Probability and Statistics Homework #13

Ari Feiglin

Question 13.1:

Suppose X is a random variable with a moment generating function:

$$M_X(t) = \frac{1}{4}e^t + \frac{1}{2}e^{2t} + \frac{1}{4}e^{-t}$$

(1) What is the expected value and variance of X?

(2) What is X's distribution?

(1) We know that the kth moment of X is equal to $M_X^{(k)}(0)$. We need to find the first and second moment ($\mathbb{E}[X]$ and $\mathbb{E}[X^2]$) for this part. So:

$$M'_X(t) = \frac{1}{4}e^t + e^{2t} - \frac{1}{4}e^{-t}$$

And

$$M''_X(t) = \frac{1}{4}e^t + 2e^{2t} + \frac{1}{4}e^{-t}$$

Therefore:

$$\mathbb{E}[X] = \frac{1}{4} + 1 - \frac{1}{4} = 1$$

And:

$$\mathbb{E}\left[X^2\right] = \frac{1}{4} + 2 + \frac{1}{4} = 2.5$$

This means that $\mathrm{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = 2.5 - 1 = 1.5$, so all in all:

$$\mathbb{E}[X] = 1 \qquad \mathrm{Var}(X) = 1.5$$

(2) Notice that adding the coefficients of $M_x(t)$ yields 1 ($\frac{1}{4}+\frac{1}{2}+\frac{1}{4}$), and since we know (assuming X is discrete):

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \sum_{x\in\mathbb{R}} \mathbb{P}(X = x) \cdot e^{tx}$$

So we can define $\mathbb{P}(X)$ as the coefficients of the terms in $M_X(t)$. And so if we define $\mathbb{P}(X = 1) = \frac{1}{4}$, $\mathbb{P}(X = 2) = \frac{1}{2}$, and $\mathbb{P}(X = -1) = e^{-t}$, we get that

$$M_X(t) = \frac{1}{4}e^t + \frac{1}{2}e^{2t} + \frac{1}{4}e^{-t}$$

As required.

Question 13.2:

A random variable X has a **Gamma Distribution** $\Gamma(n, \lambda)$ if it has a distribution:

$$f_X(t) = \begin{cases} \frac{\lambda^n}{(n-1)!} t^{n-1} e^{-\lambda t} & t > 0 \\ 0 & \text{else} \end{cases}$$

(1) Find the moment generating function of X.

(2) Show that if $X_1, \ldots, X_n \sim \mathrm{Exp}(\lambda)$ and are independent, then their sum has a distribution of $\Gamma(n, \lambda)$.

(3) Find the expected value and variance of X in two ways: using the moment generating function and the previous subquestion.

(4) Use Chernoff's inequality to show that for every $a > \frac{n}{\lambda}$, then

$$\mathbb{P}(X \geq a) \leq \left(\frac{a\lambda e}{n}\right)^n e^{-a\lambda}$$

(1) We know that:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx = \frac{\lambda^n}{(n-1)!} \cdot \int_0^{\infty} x^{n-1} e^{tx} e^{-\lambda x}\, dx = \frac{\lambda^n}{(n-1)!} \cdot \int_0^{\infty} x^{n-1} e^{x(t-\lambda)}\, dx$$

Note that if $t = \lambda$, this equals to the integral of x^{n-1} from 0 to ∞, which diverges. And if $t > \lambda$, then $x^{n-1} e^{x(t-\lambda)} \longrightarrow \infty$, so the integral diverges (as at some point the function is greater than 1, and the integral of 1 diverges). Otherwise, substituting $u = x(\lambda - t)$ gives $dx = \frac{du}{\lambda - t}$ so this integral equals:

$$= \frac{\lambda^n}{(n-1)!} \cdot \int_0^{\infty} \frac{u^{n-1}}{(\lambda-t)^{n-1}} e^{-u} \frac{1}{\lambda - t}\, du = \left(\frac{\lambda}{\lambda - t}\right)^n \cdot \frac{1}{(n-1)!} \cdot \int_0^{\infty} u^{n-1} e^{-u}\, du =$$

$$= \left(\frac{\lambda}{\lambda - t}\right)^n \cdot \frac{1}{(n-1)!} \cdot (n-1)! = \left(\frac{\lambda}{\lambda - t}\right)^n$$

So for $t < \lambda$ $M_X(t)$ is defined and equal to:

$$M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$$

(2) We know that:

$$M_{\sum X_i} = \prod M_{X_i}$$

So:

$$M_{\sum X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t)$$

And recall that $M_{X_i}(t) = \frac{\lambda}{t - \lambda}$ for $t < \lambda$, so:

$$M_{\sum X_i} = \left(\frac{\lambda}{t - \lambda}\right)^n$$

Which is the moment generating function of $\Gamma(n, \lambda)$, as required.

(3) Using $M_X(t)$ we know:

$$M_X'(t) = -n \cdot \left(-\frac{1}{\lambda}\right) \cdot \left(1 - \frac{t}{\lambda}\right)^{-n-1} = \frac{n}{\lambda}\left(1 - \frac{t}{\lambda}\right)^{-n-1}$$

And:

$$M_X''(t) = \frac{n}{\lambda} \cdot (-n-1) \cdot \left(-\frac{1}{\lambda}\right) \cdot \left(1 - \frac{t}{\lambda}\right)^{-n-2} = \frac{n(n+1)}{\lambda^2} \cdot \left(1 - \frac{t}{\lambda}\right)^{-n-2}$$

And we know that $\mathbb{E}\left[X\right] = M_X'(0) = \frac{n}{\lambda}$ and $\mathbb{E}\left[X^2\right] = M_X''(0) = \frac{n(n+1)}{\lambda^2}$. So:

$$\mathrm{Var}\left(X\right) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = \frac{n(n+1) - n^2}{\lambda^2} = \frac{n}{\lambda^2}$$

So all in all:

$$\mathbb{E}\left[X\right] = \frac{n}{\lambda} \quad \mathrm{Var}\left(X\right) = \frac{n}{\lambda^2}$$

On the other hand, we know that $X \overset{d}{=} \sum_{i=1}^n X_i$ if $X_i \sim \mathrm{Exp}\left(\lambda\right)$, so:

$$\mathbb{E}\left[X\right] = \sum_{i=1}^n \mathbb{E}\left[X_i\right] = \sum_{i=1}^n \frac{1}{\lambda} = \frac{n}{\lambda}$$

And:

$$\mathrm{Var}\left(X\right) = \sum_{i=1}^n \mathrm{Var}\left(X_i\right) = \sum_{i=1}^n \frac{1}{\lambda^2} = \frac{n}{\lambda^2}$$

(4) Recall that Chernoff's inequality is that for every $a \in \mathbb{R}$ and for every $t > 0$ where $M_X(t)$ is defined:

$$\mathbb{P}\left(X \geq a\right) \leq M_X(t)e^{-ta}$$

Let

$$f(t) = M_X(t)e^{-ta}$$

We want to find the minimum for $f(t)$. Computing its derivative yields:

$$f'(t) = M_X'(t)e^{-ta} - aM_X(t)e^{-ta} = e^{-ta}\left(1 - \frac{t}{\lambda}^{-n-1} - a\left(1 - \frac{t}{\lambda}\right)^{-n}\right) =$$

$$= e^{-ta} \cdot \left(\frac{\lambda}{\lambda - t}\right)\left(\frac{n}{\lambda} \cdot \frac{\lambda}{\lambda - t} - a\right) = e^{-ta} \cdot \left(\frac{\lambda}{\lambda - t}\right)\left(\frac{n}{\lambda - t} - a\right)$$

So $f'(t) = 0$ if $\frac{n}{\lambda - t} - a = 0 \iff t = \lambda - \frac{n}{a}$. But recall that $0 < t$ (this is sufficient as $M_X(t)$ is defined for $t > 0$ and $t < \lambda$ here already since $\lambda - \frac{n}{a} < \lambda$), so this is only true if:

$$\lambda - \frac{n}{a} > 0 \iff a\lambda > n \iff a > \frac{n}{\lambda}$$

And if this is true, we get that:

$$\mathbb{P}\left(X \geq a\right) \leq \left(\frac{\lambda}{\frac{n}{a}}\right)^n \cdot e^{n - a\lambda} = \left(\frac{a\lambda e}{n}\right)^n e^{-a\lambda}$$

As required.

Question 13.3:

We divide n balls into n urns randomly and independently. Let X be the number of balls in the first urn, and let m be the minimum integer such that $\mathbb{P}(X \geq m) \leq \frac{1}{n^2}$.

(1) Use Hoeffding's inequality to show that $m \leq 1 + 2\sqrt{n + \log n}$.

(2) Find an expression for m relative to n under the assumption that n is large enough.

(1) Recall that Hoeffding's inequality is that if X_k are independent and $|X_k - \mathbb{E}[X_k]| \leq M$ for some M then:

$$\mathbb{P}\left(\left|\sum_{k=1}^{n} X_k\right| \geq a\right) \leq e^{-\frac{a^2}{2nM^2}}$$

Let X_i indicate if the ith ball went into the first urn, then $X_i \sim \text{Ber}\left(\frac{1}{n}\right)$. And we know that X is the sum of X_is (this means that $X \sim \text{Bin}\left(n, \frac{1}{n}\right)$). Notice then that:

$$|X_i - \mathbb{E}[X_i]| = \left|X_i - \frac{1}{n}\right| \leq 1$$

Since $X_i \in \{1, 0\}$. Also note that $\mathbb{E}[X] = 1$ (since X is binomial). This means that:

$$\mathbb{P}(|X - 1| \geq a) \leq e^{-\frac{a^2}{2n}}$$

This means that:

$$\mathbb{P}(X \geq a + 1) \leq e^{-\frac{a^2}{2n}}$$

So if we require that $e^{-\frac{a^2}{2n}} = \frac{1}{n^2}$ we can get an upper bound for m:

$$\frac{a^2}{2n} = 2\log n \implies a^2 = 4n\log n \implies a = 2\sqrt{n\log n}$$

This means that:

$$\mathbb{P}\left(X \geq 1 + 2\sqrt{n\log n}\right) \leq \frac{1}{n^2}$$

Since m is the minimum (integer) where this occurs, this means that:

$$m \leq 1 + 2\sqrt{n\log n}$$

As required.

(2) Recall that $\mathbb{E}[X_i] = \frac{1}{n}$ and $\text{Var}(X_i) = \frac{1}{n} - \frac{1}{n^2}$. Using the central limit theorem, we get that X has an approximate distribution of $\mathcal{N}\left(1, 1 - \frac{1}{n}\right)$. Therefore

$$\mathbb{P}(X \geq m) \approx 1 - \Phi\left(\frac{m-1}{\sqrt{1 - \frac{1}{n}}}\right)$$

So we get that the probability is less than $\frac{1}{n^2}$ if:

$$1 - \Phi\left(\frac{m-1}{\sqrt{1 - \frac{1}{n}}}\right) \leq \frac{1}{n^2} \iff \Phi\left(\frac{m-1}{\sqrt{1 - \frac{1}{n}}}\right) \geq 1 - \frac{1}{n^2}$$

Since Φ is (strictly) monotonic increasing, it has an inverse which is also monotonic increasing and thus

$$m \geq \Phi^{-1}\left(1 - \frac{1}{n^2}\right) \cdot \sqrt{1 - \frac{1}{n}} + 1$$

Since m is the minimum *integer* which satisfies this, we get that:

$$m = \left\lceil \Phi^{-1}\left(1 - \frac{1}{n^2}\right) \cdot \sqrt{1 - \frac{1}{n}} + 1 \right\rceil$$

Question 13.4:

Suppose $X_1, \ldots, X_{20} \sim \text{Poi}(1)$ are independent. Let $S = \sum_{i=1}^{20} X_i$.

(1) Use Markov's and Chebyshev's inequalities to bound $\mathbb{P}(X > 30)$.

(2) Use the central limit theorem to estimate $\mathbb{P}(X > 30)$.

(1) We know $\mathbb{E}[X] = 20 \cdot \mathbb{E}[X_i] = 20$, so:

$$\mathbb{P}(X \geq 30) \leq \frac{20}{30} = \frac{2}{3}$$

And since $\mathbb{P}(X > 30) \leq \mathbb{P}(X \geq 30)$:

$$\mathbb{P}(X > 30) \leq \frac{2}{3}$$

We know that $\text{Var}(X) = 20 \cdot \text{Var}(X_i) = 20$, so: And using Chebyshev's inequality, we see that:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq 10) = \mathbb{P}(|X - 20| \geq 10) \leq \frac{\text{Var}(X)}{100} = \frac{2}{10}$$

So:

$$\mathbb{P}(X \geq 30) \leq \frac{2}{10} \implies \mathbb{P}(X > 30) \leq \frac{2}{10}$$

(2) Since $\mathbb{E}[X_i] = 1$ and $\text{Var}(X_i) = 1$, $X \sim \mathcal{N}(20, 20)$ so:

$$\mathbb{P}(X \geq 30) \approx 1 - \Phi\left(\frac{30 - 20}{\sqrt{20}}\right) = 1 - \Phi\left(\sqrt{5}\right) = 1 - 0.987 = 0.013$$

Question 13.5:

Sharon has a 6 year old son. On an average day, he spends 8 hours in front of a computer, with a standard deviation of 4 hours. She decided to measure the number of hours he spends on a computer each day over a course of 50 days. If the average number of hours is above 9, she will take away his computer.

(1) What is the probability she takes away his computer?

(2) What is the probability the averga is above 12 hours?

(1) Let X_i equal the number of hours her son spends on the computer on the ith day. This means that $\mathbb{E}[X_i] = 8$ and $\mathrm{Var}(X_i) = 4^2 = 16$. So the probability she takes away his computer is:

$$\mathbb{P}\left(\frac{1}{50}\sum_{i=1}^{50} X_i \geq 9\right)$$

Assuming that 50 is sufficiently large, using the central limit theorem, we can approximate that:

$$\frac{1}{50}\sum_{i=1}^{50} X_i \overset{\text{approx.}}{\sim} \mathcal{N}\left(8, \frac{4}{50}\right)$$

So this probability is approximately:

$$1 - \Phi\left(\frac{9-8}{\frac{4}{\sqrt{50}}}\right) = 1 - \Phi\left(\frac{\sqrt{50}}{4}\right) \approx 0.0385$$

(2) Again, we can assume that

$$\frac{1}{50}\sum_{i=1}^{50} X_i \overset{\text{approx.}}{\sim} \mathcal{N}\left(8, \frac{16}{50}\right)$$

So:

$$\mathbb{P}\left(\frac{1}{50}\sum_{i=1}^{50} X_i \geq 12\right) \approx 1 - \Phi\left(\frac{12-8}{\frac{4}{\sqrt{50}}}\right) = 1 - \Phi\left(\sqrt{50}\right) \approx 0$$

Question 13.6:

Each analyst at a company creates on average 30 graphs per week, with a standard deviation of 5 graphs. The company has 100 analysts.

(1) What is the probability that in total they drew at least 3,040 graphs in a week?

(2) What is the probability that they drew exactly 3,040 graphs?

(1) Let X_i be the number of graphs the ith analyst drew. This means that $\mathbb{E}\left[X_i\right] = 30$ and $\mathrm{Var}\left(X_i\right) = 5^2 = 25$. Let $X = \sum_{i=1}^{100} X_i$. The probability that they drew at least 3,040 graphs is:

$$\mathbb{P}\left(X \geq 3{,}040\right)$$

Using the central limit theorem, we can infer that X has an approximate distribution of $\mathcal{N}\left(3000, 2500\right)$. Therefore the probability is approximately:

$$1 - \Phi\left(\frac{3040 - 3000}{50}\right) = 1 - \Phi\left(\frac{4}{5}\right) \approx 0.21186$$

(2) We know that:

$$\mathbb{P}\left(X = 3040\right) = \mathbb{P}\left(3039.5 \leq X \leq 3040.5\right) = \mathbb{P}\left(X \leq 3040.5\right) - \mathbb{P}\left(X < 3039.5\right)$$

Since X has an approximate distribution of $\mathcal{N}\left(3000, 2500\right)$ this is approximately equal to:

$$\Phi\left(\frac{40.5}{50}\right) - \Phi\left(\frac{39.5}{50}\right) = \Phi\left(0.81\right) - \Phi\left(0.79\right) \approx 0.00579$$

Question 13.7:

The ratios of people who have blue, green, brown, and hazel eyes is 1 : 2 : 3 : 4 respectively. In a group there are 400 people.

(1) What is the probability that at least 90 people have green eyes?

(2) What is the probability that the number of people with hazel eyes is at least 30 more than the number of people with brown and blue eyes together?

(1) Let Blue_i be the probability that the ith person has blue eyes, so $\text{Blue}_i \sim \text{Ber}\left(\frac{1}{0}\right)$. Similarly for the rest of the colors. And let Blue equal the number of people with blue eyes, so $\text{Blue} = \sum_{i=1}^{400} \text{Blue}_i$. Similarly for the rest of the colors.
So we want to compute:

$$\mathbb{P}\left(\text{Green} \geq 90\right)$$

Since $\mathbb{E}\left[\text{Green}_i\right] = 0.2$ and $\text{Var}\left(\text{Green}_i\right) = 0.16$, Green has an approximate distribution of

$$\mathcal{N}\left(400 \cdot 0.2, 400 \cdot 0.16\right) = \mathcal{N}\left(80, 64\right).$$

So:

$$\mathbb{P}\left(\text{Green} \geq 90\right) \approx 1 - \Phi\left(\frac{90-80}{8}\right) = 1 - \Phi\left(1.25\right) = 0.106$$

(2) We want to compute $\mathbb{P}\left(\text{Hazel} \geq 30 + \text{Brown} + \text{Blue}\right)$. Notice that $\text{Brown} + \text{Blue} + \text{Hazel} + \text{Green} = 400$, so this is equal to $\mathbb{P}\left(2\text{Hazel} + \text{Green} \geq 430\right)$. Let us define:

$$X_i = \begin{cases} 2 & \text{the } i\text{th person has hazel eyes} \\ 1 & \text{the } i\text{th person has green eyes} \\ 0 & \text{else} \end{cases}$$

This means that:

$$2\text{Hazel} + \text{Green} = \sum_{i=1}^{400} X_i$$

Since the sum of X_i doubly counts the number of people with hazel eyes and counts the number of people with green eyes. We know that $\mathbb{E}\left[X_i\right] = 2 \cdot \frac{4}{10} + 1 \cdot \frac{2}{10} = 1$ and

$$\mathbb{E}\left[X_i^2\right] = 4 \cdot \frac{4}{10} + 1 \cdot \frac{2}{10} = 1.8$$

So:

$$\text{Var}\left(X_i\right) = 1.8 - 1 = 0.8$$

This means that $2\text{Hazel} + \text{Green}$ has an approximate distribution of $\mathcal{N}\left(400, 400 \cdot 0.8\right) = \mathcal{N}\left(720, 320\right)$. Therefore:

$$\mathbb{P}\left(2\text{Hazel} + \text{Green} \geq 430\right) \approx 1 - \Phi\left(\frac{430-400}{\sqrt{320}}\right) = 1 - \Phi\left(1.677\right) \approx 0.04648$$

Question 13.8:

Suppose $\{X_i\}_{i=1}^{\infty}$ is a series of independent random variables such that $X_i \sim \text{Geo}\left(\frac{1}{2}\right)$. Compute:

$$\lim \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq 2\right)$$

So we know that $\mathbb{E}[X_i] = 2$ and $\text{Var}(X_i) = \frac{1-\frac{1}{2}}{\frac{1}{4}} = 2$. This means that:

$$\frac{\sum_{i=1}^{n} X_i - 2n}{\sqrt{2n}} \xrightarrow{d} \mathcal{N}(1, 0)$$

This means that:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - 2n}{\sqrt{2n}} \leq a\right) = \Phi(a)$$

And therefore:

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - 2n}{\sqrt{2n}} \geq a\right) = 1 - \Phi(a)$$

Notice that by using a wee bit of algebraic manipulation:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq 2\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - 2n}{\sqrt{2n}} \geq \frac{2n-2n}{\sqrt{2n}}\right) = \mathbb{P}\left(\frac{\sum_{i=1}^{n} X_i - 2n}{\sqrt{2n}} \geq 0\right)$$

And as explained above, the limit of this is $1 - \Phi(0) = 1 - \frac{1}{2} = \frac{1}{2}$. So all in all:

$$\lim \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq 2\right) = \frac{1}{2}$$