon = 0.01$, at least how many samples do we need to approximate all quantiles of a distribution so that the probability of failure $\delta = 0.05$?

解： 为了使用经验分布函数一致逼近真实分布函数的所有分位数，我们使用 Dvoretzky-Kiefer-Wolfowitz (DKW) 不等式：

$$P\left(\sup_x |F_n(x) - F(x)| > \varepsilon\right) \leq 2\exp(-2n\varepsilon^2)$$

其中 $F_n(x)$ 是经验分布函数，$F(x)$ 是真实分布函数，n 是样本数量。
要确保失败概率不超过 $\delta = 0.05$，我们需要：

$$2\exp(-2n\varepsilon^2) \leq \delta$$

代入 $\varepsilon = 0.01$ 和 $\delta = 0.05$：

$$2\exp(-2n(0.01)^2) \leq 0.05$$

$$-0.0002n \leq \ln(0.025)$$

由于 $\ln(0.025) \approx -3.688879$，所以：

$$n \geq \frac{3.688879}{0.0002} = 18444.395$$

因此，至少需要 18445 个样本才能以 95% 的置信水平确保所有分位数的估计误差不超过 0.01。