Handout 1: RVs - Inequalities - Limit theorems

Due date : February 3^{rd}

Brief recap

1. PMFs are usually more intuitive to retrieve for discrete rvs, whereas CDFs are more intuitive for continuous rvs.

2. A PMF is valid iif, $p(e) \geq 0$ for any event, and $\sum_{i \in E} = 1$, where E is the sample space.

3. A PDFs f is valid iif : $f \geq 0$ for any event, f is piecewise continuous, and $\int_{-\infty}^{\infty} f(x)dx = 1$.

4. A Bern(p) r.v. is the indicator of success in a Bernoulli trial with probability of success p.

5. A Bin(n, p) r.v. is the number of successes in n independent Bernoulli trials, all with the same probability p of success.

6. Cauchy-Schwarz and Jensen inequalities give bounds on expectations. Markov and Chebyshev inequalities give bounds on tail probabilities.

7. The LLN and the CLT describe the behavior of the sample mean $\bar{X}_n$ of i.i.d. rvs $X_1, X_2, \ldots$ with mean μ and variance σ^2. The LLN says that $P(\lim_{n\to\infty} \bar{X}_n = \mu$ with probability 1. The CLT says that

$$\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \to \mathcal{N}(0,1)$$

in distribution, which can be recast as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).$$

Exercise 1. Consider a sequence of independent Bernoulli trials, each with the same success probability $p \in]0,1[$, with trials performed until a success occurs. Let X be the number of failures before the first successful trial.

1. Determine $P(X = k)$ and show that it is a valid PMF.

2. Calculate $E[X]$ and $V[X]$.

This is the geometric distribution $X \sim \text{Geom}(p)$.

Exercise 2.

1. Show that $p(n) = (1/2)^{n+1}$ for $n = 0, 1, 2, \ldots$ is a valid PMF for a discrete r.v.

2. Find the CDF of a random variable with the PMF in 1.

Exercise 3.

1. Benford's law states that in a very large variety of real-life data sets, the first digit approximately follows a particular distribution with about a 30% chance of a 1, an 18% chance of a 2, and in general

$$P(D = j) = \log_{10}\left(\frac{j+1}{j}\right), \quad 1 \leq j \leq 9.$$

where D is the first digit of a randomly chosen element. Check that this is a valid PMF.

Exercise 4. There are n eggs, each of which hatches a chick with probability p (independently). Each of these chicks survives with probability r, independently.

1. What is the distribution of the number of chicks that hatch?

2. What is the distribution of the number of chicks that survive?

Give the PMFs; also give the names of the distributions and their parameters, if applicable.

Exercise 5. Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, independent of X.

1. Show that $X + Y$ is Binomial and give its distribution.

2. Is $X - Y$ Binomial? If yes, give its distribution. If not, why?

Exercise 6. There are two coins, one with probability p_1 of Heads and the other with probability p_2 of Heads. One of the coins is randomly chosen (with equal probabilities for the two coins). It is then flipped $n \geq 2$ times. Let X be the number of times it lands Heads.

1. Find the PMF of X.
2. What is the distribution of X if $p_1 = p_2$?

Exercise 7. Student A flips a fair coin n times and student B flips another fair coin $n+1$ times, resulting in independent random variables $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim \text{Bin}(n+1, \frac{1}{2})$.

1. Show that $P(X < Y) = 1 - P(n - X < n + 1 - Y)$.
2. Compute $P(X < Y)$.

Exercise 8. Prove that $E[X]^2 \leq E[X^2]$.

Exercise 9. For a non-negative rv X and $a \in \mathbb{R}_+^*$

1. Show that a $\mathbb{I}_{|X| \geq a} \leq X$ ($\mathbb{I}$ is the indicator rv).
2. Prove the Markov inequality using 1.

Exercise 10. Let X be the number of purchases you will make on the website of a company in a specified time period. Let the PMF of X be

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots$$

This distribution is called the Poisson distribution with parameter λ, and it will be studied extensively in later chapters.

1. Find $P(X \geq 1)$ and $P(X \geq 2)$.
2. Find the conditional PMF of X given $X \geq 1$. (This conditional distribution is called a truncated Poisson distribution.)

Exercise 11. In a national survey, a random sample of people are chosen and asked whether they support a certain policy. Assume that everyone in the population is equally likely to be surveyed at each step, and that the sampling is with replacement. Let n be the sample size, and let $\hat{p}$ and p be the proportion of people who support the policy in the sample and in the entire population, respectively. Show that

$$\forall c > 0, \quad P(|\hat{p} - p| > c) \leq \frac{1}{4nc^2}. \qquad (1)$$

Exercise 12. For i.i.d. r.v.s $X_1, \ldots, X_n$ with mean μ and variance σ^2, give a value of n that will ensure that there is at least a 99% chance that the sample mean will be within 2 standard deviations of the true mean μ.

Exercise 13. Let $X_1, X_2, \ldots$ be i.i.d. positive random variables with mean 2. Let $Y_1, Y_2, \ldots$ be i.i.d. positive random variables with mean 3.

1. Show that
$$\frac{X_1 + X_2 + \ldots + X_n}{Y_1 + Y_2 + \ldots + Y_n} \underset{n \to \infty}{\to} \frac{2}{3}$$
with probability 1. Does it matter whether the X_i are independent of the Y_j ?

Exercise 14. Let $Y_1, Y_2, \ldots, U_{100}$ be i.i.d. r.v.s with distribution Unif$(0, 1)$ and $X = Y_1 + Y_2 + \ldots + U_{100}$.

1. Which important distribution is the distribution of X very close to? Specify what the parameters are, and state which theorem justifies your choice.
2. Give a simple but accurate approximation for $P(X > 17)$. Justify briefly.

Exercise 15. Let $Y \sim \text{Bin}(n, p)$, with $n \in \mathbb{N}^*$ and $p \in [0, 1]$. Show that

$$Y \underset{n \to \infty}{\sim} \mathcal{N}(np, np(1-p)).$$

Exercise 16.

1. Let $Y = e^X$, with $X \sim \text{Expo}(3)$. Find the mean and variance of Y.
2. For $Y_1, \ldots, Y_n$ i.i.d. with the same distribution as Y from the previous question, what is the approximate distribution of the sample mean $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$ when n is large?

References: Introduction to probability (Blitzstein and Huang), Stochastic Processes (Gallager).