STAT201A: Introduction to Probability at an Advanced Level (Fall 2024)
UC Berkeley

Problem Set 2

Due: 10:00pm, Friday, September 27, 2024 (via Gradescope)

1. **(Binomial tail bounds)** Let S_n have the Binomial(n, p_i) distribution of the number of successes in n independent Bernoulli(p_i) trials. Use a suitable computational environment to evaluate the right tail probabilities

$$\mathbb{P}\left(\frac{S_n}{n} \geq p_i + \epsilon\right)$$

for $n = 100$ and $p_i = i/10$ for $i = 1, 2, \ldots, 9$, and $\epsilon = 1/10$, together with various approximations and upper bounds as indicated. In each case,

- give an exact mathematical formula for the function of i you are computing;
- indicate suitable code for evaluating the formula in your preferred environment and **attach the code at the end of the homework**;
- give the numerical values correct to two significant decimal place.

(a) The exact probabilities.
For a binomial random variable $S_n \sim$ Binomial(n, p_i), the exact probability is

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) = \sum_{k=[n(p_i+\epsilon)]}^{n} \binom{n}{k} p_i^k \left(1 - p_i\right)^{n-k}$$

(b) Markov's upper bounds for these probabilities.
Markov's inequality provides a simple upper bound on the tail probability, which is

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \frac{\mathbb{E}\left[S_n\right]}{n\left(p_i + \epsilon\right)} = \frac{np_i}{n\left(p_i + \epsilon\right)} = \frac{p_i}{p_i + \epsilon}$$

(c) Chebychev's upper bounds for these probabilities (which can be halved for $i = 5$ only: explain why).
Chebyshev's inequality uses the variance of S_n, which is $\text{Var}\left(S_n\right) = np_i\left(1 - p_i\right)$. It provides an upper bound on the tail probability as

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \frac{\text{Var}\left(S_n\right)}{n^2\epsilon^2} = \frac{p_i\left(1 - p_i\right)}{n\epsilon^2}$$

When $p_i = 0.5$, the probability distribution for S_n is symmetric around its expectation $0.5n$

$$\begin{aligned}\mathbb{P}\left(\left|\frac{S_n}{n} - 0.5\right| \geq \epsilon\right) &= \mathbb{P}\left(S_n \geq n(0.5 + \epsilon)\right) + \mathbb{P}\left(S_n \leq n(0.5 - \epsilon)\right) \\ &= 2\mathbb{P}\left(S_n \geq n(0.5 + \epsilon)\right).\end{aligned}$$

When $p_i \neq 0.5$, the distribution is not symmetric and that's why the bound can be halved for $i = 5$ only.

(d) Hoeffding's upper bounds.

Hoeffding's inequality provides an upper bound for the sum of independent bounded random variables, such as a binomial variable S_n. The upper bound is

$$\mathbb{P}\left(\frac{S_n}{n} \geq p_i + \epsilon\right) \leq \exp\left(-2n\epsilon^2\right)$$

(e) Chernoff's upper bounds.

Here we use KL divergence form for Chernoff's bound

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \exp\left(-nD\left(p_i + \epsilon \| p_i\right)\right)$$

where the KL divergence $\mathrm{KL}(q\|p)$ between two Bernoulli distributions with parameters q and p is given by

$$\mathrm{KL}(q\|p) = q \log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$$

For $q = p_i + \epsilon$ and $p = p_i$, the bound becomes:

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \exp\left(-n\left[\left(p_i + \epsilon\right)\log\left(\frac{p_i + \epsilon}{p_i}\right) + \left(1 - p_i - \epsilon\right)\log\left(\frac{1 - p_i - \epsilon}{1 - p_i}\right)\right]\right)$$

Remark: The proof of equivalence of the standard form and the KL divergence form of Chernoff's bound in Binomial distribution.

The standard form of Chernoff's bound gives an exponential decay for the upper tail of a sum of independent random variables. For $S_n \geq n\left(p_i + \epsilon\right)$, the bound is

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \min_{t>0}\left(\mathbb{E}\left[e^{tS_n}\right] e^{-\operatorname{tn}\left(p_i + \epsilon\right)}\right)$$

For binomial random variables, the moment generating function $\mathbb{E}\left[e^{tS_n}\right]$ is

$$\mathbb{E}\left[e^{tS_n}\right] = \left(p_i e^t + \left(1 - p_i\right)\right)^n$$

Thus, the bound can be expressed as

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \left(p_i e^t + \left(1 - p_i\right)\right)^n e^{-\operatorname{tn}\left(p_i + \epsilon\right)}$$

Optimizing for t using calculus, we get that the right-hand side is minimized if

$$e^t = \frac{\left(1 - p_i\right)\left(p_i + \epsilon\right)}{p_i\left(1 - p_i - \epsilon\right)}$$

Substituting this back into the bound, we obtain the KL divergence form of the Chernoff bound, which is equivalent to

$$\mathbb{P}\left(S_n \geq n\left(p_i + \epsilon\right)\right) \leq \exp\left(-nD\left(p_i + \epsilon \| p_i\right)\right)$$

2. **(LLN)** Suppose that X_1, $X_2, \ldots$ form an i.i.d. sequence of random variables with $\mathbb{E}[X_i] = \mu < \infty$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Evaluate

$$\lim_{n\to\infty} \frac{1}{\binom{n}{2}} \sum_{i,j:1\le i<j\le n} (X_i - X_j)^2.$$

First notice that $\sum_{1\le i<j\le n}(X_i - X_j)^2 = n\sum_{k=1}^n X_k^2 - \left(\sum_{k=1}^n X_k\right)^2$. Now using the law of larges numbers we get

$$\binom{n}{2}^{-1} n \sum_{k=1}^n X_k^2 = \frac{2n}{n-1} \frac{\sum_{k=1}^n X_k^2}{n} \to 2(\sigma^2 + \mu^2).$$

and that

$$\binom{n}{2}^{-1} (\sum_{k=1}^n X_k)^2 = \frac{2n}{n-1} \left(\frac{\sum_{k=1}^n X_k}{n}\right)^2 \to 2\mu^2.$$

We then have that $\binom{n}{2}^{-1} \sum_{1\le i<j\le n}(X_i - X_j)^2 \to 2(\sigma^2 + \mu^2 - \mu^2) = 2\sigma^2$.

3. **(Chebyshev & CLT)** Let $X_1, X_2, X_3, \ldots$ be i.i.d. random variables with mean zero and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Determine the limits below, with precise justifications.

   (a) $\lim_{n\to\infty} \mathbb{P}(S_n \ge 0.01n)$.

   (b) $\lim_{n\to\infty} \mathbb{P}(S_n \ge 0)$.

   (c) $\lim_{n\to\infty} \mathbb{P}(S_n \ge -0.01n)$.

   (a) Using Chebyshev's inequality, we get that

$$\mathbb{P}\left(S_n \ge 0.01n\right) \le \frac{n\sigma^2}{n^2/10^4} \to 0.$$

   (b) Using CLT, we get that

$$\mathbb{P}\left(\frac{S_n}{\sqrt{n}} \ge 0\right) \to \frac{1}{2}.$$

   (b) Using Chebyshev's inequality, we get that

$$\mathbb{P}\left(S_n \le -0.01n\right) \le \frac{n\sigma^2}{n^2/10^4} \to 0.$$

   Therefore,

$$\mathbb{P}\left(S_n \ge -0.01n\right) \to 1.$$

4. **(Convolution & MGF)** The Laplace distribution has density $f_Z(z) = \frac{\lambda}{2}\exp(-\lambda|z|)$ and MGF $M_Z(t) = \frac{\lambda^2}{\lambda^2 - t^2}$, where $\lambda > 0$. Let $X, Y \overset{\text{iid}}{\sim} \mathrm{Exp}(\lambda)$. Prove that $Z = X - Y$ follows a Laplace distribution by using:

   (a) Moment generating functions.

(b) The convolution formula.

(a) Our approach is to find the MGF of $Z = X - Y$ and match it to the moment generating function of the Laplace distribution, which is given to be $\frac{\lambda^2}{\lambda^2 - t^2}$. We have:

$$\begin{aligned} M_Z(t) &= M_X(t)M_{-Y}(t) \\ &= M_X(t)\mathbb{E}[e^{-tY}] \\ &= M_X(t)M_Y(-t) \\ &= \frac{\lambda}{\lambda - t}\frac{\lambda}{\lambda - (-t)} = \frac{\lambda^2}{\lambda^2 - t^2}, \end{aligned}$$

and therefore Z is Laplace.

(b) Using the convolution formula:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^{\infty} f_X(x)f_{-Y}(z-x)\,dx \\ &= \int_{-\infty}^{\infty} \lambda e^{-\lambda x} I_{(x>0)} \lambda e^{-\lambda(-(z-x))} I_{(z-x<0)} dx \\ &= \int_{-\infty}^{\infty} \lambda e^{-\lambda x} I_{(x>0)} \lambda e^{-\lambda(x-z)} I_{(x-z>0)} dx \\ &= \lambda^2 e^{\lambda z} \int_{-\infty}^{\infty} e^{-2\lambda x} I_{(x>0)} I_{(x-z>0)} dx \end{aligned}$$

We can simplify the indicators:

$$I_{(x>0)} \cdot I_{(x-z>0)} = I_{(x>0)\&(x>z)} = I_{x>\max(0,z)}.$$

Therefore:

$$\begin{aligned} f_Z(z) &= \lambda^2 e^{\lambda z} \int_{\max(z,0)}^{\infty} e^{-2\lambda x} dx \\ &= \lambda^2 e^{\lambda z} \left[-\frac{1}{2\lambda} e^{-2\lambda x} \right]_{x=\max(z,0)}^{x=\infty} \\ &= \lambda^2 e^{\lambda z} \cdot \frac{1}{2\lambda} e^{-2\lambda \max(z,0)} \\ &= \frac{\lambda}{2} e^{\lambda z} \cdot e^{-2\lambda \max(z,0)} \\ &= \frac{\lambda}{2} e^{-\lambda(2\max(z,0) - z)} \\ &= \frac{\lambda}{2} e^{-\lambda|z|} \end{aligned}$$

which is the Laplace(λ) PDF.

5. **(Moments & MGF)** Let X be a random variable with p.d.f. given by

$$f_X(x) = \begin{cases} 2/9, & \text{if } 0 \le x \le 1, \\ (4 - |4 - 2x|)/9, & \text{if } 1 < x \le 4, \\ 0, & \text{otherwise.} \end{cases}$$

(a) Verify that this is actually a p.d.f.

(b) Find the moment generating function of X.

(c) Find $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.

(d) Find a formula for the moments of X.

First note that the p.d.f. can also be written as

$$f_X(x) = \begin{cases} 2/9 \text{ if } 0 \le x \le 1 \\ 2x/9 \text{ if } 1 < x \le 2 \\ (8 - 2x)/9 \text{ if } 2 < x \le 4 \\ 0 \text{ else.} \end{cases}$$

(a) Let's verify that $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x)\,dx &= \int_0^1 2/9\,dx + \int_1^2 2x/9\,dx + \int_2^4 (8-2x)/9\,dx \\ &= \frac{2}{9}\Big([x]_0^1 + [x^2/2]_1^2 + [4x - x^2/2]_2^4\Big) \\ &= \frac{2}{9}\Big((1-0) + (2-1/2) + (8-6)\Big) = 1. \end{aligned}$$

(b) Let's do the full calculation,

$$\begin{aligned} \mathbb{E}[e^{tX}] &= \int_{-\infty}^{\infty} e^{tx} f_X(x)\,dx \\ &= \frac{2}{9}\Big(\int_0^1 e^{tx}\,dx + \int_1^2 xe^{tx}\,dx + \int_2^4 (4-x)e^{tx}\,dx\Big) \\ &= \frac{2}{9}\Big([\frac{e^{tx}}{t}]_0^1 + [\frac{(tx-1)e^{tx}}{t^2}]_1^2 + [\frac{(1+4t-tx)e^{tx}}{t^2}]_2^4\Big) \\ &= \frac{2}{9}\Big(\frac{e^t - 1}{t} + \frac{(2t-1)e^{2t} - (t-1)e^t}{t^2} + \frac{(1+4t-4t)e^{4t} - (1+4t-2t)e^{2t}}{t^2}\Big) \\ &= \frac{2}{9} \cdot \frac{e^{4t} + e^t - 2e^{2t} - t}{t^2} \end{aligned}$$

(c) One way to solve this problem is to directly calculate $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ using the formulas $\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,dx$ and $\mathbb{E}[X] = \int_{-\infty}^{\infty} x^2 f_X(x)\,dx$ and then $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. Another way to solve this part is using part 4. We obtain $\mathbb{E}[X] = \frac{49}{27}$, $\mathbb{E}[X^2] = \frac{25}{6}$ and then $\mathrm{Var}(X) = \frac{1273}{1458}$.

(d) Let's use the Taylor expansion for the exponential. We have that $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$. Replacing the series on the formula obtained in part 2. we obtain the following.

$$\begin{aligned}
\mathbb{E}[e^{tx}] &= \frac{2}{9} \cdot \frac{e^{4t} + e^t - 2e^{2t} - t}{t^2} \\
&= \frac{2}{9} \cdot \frac{\sum_{n=0}^{\infty} \frac{4^n t^n}{n!} + \sum_{n=0}^{\infty} \frac{t^n}{n!} - 2\sum_{n=0}^{\infty} \frac{2^n t^n}{n!} - t}{t^2} \\
&= \frac{2}{9} \cdot \frac{\sum_{n=0}^{\infty} \frac{1}{n!}(1 + 4^n - 2^{n+1})t^n - t}{t^2} \\
&= \frac{2}{9} \sum_{n=2}^{\infty} \frac{1 + 4^n - 2^{n+1}}{n!} t^{n-2} = \frac{2}{9} \sum_{n=0}^{\infty} \frac{1 + 4^{n+2} - 2^{n+3}}{(n+2)!} t^n
\end{aligned}$$

We conclude from this that the general formula for the moments is

$$\mathbb{E}(X^n) = \frac{2(1 + 4^{n+2} - 2^{n+3})}{9(n+1)(n+2)}.$$

6. **(Distribution of sums using MGFs)** Let $S_n := X_1 + \cdots + X_n$ for independent $X_1, \ldots, X_n$. Use MGFs to find the distribution of S_n

(a) For X_i with Normal $\left(\mu_i, \sigma_i^2\right)$ distribution.
The MGF of a normal random variable $X_i \sim$ Normal $\left(\mu_i, \sigma_i^2\right)$ is given by

$$M_{X_i}(t) = \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right)$$

Thus, for the sum $S_n = X_1 + X_2 + \cdots + X_n$, the MGF is

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \exp\left(\mu_i t + \frac{\sigma_i^2 t^2}{2}\right) = \exp\left(t\sum_{i=1}^{n} \mu_i + \frac{t^2}{2}\sum_{i=1}^{n} \sigma_i^2\right)$$

This is the MGF of a normal distribution with mean $\sum_{i=1}^{n} \mu_i$ and variance $\sum_{i=1}^{n} \sigma_i^2$. Therefore,

$$S_n \sim \text{Normal}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

(b) For X_i with Gamma (r_i, λ) distribution.
The MGF of a gamma random variable $X_i \sim$ Gamma (r_i, λ) (where r_i is the shape parameter and λ is the rate parameter) is given by

$$M_{X_i}(t) = \left(1 - \frac{t}{\lambda}\right)^{-r_i}, \quad \text{for } t < \lambda$$

Thus, for the sum $S_n = X_1 + X_2 + \cdots + X_n$, the MGF is

$$M_{S_n}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \prod_{i=1}^{n} \left(1 - \frac{t}{\lambda}\right)^{-r_i} = \left(1 - \frac{t}{\lambda}\right)^{-\sum_{i=1}^{n} r_i}$$

This is the MGF of a gamma distribution with shape parameter $\sum_{i=1}^{n} r_i$ and rate parameter λ. Therefore,

$$S_n \sim \text{Gamma}\left(\sum_{i=1}^{n} r_i, \lambda\right)$$

(c) For $X_i = Z_i^2$ with $Z_i \sim \text{Normal}(0, 1)$.
The MGF of X_i is the expectation of $e^{tZ_i^2}$. Applying the density function for Normal distribution, we have

$$\begin{aligned} M_{X_i(t)} &= \mathbb{E}\left[e^{tZ_i^2}\right] = \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}(1-2t)} dz \\ &= \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\frac{2\pi}{1-2t}} \\ &= \frac{1}{\sqrt{1-2t}}, \end{aligned}$$

which is valid for $t < \frac{1}{2}$. This is exactly the MGF for χ_1^2.

$$M_{S_n}(t) = \left((1-2t)^{-\frac{1}{2}}\right)^n = (1-2t)^{-\frac{n}{2}}, \quad \text{for } t < \frac{1}{2}$$

Therefore, S_n follows a chi-squared distribution with n degrees of freedom $S_n \sim \chi_n^2$.