SI140 Discussion 06

Li Zeng, Tao Huang, Xinyi Liu

ShanghaiTech University, China
{zengli,huangtao1,liuxy10}@shanghaitech.edu.cn

1 The Law of Small Numbers

In 1898, Ladislaus von Bortkiewicz published Das Gesetz der kleinen Zahlen [The Law of Small Numbers]. He described a number of observations on the frequency of occurrence of **rare events** that appear to follow a Poisson distribution.

Proposition 1 (The Law of Small Numbers). *Given independent random variables* $Y_1, \cdots, Y_n$ *such that for any* $1 \leq i \leq n, P\left(Y_i=1\right)=p_i$ *and* $P\left(Y_i=0\right)=1-p_i$. *Let* $S_n=Y_1+\cdots+Y_n$. *Suppose*

$$\sum_{i=1}^n p_i \rightarrow \lambda \in(0, \infty) \quad \text{as } n \rightarrow \infty$$

and

$$\max_{1 \leq i \leq n} p_i \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

then

$$d_{TV}\left(S_n, \operatorname{Poi}(\lambda)\right) \rightarrow 0 \quad \text{as } n \rightarrow \infty$$

LSN is in fact an extension of Poisson approximation to Binomial which is a special case when $p_i=\frac{\lambda}{n}$. In LSN setting, we assume numerous times ($n \rightarrow \infty$)independent but not identical Bernoulli trails with varying parameter p_i for the i-th experiment. As long as the summation of p_i converges to a constant λ, then the sum of these r.v. converges to a Poisson r.v. with mean λ without requiring p_i a constant.

2 Probability Generating Functions

Definition 1. *The probability generating function (PGF) of a nonnegative integer-valued r.v. X with* PMF $p_k=P(X=k)$ *is the generating function of the PMF. By LOTUS, this is*

$$E\left(t^X\right)=\sum_{k=0}^{\infty} p_k t^k$$

Property 1. The PGF converges to a value in [-1,1] for all t in [-1,1] since $\sum_{k=0}^{\infty} p_k=1$ and $\left|p_k t^k\right| \leq p_k$ for $|t| \leq 1$.

Property 2. It is to be noted that $G(0)=\mathrm{P}(X=0)$ and, rather more importantly, that:

$$G(1)=\mathrm{P}(X=0)+\mathrm{P}(X=1)+\mathrm{P}(X=2)+\cdots=\sum_r \mathrm{P}(X=r)=1$$

Property 3. Consider $G(\eta)$ together with its first and second derivatives $G'(\eta)$ and $G''(\eta)$ when $\eta=1$

$$G(1)=\sum_r \mathrm{P}(X=r)=1$$

$$G'(1)=\sum_r r . \mathrm{P}(X=r)=\mathrm{E}(X)$$

$$G''(1)=\sum_r r(r-1) \cdot \mathrm{P}(X=r)=\mathrm{E}(X(X-1))$$

Property 4. By differentiating the generating function one can directly obtain the expectation $\mathrm{E}(X)$. By differentiating again one can obtain $\mathrm{E}(X(X-1))$. These two results lead to the rapid derivation of $\mathrm{V}(X)$. Note:

$$\begin{aligned} G''(1)+G'(1)-\left(G'(1)\right)^2 &= \mathrm{E}(X(X-1))+\mathrm{E}(X)-(\mathrm{E}(X))^2 \\ &= \mathrm{E}\left(X^2\right)-\mathrm{E}(X)+\mathrm{E}(X)-(\mathrm{E}(X))^2 \\ &= \mathrm{E}\left(X^2\right)-(\mathrm{E}(X))^2 \\ &= \mathrm{Var}(X) \end{aligned}$$

Property 5 (Sum of Discrete **RVs***).* If X and Y are independent, then $G_{X+Y}(t)=G_X(t)G_Y(t)$. (can be extended to $n \geq 2$ case)

Proof.

$$\begin{aligned} G_X(z)G_Y(z) &= \left(\sum_{k=0}^{\infty}\Pr[X=k]z^k\right)\left(\sum_{k=0}^{\infty}\Pr[Y=k]z^k\right) \\ &= \sum_{k=0}^{\infty} z^k\left(\sum_{\ell=0}^{k}\Pr[X=\ell]\Pr[Y=k-\ell]\right) \\ &= \sum_{k=0}^{\infty} z^k\left(\sum_{\ell=0}^{k}\Pr[X=\ell,Y=k-\ell]\right) \\ &= \sum_{k=0}^{\infty}\sum_{\ell=0}^{k}\Pr[X+Y=k]z^k = G_{X+Y}(z) \end{aligned}$$

Property 6. $P(X=n)=\frac{G^{(n)}(0)}{n!}$, for every $n \in \{0,1,2,\cdots\}$, where $G^{(n)}(t) := \frac{\mathrm{d}^n G(t)}{\mathrm{d}z^n}$.

Property 7 (Linear Transformation). If $Y=aX+b$ for scalars $a,b \in R$, then $G_Y(z)=z^aG_X(z^b)$.

Property 8 (Random Sum of discrete **RVs***).* Let $Y=\sum_{i=1}^{N}X_i$, where X_i 's are i.i.d discrete positive integer valued random variables and N is independent of X_i 's. The PGF of Y is $G_Y(z)=G_N\left(G_X(z)\right)$.

Proof. $G_Y(z)=E\left[z^Y\right]=E\left[E\left[z^Y \mid N\right]\right]$ (By law of total expectation). Now,

$$E\left[z^Y \mid N=n\right]=E\left[z^{\sum x_i} \mid N=n\right]=E\left[G_X(z)^N\right]$$

This implies that

$$G_Y(z)=G_N\left(G_X(z)\right)$$

Theorem 1 (Uniqueness Theorem for PGF). *Let $X,Y \geq 0$ be discrete random variables. Then X and Y have the same distribution if and only if $G_X=G_Y$.*

Remark 1. **If two power series agree on any interval containing 0, however small, then all terms of the two series are equal**. Mathematically, let $A(s)$ and $B(s)$ be PGFs with $A(s)=\sum_{n=0}^{\infty}a_ns^n, B(s)=\sum_{n=0}^{\infty}b_ns^n$. If there exists some $R>0$ such that $A(s)=B(s)$ for all $s \in (-R,R)$, then $a_n=b_n$ for all n.

Remark 2. The practical use of this theorem is that if we can show that two random variables have the same PGF in some interval containing 0, then we have shown that the two random variables have the same distribution.

Example 1. Reconsider the property 8 (random sum of discrete RVs), i.e. let N be a positive integer random variable and, given N, let $X_1,X_2,\cdots$ be i.i.d.and independent of N. To move a step further, prove that:

1. $E(S_N)=E(N)E(X_1)$
2. $\mathrm{Var}(S_N)=E(N)\,\mathrm{Var}(X_1)+\mathrm{Var}(N)(E(X_1))^2$

Solution 1. 1. Recall that $E(S_N) = G'_{S_N}(1)$, by the chain rule

$$\begin{aligned} G'_{S_N}(1) &= G'_N(G_{X_1}(1))G'_{X_1}(1) \\ &= G'_N(1)E(X_1) \\ &= E(N)E(X_1) \end{aligned}$$

2. Recall that $\text{Var}(S_N) = G''_{S_N}(1) + G'_{S_N}(1) - (G'_{S_N}(1))^2$, by the chain rule

$$\begin{aligned} G''_{S_N}(1) &= G'_N(G_{X_1}(1))G''_{X_1}(1) + G''_N(G_{X_1}(1))(G'_{X_1}(1))^2 \\ &= E(N)G''_{X_1}(1) + G''_N(1)(E(X_1))^2 \\ &= E(N)\left((E(X_1^2)) - E(X_1)\right) + \left((E(N^2)) - E(N)\right)(E(X_1))^2 \end{aligned}$$

$$\begin{aligned} \text{Var}(S_N) &= G''_{S_N}(1) + G'_{S_N}(1) - (G'_{S_N}(1))^2 \\ &= E(N)\left(E(X_1^2) - E(X_1)\right) + \left(E(N^2) - E(N)\right)(E(X_1))^2 + E(N)E(X_1) - (E(N)E(X_1))^2 \\ &= E(N)\left(E(X_1^2) - (E(X_1))^2\right) + E(X_1^2)\left(E(N^2) - (E(N))^2\right) \\ &= E(N)\,\text{Var}(X_1) + \text{Var}(N)(E(X_1))^2 \end{aligned}$$

3 Continuous Distributions

So far, all random variables we have seen have been discrete. In all the cases we have seen in SI140, this meant that our RVs could only take on integer values. Now it's time for continuous random variables, which can take on values in the real number domain (R). Continuous random variables can be used to represent measurements with arbitrary precision (e.g., height, weight, or time).

3.1 Probability Density Functions

Definition 2 (PDF). *For a continuous r.v. X with CDF F, the probability density function (PDF) of X is the derivative f of the CDF, given by $f(x) = F'(x)$. The support of X, and of its distribution,is the set of all x where $f(x) > 0$.*

Motivation of PDF

In the world of discrete random variables, the most important property of a random variable was its probability mass function (PMF), which told you the probability of the random variable taking on a certain value. When we move to the world of continuous random variables, we are going to need to rethink this basic concept. If I were to ask you what the probability is of a child being born with a weight of exactly 3.523112342234 kilograms, you might recognize that question as ridiculous. No child will have precisely that weight. Real values are defined with infinite precision; as a result, the probability that a random variable takes on a specific value is not very meaningful when the random variable is continuous. The PMF doesn't apply. We need another idea.

In the continuous world, every random variable has a probability density function (PDF), which says how likely it is that a random variable takes on a particular value, relative to other values that it could take on. The PDF has the nice property that you can integrate over it to find the probability that the random variable takes on values within a range (a, b).

Properties of PDF

X is a continuous random variable if there is a function $f(x)$ for $-\infty \leq x \leq \infty$, called the probability density function (PDF), such that:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

To preserve the axioms that guarantee $P(a \leq X \leq b)$ is a probability, the following properties must also hold:

$$0 \leq P(a \leq X \leq b) \leq 1$$
$$P(-\infty < X < \infty) = 1$$

Remark 3. A common **misconception** is to think of $f(x)$ as a probability. It is instead what we call a probability density. It represents probability divided by the units of X. Generally this is only meaningful when we either take an integral over the PDF or we compare probability densities. As we mentioned when motivating probability densities, the probability that a continuous random variable takes on a specific value (to infinite precision) is 0.

$$P(X = a) = \int_a^a f(x)dx = 0$$

This is very different from the discrete setting, in which we often talked about the probability of a random variable taking on a particular value exactly.

3.2 Cumulative Distribution Function

Definition 3 (CDF). *The CDF is a function which takes in a number and returns the probability that a random variable takes on a value less than (or equal to) that number. If we have a CDF for a random variable, we don't need to integrate to answer probability questions! For a continuous random variable X, the cumulative distribution function is:*

$$F_X(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$$

This can be written $F(a)$, without the subscript, when it is obvious which random variable we are using.

Probability Query	**Solution**	**Explanation**
$P(X \leq a)$	$F(a)$	This is the definition of the CDF
$P(X < a)$	$F(a)$	Note that $P(X = a) = 0$
$P(X > a)$	$1 - F(a)$	$P(X \leq a) + P(X > a) = 1$
$P(a < X < b)$	$F(b) - F(a)$	$F(a) + P(a < X < b) = F(b)$

Exercise 1 (Disk crashes). Let X be a RV representing the number of days of use before your disk crashes, with PDF:

$$f(x) = \begin{cases} \lambda e^{-x/100} & \text{when } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

First, determine λ. Recall that $\int Ae^{Au}du = e^{Au}$:

$$\begin{aligned} \int_0^{\infty} \lambda e^{-x/100} dx &= 1 \\ -100\lambda \int_0^{\infty} \frac{-1}{100} e^{-x/100} dx &= 1 \\ -100\lambda \cdot e^{-x/100} \Big|_{x=0}^{\infty} &= 1 \\ 100\lambda \cdot 1 &= 1 \quad \Rightarrow \quad \lambda = 1/100 \end{aligned}$$

3.3 Expectation and Variance

For continuous RV X :

$$E[X] = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)dx$$

$$E\left[X^n\right]=\int_{-\infty}^{\infty} x^n \cdot f(x) d x$$

For both continuous and discrete RVs:

$$E[a X+b]=a E[X]+b$$
$$\operatorname{Var}(X)=E\left[(X-\mu)^2\right]=E\left[X^2\right]-(E[X])^2$$
$$\operatorname{Var}(a X+b)=a^2 \operatorname{Var}(X)$$

Expectation via Survival Function Let X be a continuous and **non-negative** r.v. Let F be the CDF of X, and $G(x)=1-F(x)=P(X>x)$. The function G is called the survival function of X. Then

$$E(X)=\int_0^{\infty} G(x) \mathrm{d} x$$

Proof. With $G(x)=P(X \geq x)=\int_x^{\infty} f(y) \mathrm{d} y$

$$\begin{aligned}
\int_0^{\infty} G(x) \mathrm{d} x & =\int_0^{\infty} \int_x^{\infty} f(y) \mathrm{d} y \mathrm{~d} x \\
& =\int_0^{\infty} \int_0^y f(y) \mathrm{d} x \mathrm{~d} y \\
& =\int_0^{\infty}\left(\int_0^y \mathrm{~d} x\right) f(y) \mathrm{d} y \\
& =\int_0^{\infty} y f(y) \mathrm{d} y \\
& =E(X)
\end{aligned}$$

3.4 Uniform Random Variable

The most basic of all the continuous random variables is the uniform random variable, which is equally likely to take on any value in its range (α, β) X is a uniform random variable ($X \sim \operatorname{Uni}(\alpha, \beta)$) if it has PDF:

$$f(x)= \begin{cases}\frac{1}{\beta-\alpha} & \text { when } \alpha \leq x \leq \beta \\ 0 & \text { otherwise }\end{cases}$$

Notice how the density $1 /(\beta-\alpha)$ is exactly the same regardless of the value for x. That makes the density uniform. So why is the PDF $1 /(\beta-\alpha)$ and not 1 ? That is the constant that makes it such that the integral over all possible inputs evaluates to 1 . The key properties of this RV are:

$$P(a \leq X \leq b)=\int_a^b f(x) d x=\frac{b-a}{\beta-\alpha}(\text { for } \alpha \leq a \leq b \leq \beta)$$
$$E[X]=\int_{-\infty}^{\infty} x \cdot f(x) d x=\int_\alpha^\beta \frac{x}{\beta-\alpha} d x=\left.\frac{x^2}{2(\beta-\alpha)}\right|_{x=\alpha} ^\beta=\frac{\alpha+\beta}{2}$$
$$\operatorname{Var}(X)=\frac{(\beta-\alpha)^2}{12}$$

3.5 Universality of the Uniform[1]

To briefly summary,

- When you plug a random variable into its own CDF, you get a Uniform. $F(X) \sim U$. We can apply F^{-1} to both sides to get the second result.
- When you plug a Uniform into the inverse CDF of X, you get a random variable distributed like X. $X \sim F^{-1}(U)$

[1] Brian Zhang has a video explaining this graphically here.