Probability (Level M) [STATS 5024]

Lecture 18

Sums and Means of Random Variables: Exact Results

Sums and Means of Random Variables

- We are often interested in the properties of the sum or mean of a sequence of random variables
- Today, we will introduce methods we can sometimes use to obtain exact probability information about these derived random variables
- Next week, we will discuss methods that, although more generally applicable, give only approximations to probabilities of interest

Recap

Recall for independent random variables $X_1, X_2, \ldots, X_p$ and real constants $a_0, a_1, \ldots, a_p$:

$$\mathrm{E}(X_1 + \cdots + X_p) = \sum_{i=1}^{p} \mathrm{E}(X_i)$$

$$\mathrm{E}(a_0 + a_1 X_1 + \cdots + a_p X_p) = a_0 + \sum_{i=1}^{p} a_i \mathrm{E}(X_i)$$

$$\mathrm{Var}(X_1 + \cdots + X_p) = \sum_{i=1}^{p} \mathrm{Var}(X_i)$$

$$\mathrm{Var}(a_0 + a_1 X_1 + \cdots + a_p X_p) = \sum_{i=1}^{p} a_i^2 \mathrm{Var}(X_i)$$

Example

Example 1

- A CD player has a shuffle feature
- The next track to be played is always selected randomly from all the tracks on the CD, so that each track is equally-likely to be played
- We will call the playing of a randomly-selected track a trial
- Suppose there are ten tracks on a CD.
- What can be said about the random variable, S, the number of trials required until every track on the CD is played at least once?

Solution I

- Let the discrete random variable X_1 be the number of trials required until the first new track is played

Solution I

- Let the discrete random variable X_1 be the number of trials required until the first new track is played
  - X_1 takes the value 1 with probability 1

Solution I

- Let the discrete random variable X_1 be the number of trials required until the first new track is played
  - X_1 takes the value 1 with probability 1
- Let the discrete random variable X_2 be the number of further trials required, after the first track is played, until another new track is played

Solution I

- Let the discrete random variable X_1 be the number of trials required until the first new track is played
  - X_1 takes the value 1 with probability 1
- Let the discrete random variable X_2 be the number of further trials required, after the first track is played, until another new track is played
  - On each trial, there is probability 0.1 of playing the track that has already been played and probability 0.9 of playing a new track
  - Trials are independent
  - So, $X_2 \sim \text{Geo}(0.1)$

Solution II

- Let the random variable X_i $(i = 3, \ldots, 10)$ be the number of further trials required, after the $(i-1)$th different track is played for the first time, until the ith different track is played

Solution II

- Let the random variable X_i $(i = 3, \ldots, 10)$ be the number of further trials required, after the $(i-1)$th different track is played for the first time, until the ith different track is played
- Then $X_i \sim \mathrm{Geo}(\theta_i)$ where $\theta_i = (i-1)/10$

Solution II

- Let the random variable X_i $(i = 3, \dots, 10)$ be the number of further trials required, after the $(i-1)$th different track is played for the first time, until the ith different track is played
- Then $X_i \sim \mathrm{Geo}(\theta_i)$ where $\theta_i = (i-1)/10$

$$\mathrm{E}(X_i) = \frac{1}{1-\theta_i} = \frac{1}{1-(i-1)/10} = \frac{10}{11-i}$$

$$\mathrm{Var}(X_i) = \frac{\theta_i}{(1-\theta_i)^2} = \frac{(i-1)/10}{(11-i)^2/100} = \frac{10(i-1)}{(11-i)^2}$$

Solution II

- Let the random variable X_i $(i = 3, \dots, 10)$ be the number of further trials required, after the $(i-1)$th different track is played for the first time, until the ith different track is played
- Then $X_i \sim \mathrm{Geo}(\theta_i)$ where $\theta_i = (i-1)/10$

$$\mathrm{E}(X_i) = \frac{1}{1-\theta_i} = \frac{1}{1-(i-1)/10} = \frac{10}{11-i}$$

$$\mathrm{Var}(X_i) = \frac{\theta_i}{(1-\theta_i)^2} = \frac{(i-1)/10}{(11-i)^2/100} = \frac{10(i-1)}{(11-i)^2}$$

- $S = X_1 + X_2 + \dots + X_{10}$ $(X_1, \dots, X_{10}$ independent). So

$$\mathrm{E}(S) = \frac{10}{10} + \frac{10}{9} + \dots + \frac{10}{1} = 29.29$$

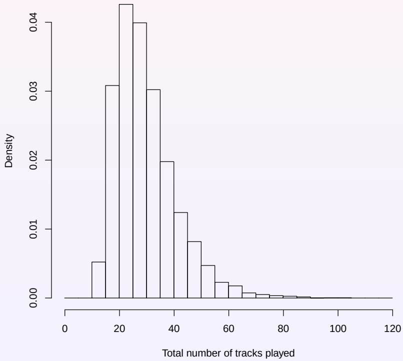
$$\mathrm{Var}(S) = \frac{10 \times 0}{10^2} + \frac{10 \times 1}{9^2} + \dots + \frac{10 \times 9}{1^2} = 125.69$$

Solution III

- On average, 29.29 tracks have to be played in order to hear all ten different tracks on the CD at least once
- This might not seem a lot
- However, the variance is very large relative to the expected value so there is a good chance that many more than this average number will be required on a particular occasion

Solution IV

- We are not in a position to obtain probabilities related to S algebraically, but the results from $10,000$ simulations can be plotted:



- The distribution of S is highly skewed towards larger values

m.g.f. to the rescue

There are many important cases where we can use moment-generating functions to determine the entire probability distribution of the sum, S, of a sequence of random variables

Proposition 4.1

Suppose that $X_1, \dots, X_n$ are independent random variables, each with a finite moment generating function $M_i(t)$. Then the moment-generating function of

$$S = X_1 + \cdots + X_n$$

is

$$M_S(t) = \prod_{i=1}^{n} M_i(t)$$

Proof

$$
\begin{aligned}
M_S(t) = \mathrm{E}(e^{St}) &= \mathrm{E}\left(e^{(X_1+\cdots+X_n)t}\right) \\
&= \mathrm{E}\left(e^{X_1 t} e^{X_2 t} \cdots e^{X_n t}\right) \\
&= \mathrm{E}\left(e^{X_1 t}\right) \mathrm{E}\left(e^{X_2 t}\right) \cdots \mathrm{E}\left(e^{X_n t}\right) \qquad \text{independence} \\
&= \prod_{i=1}^{n} M_i(t)
\end{aligned}
$$

Proposition 4.1 allows us to obtain a variety of important results, known as **reproductive properties**, that arise when the sum of independent random variables drawn from a given family of distributions has a distribution that also belongs to that family

Example

Example 2

Suppose that $X_1, \ldots, X_n$ are independent random variables and that $X_i \sim \text{Poi}(\lambda_i)$ for $\lambda_i > 0$ $(i = 1, \ldots, n)$. Let $S = X_1 + \cdots + X_n$. What is the distribution of S?

Example

Example 2

Suppose that $X_1, \ldots, X_n$ are independent random variables and that $X_i \sim \mathrm{Poi}(\lambda_i)$ for $\lambda_i > 0$ $(i = 1, \ldots, n)$. Let $S = X_1 + \cdots + X_n$. What is the distribution of S?

$$M_i(t) = \exp[\lambda_i(e^t - 1)]$$

Example

Example 2

Suppose that $X_1, \ldots, X_n$ are independent random variables and that $X_i \sim \text{Poi}(\lambda_i)$ for $\lambda_i > 0$ $(i = 1, \ldots, n)$. Let $S = X_1 + \cdots + X_n$. What is the distribution of S?

$$M_i(t) = \exp[\lambda_i(e^t - 1)]$$

$$M_S(t) = \prod_{i=1}^{n} M_i(t) = \prod_{i=1}^{n} \exp[\lambda_i(e^t - 1)]$$

$$= \exp\left(\sum_{i=1}^{n} \lambda_i(e^t - 1)\right) = \exp\left((e^t - 1)\sum_{i=1}^{n} \lambda_i\right)$$

Example

Example 2

Suppose that $X_1, \ldots, X_n$ are independent random variables and that $X_i \sim \text{Poi}(\lambda_i)$ for $\lambda_i > 0$ $(i = 1, \ldots, n)$. Let $S = X_1 + \cdots + X_n$. What is the distribution of S?

$$\begin{aligned} M_i(t) &= \exp[\lambda_i(e^t - 1)] \\ M_S(t) &= \prod_{i=1}^{n} M_i(t) = \prod_{i=1}^{n} \exp[\lambda_i(e^t - 1)] \\ &= \exp\left(\sum_{i=1}^{n} \lambda_i(e^t - 1)\right) = \exp\left((e^t - 1)\sum_{i=1}^{n} \lambda_i\right) \end{aligned}$$

This is the m.g.f. of the Poisson distribution with parameter $\sum_{i=1}^{n} \lambda_i$. By the uniqueness property of m.g.f.s, $S \sim \text{Poi}\left(\sum_{i=1}^{n} \lambda_i\right)$

Example

Example 3

Suppose that $X_1, \ldots, X_n$ are independent random variables and that $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. What is the distribution of $S = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$, where $a_0, a_1, a_2, \ldots, a_n$ are real constants (not all of which are zero)

Recall

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

Solution I

$$S = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

Solution I

$$S = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

- Let $Y_i = a_iX_i$ and then $M_{Y_i}(t) = M_{X_i}(a_it)$ (Proposition 2.7)

Solution I

$$S = a_0 + a_1X_1 + a_2X_2 + \cdots + a_nX_n$$

- Let $Y_i = a_iX_i$ and then $M_{Y_i}(t) = M_{X_i}(a_it)$ (Proposition 2.7)
- Let $Z = \sum_{i=1}^n Y_i$, then

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

Solution I

$$S = a_0 + a_1 X_1 + a_2 X_2 + \cdots + a_n X_n$$

- Let $Y_i = a_i X_i$ and then $M_{Y_i}(t) = M_{X_i}(a_i t)$ (Proposition 2.7)
- Let $Z = \sum_{i=1}^{n} Y_i$, then

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

- $S = a_0 + Z$, so (Proposition 2.7)

$$\begin{aligned} M_S(t) &= e^{a_0 t} \prod_{i=1}^{n} \exp\left(\mu_i a_i t + \frac{1}{2}\sigma_i^2 a_i^2 t^2\right) \\ &= \exp\left[(a_0 + a_1\mu_1 + \cdots + a_n\mu_n)t + \frac{1}{2}(a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)t^2\right] \end{aligned}$$

Solution II

$$M_S(t) = \exp\left[(a_0 + a_1\mu_1 + \cdots + a_n\mu_n)t + \frac{1}{2}(a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2)t^2\right]$$

- This is the moment-generating function of the Normal distribution with:
  - expected value $a_0 + a_1\mu_1 + \cdots + a_n\mu_n$ and
  - variance $a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2$
- Using the Uniqueness Property of moment-generating functions, $S \sim \mathrm{N}\left(a_0 + a_1\mu_1 + \cdots + a_n\mu_n, a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2\right)$
- Note we already knew these expressions for $\mathrm{E}(S)$ and $\mathrm{Var}(S)$; it is the fact that S is normally distributed that is the new result

Sums and means of i.i.d. normal random variables

- Now suppose that $X_1, \ldots, X_n$ are identically distributed, as well as independent (i.i.d.)
- This means that $\mu_1 = \cdots = \mu_n \equiv \mu$ and $\sigma_1 = \cdots = \sigma_n \equiv \sigma$
- Consider the sum of the random variables $S = X_1 + \cdots + X_n$
- Setting $a_0 = 0$ and $a_1 = \cdots = a_n = 1$, the general result we have just proved shows that

$$S \sim N(n\mu, n\sigma^2)$$

- Consider next the sample mean of the random variables, $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$
- Setting $a_0 = 0$ and $a_1 = \cdots = a_n = \frac{1}{n}$, then the same general result shows that

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$