8th problem set for Probability and Statistics — April 20/22

x	-4	-3	-2	-1	0	1	2	3	4
$\Phi(x)$	0.00003	0.00135	0.02275	0.15866	0.500000	0.84135	0.97725	0.99865	0.99997

Further values see https://en.wikipedia.org/wiki/Standard_normal_table – section Cumulative.

1. Let $Z \sim N(0,1)$. Use the Φ function table to calculate:
(a) $P(|Z| \leq 1)$
(b) $P(|Z| \leq 2)$
(c) $P(|Z| \leq 3)$

2. Let X and Y be two random variables with $\mathbb{E}(X) = 1$, $\sigma_X = 2$, and $\mathbb{E}(Y) = 2$, $\sigma_Y = 1$. Find the maximum possible value for $\mathbb{E}(XY)$. Also, express Y as a function of X for which this maximum is achieved. (Hint: The best bound is not obtained by applying the Cauchy-Schwarz inequality directly, but by considering the correlation coefficient.)

Applications of Inequalities and Central Limit Theorem (CLT)

- **Markov's inequality:** $P(X \geq a\mathbb{E}(X)) \leq \frac{1}{a}$ for $X \geq 0$.
- **Chebyshev's inequality:** $P(|X - \mathbb{E}(X)| \geq t\sigma_X) \leq \frac{1}{t^2}$.
- **Central Limit Theorem:** Let $X_1, X_2, \ldots$ a sequence of i.i.d. $L^2(\Omega, F, P)$ random variables with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2 > 0$. Then for the sequence $(S_n^*)_{n=1}^{\infty}$, where

$$S_n^* = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu}{\sigma}$$

we have

$$S_n^* \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

3. We roll a die, and if the result is 1 or 2, we get a single point. Let X be the number of points obtained after n (independent) rolls. Estimate the probability that $X \geq n/2$:
(a) Using Markov's inequality.
(b) Using Chebyshev's inequality.
(c) For a specific n, how can this value be determined exactly? For comparison, for $n = 100$ the actual probability is $1 - F_X(49)$, where $X \sim Bin(100, 1/3)$, which is less than 0.4

4. We know that the average number of points on a test was 40 (out of 100). Estimate the proportion of students with at least 80 points. Improve the estimate if you know that the standard deviation of the number of points is 10.

5. Estimate $\binom{100}{40}$ using the central limit theorem. (Hint: use CLT to estimate $P(39.5 < X < 40.5)$ for a suitable random variable X. On the other hand, for $P(X = 40)$, we have the formula $\binom{100}{40}/2^{100}$ from the binomial distribution.)

6. Let $S = \sum_{k=0}^{30} \binom{100}{k}$. Also, let $X = \sum_{i=1}^{100} X_i$, where X_i is 0 or 1, both with probability 1/2 and the variables $X_1, \ldots, X_n$ are independent. Thus, $X \sim Bin(100, 1/2)$.
(a) Express S using the cumulative distribution function F_X.
(b) Use CLT to estimate this probability.

7. We want to estimate whether our coin (and the way we flip it) is fair. If we get more than 55 heads out of a hundred flips, we'll say it's not fair. What is the probability of making a mistake? That is, if we have a fair coin, what is the probability that we get more than 55 heads out of a hundred flips?

More practice problems

8. You're throwing a party for 100 guests and wondering how many sandwiches to order. You know from experience that the number of sandwiches eaten by a random guest follows a Poisson distribution with a mean of 3. Approximately how many sandwiches do you need to order so that with probability 0.95 no guest will go hungry? (Hint: Use an appropriate limit theorem.)

9. A biased coin, which lands heads with probability $1/10$ each time it is flipped, is flipped 200 times consecutively. Give an upper bound on the probability that it lands heads at least 120 times using

(a) Markov's inequality.

(b) Chebyshev's inequality.

Bonus problems

10. Suppose that each of $m \geq 1$ pigeons independently and at random enter one of $n \geq 1$ pigeonholes. If $m \geq 1.2\sqrt{n} + 1$, then show that the probability that two pigeons go into the same pigeonhole is greater than $1/2$.

11. A statistician wants to estimate the average height h (in meters) of people in a population using n independent samples $X_1, \ldots, X_n$, randomly selected from all possible people. For estimation, he uses the sample mean $\bar{X}_n = (X_1 + \cdots + X_n)/n$. He estimates that the standard deviation of a single measurement is at most 1 meter.

(a) What value of n guarantees that the standard deviation of $\bar{X}_n$ is at most 1 cm?

(b) For what n does Chebyshev's inequality guarantee that $\bar{X}_n$ differs from h by at most 5 cm with at least 99% probability?

(c) The statistician notices that all measured people have heights in the interval $(1.4, 2.1)$. How should he adjust the estimate of the standard deviation? How will the answers to the previous questions change?