Elements of Probability Summer 2019

Expected values of random variables — 28.03, 2019

prof. dr hab Jacek Cichoń *Scribe: Krzysztof Agieńczuk*

1 Overview

In the last lecture we talked about geometric distribution and Bayes theorem. We also defined random variables.

In this lecture we covered: discussed coupon collector problem and expected value of random variable

2 Sum of expected values

Fact Suppose we have a family of random variables $X_1, ..., X_n$ in the same probability space. Then

$$E(X_1 + ..., X_n) = E(X_1) + ... + E(X_n) \tag{1}$$

Proof Let's assume Z for X+Y. $\Omega = \{\omega_1, ..., \omega_n\}$, $X, Y : \Omega \to \mathbb{R}$ Then

$$E(X+Y) = E(Z) = \sum_{i=1}^{k}(\omega_i) * Pr(\{\omega_i\}) = \sum_{i=1}^{k}(X(\omega_i) + Y(\omega_i)) * Pr(\{\omega_i\}) = \tag{2}$$

$$\sum_{i=1}^{k}(X(\omega_i) * Pr(\{\omega_i\}) * (Y(\omega_i) * Pr(\{\omega_i\})) = \tag{3}$$

$$\sum_{i=1}^{k}(X(\omega_i) * Pr(\{\omega_i\}) + \sum_{i=1}^{k}(Y(\omega_i) * Pr(\{\omega_i\}) = \tag{4}$$

$$E(X) + E(Y) \tag{5}$$

3 Coupon Collector Problem

Imagine we have n urns and some number of balls. We throw balls into urns at random (this means we will "use" normal distribution). By T_n we will denote number of ball after all urns are non-empty.
Let L_k denote number of steps number of steps to fill k-1 urns.

With each throw we have some probability p of filling the next urn and 1-p probability of choosing the same as previously.

$$L_K : Pr(X \notin \{1, ..., k\}) = \frac{n-(k-1)}{n} = 1 - \frac{k-1}{n} \tag{6}$$

We can clearly see that

$$L_K \sim Geo(1 - \frac{k-1}{n}) \tag{7}$$

Theorem When $X \sim Geo(p)$ then $E(X) = \frac{1}{p}$

Let's check it. In our previous problem we had $E[L_K] = \frac{n}{n-(k-1)}$ and $T_n = L_1, ..., L_n$. Then

$$E[T_n] = \sum_{k=1}^{n} E[L_K] = \sum_{k=1}^{n} \frac{n}{n-(k-1)} = n \sum_{k=1}^{n} \frac{1}{n-(k-1)} = \tag{8}$$

$$n \sum_{l=1}^{n} \frac{1}{l} \tag{9}$$

where $l = n - (k-1)$.

Definition 3.1. *n-th harmonic number is*

$$H_n = \sum_{k=1}^{n} \frac{1}{k} \tag{10}$$

Which helps us with previous example as $E[T_n] = n * H_n$ and $H_n \sim \int_1^n \frac{1}{x} dx = \ln(x)|_1^n = \ln(n)$.

So we can say

$$E[T_n] \sim n \ln(n) \tag{11}$$

3.1 Sum up

To sum up - let's suppose n is big. We are throwing balls into urns. If there are n urns, before $\sqrt{2n}$ throws we expect no collisions. This is connected to Birthday Paradox we talked previously about. However after point $n \ln(n)$ we expect to have collisions with very big probability (ie. every new ball causes collision).
This is essentially what we call Map Reduce.

Example $\Omega = [0, 1]$, $X = [0, 1] \to \mathbb{R}$
$E(X) = \int_0^1 X(t)dt$.
If $X(t) = t$, then

$$E(X) = \int_0^1 t dt = \frac{1}{2} t^2|_0^1 = \frac{1}{2} \tag{12}$$

4 Distribution function

Definition 4.1. *If X is a random variable, then a cumulative distribution function of X is*

$$F_x(t) = Pr(X \leq t) \tag{13}$$

Basic properties of F_x

- $\lim_{t \to -\infty} F_x(t) = 0$
- $\lim_{t \to \infty} F_x(t) = 1$
- $\forall a \in \mathbb{R}(\lim_{t \to a^+} F_x(t) = F_x(a))$

This last property means that this function is right continuous.

Fact $P(X \in [a, b]) = F_x(b) - F_x(a)$ where $a \leq b$.

Proof Let's define $A : \{\omega \in \Omega : X(\omega) \leq a\}$ and $B : \{\omega \in \Omega : X(\omega) \leq b\}$
$\omega \in A \iff X(\omega) \leq a \to X(\omega) \leq b \iff \omega \in B$ for $a \leq b$.
$F_x(b) = Pr(B)$ and $F_x(a) = Pr(A)$.[1]

4.1 Density of random variable

Let's say we have a random variable $X \geq 0$; $X_n \sim X$; $X_n = \frac{\lfloor n * X \rfloor}{n}$ and $(\forall \omega)(|X_n - X| \leq \frac{1}{n})$. From definition of floor function we also know that $X - \frac{1}{n} \leq X_n \leq X$
Then

$$E(X_n) = \sum_{k=0} \frac{k}{n} Pr(X_n = \frac{k}{n}) = \sum_{k=0} \frac{k}{n} Pr(\frac{k-1}{n} \leq X \leq \frac{k}{n}) = \tag{14}$$

$$\sum_{k=0} \frac{k}{n} (F_X(\frac{k}{n}) - F_X(\frac{k-1}{n})) = \sum_{k=0} (\frac{x * F_X(\frac{k}{n}) - F_X(\frac{k-1}{n})}{\frac{1}{n}} * \frac{1}{n}) = \tag{15}$$

$$\sum_{k=0} \frac{k}{n} F'_x(\frac{k}{n}) * \frac{1}{n} = \sum_{k=0} \phi(\frac{k}{n}) * \frac{1}{n} \sim \int_0^\infty \phi(x) dx \tag{16}$$

In the last line substitution is $\phi(x) = x F'_x(x)$.

Then

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{17}$$

where $f_x(t) = F'_x(t)$ is a density of X.

[1] I may have skipped something. Personally I think I understand this fact, however I cannot get my head around this proof but want to push notes to git for you. I'll welcome pull request here.

Example Imagine we are throwing darts at the board at random. If X is a random point from B (B is field of a dart board), and $Y = \|X\|$ then

$$F_Y(r) = \frac{\pi r^2}{x * 1^2} \tag{18}$$

for $r \in [0, 1]$, 0 for $r < 0$ and 1 for $r > 1$. In such case

$$f_Y(r) = 2r \tag{19}$$

in the same boundaries. Then

$$E(Y) = \int_{-\infty}^{\infty} r f_Y(r) dr = \int_{1}^{1} r 2r dr = 2 * \frac{1}{3} = \frac{2}{3} \tag{20}$$

This is called **Curse of high dimensionality**. If you define such dart board in $\mathbb{R}^{100}$, you'll get 0.99.

In general when $dim = n$, then $c_n r^n = \frac{1}{2} c_n * 1^n$, so $r^n = \sqrt[n]{\frac{1}{2}}$, which can be approximated to

$$1 - \frac{\ln 2}{n} \tag{21}$$