Quick Review of Probability

Anil Maheshwari

School of Computer Science
Carleton University
Canada
Sample Space & Events

Random Variable

Geometric Distribution

Coupon Collector Problem
Sample Space & Events
## Definitions

Sample Space $S = \text{Set of Outcomes}$.

Events $\mathcal{E} = \text{Subsets of } S$.

Probability is a function from subsets $A \subseteq S$ to positive real numbers between $[0, 1]$ such that:

1. $Pr(S) = 1$
2. For all $A, B \subseteq S$ if $A \cap B = \emptyset$, $Pr(A \cup B) = Pr(A) + Pr(B)$.
3. If $A \subset B \subseteq S$, $Pr(A) \leq Pr(B)$.
4. Probability of complement of $A$, $Pr(\overline{A}) = 1 - Pr(A)$.
Basic Definition

Examples:

1. Flipping a fair coin:
   \[ S = \{ H, T \}; \]
   \[ \mathcal{E} = \{ \emptyset, \{ H \}, \{ T \}, S \} = \{ H, T \} \]

2. Flipping fair coin twice:
   \[ S = \{ HH, HT, TH, TT \}; \]
   \[ \mathcal{E} = \{ \emptyset, \{ HH \}, \{ HT \}, \{ TH \}, \{ TT \}, \]
   \[ \{ HH, TT \}, \{ HH, TH \}, \{ HH, HT \}, \]
   \[ \{ HT, TH \}, \{ HT, TT \}, \{ TH, TT \}, \]
   \[ \{ HH, HT, TH \}, \{ HH, HT, TT \}, \{ HH, TH, TT \}, \]
   \[ \{ HT, TH, TT \}, S = \{ HH, HT, TH, TT \} \]

3. Rolling fair die twice:
   \[ S = \{(i, j) : 1 \leq i, j \leq 6\}; \]
   \[ \mathcal{E} = \{ \emptyset, \{1, 1\}, \{1, 2\}, \ldots, S\} \]
Random Variable
Definition

A random variable $X$ is a function from sample space $S$ to Real numbers, $X : S \rightarrow \mathbb{R}$.

Expected value of a discrete random variable $X$ is given by

$$E[X] = \sum_{s \in S} X(s) \ast Pr(X = X(s)).$$

Note: It's a misnomer to say $X$ is a random variable, it's a function.

Example: Flip a fair coin and define the random variable $X : \{H, T\} \rightarrow \mathbb{R}$ as

$$X = \begin{cases} 
1 & \text{Outcome is Heads} \\
0 & \text{Outcome is Tails}
\end{cases}$$

$$E[X] = \sum_{s \in \{H,T\}} X(s) \ast Pr(X = X(s)) = 1 \ast \frac{1}{2} + 0 \ast \frac{1}{2} = \frac{1}{2}$$
**Linearity of Expectation**

**Definition**
Consider two random variables $X, Y$ such that $X, Y : S \to \mathbb{R}$, then 
$$E[X + Y] = E[X] + E[Y].$$
In general, consider $n$ random variables $X_1, X_2, \ldots, X_n$ such that 
$X_i : S \to \mathbb{R}$, then 
$$E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i].$$

**Example:** Flip a fair coin $n$ times and define $n$ random variable $X_1, \ldots, X_n$ as 
$$X_i = \begin{cases} 
1 & \text{Outcome is Heads} \\
0 & \text{Outcome is Tails} 
\end{cases}$$

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n] = \frac{1}{2} + \cdots + \frac{1}{2} = \frac{n}{2}$$
$$= \text{Expected # of Heads in } n \text{ tosses.}$$
Geometric Distribution
Geometric Distribution

**Definition**

Perform a sequence of independent trials till the first success. Each trial succeeds with probability $p$ (and fails with probability $1 - p$). A Geometric Random Variable $X$ with parameter $p$ is defined to be equal to $n \in \mathbb{N}$ if the first $n - 1$ trials are failures and the $n$-th trial is success. Probability distribution function of $X$ is $Pr(X = n) = (1 - p)^{n-1}p$.

Let $Z$ to be the r.v. that equals the # failures before the first success, i.e. $Z = X - 1$.

**Problem:** Evaluate $E[X]$ and $E[Z]$.

**To show:** $E[Z] = \frac{1-p}{p}$ and $E[X] = 1 + \frac{1-p}{p} = \frac{1}{p}$. 
Computation of $E[Z]$

$Z = \#$ failures before the first success.

Set $q = 1 - p$.

- $Pr(Z = k) = q^k p$
- $\frac{1}{1-q} = \sum_{k=0}^{\infty} q^k$ (for $0 < q < 1$)
- $\frac{1}{(1-q)^2} = \sum_{k=0}^{\infty} kq^{k-1}$

\[
E[Z] = \sum_{k=0}^{\infty} kPr(Z = k)
= \sum_{k=0}^{\infty} kq^k p
= pq \sum_{k=0}^{\infty} kq^{k-1}
= pq \frac{1}{(1-q)^2}
= \frac{1-p}{p}
\]
Examples:

1. Flipping a fair coin till we get a Head:
   
   \[ p = \frac{1}{2} \text{ and } E[X] = \frac{1}{p} = 2 \]

2. Roll a die till we see a 6:
   
   \[ p = \frac{1}{6} \text{ and } E[X] = \frac{1}{p} = 6 \]

3. Keep buying LottoMax tickets till we win (assuming we have 1 in 33294800 chance).
   
   \[ p = \frac{1}{33294800} \text{ and } E[X] = \frac{1}{p} = 33,294,800. \]
Coupon Collector Problem
Coupon’s Collector Problem

Problem Definition
There are a total of \( n \) different types of coupons (Pokemon cards). A cereal manufacturer has ensured that each cereal box contains a coupon. Probability that a box contains any particular type of coupon is \( \frac{1}{n} \). What is the expected number of boxes we need to buy to collect all the \( n \) coupons?

Define r.v. \( N_1, N_2, \ldots, N_n \), where \( N_i \) = the number of boxes bought till the \( i \)-th coupon is collected.
Each \( N_i \) is a geometric random variable.
Let $N = \sum_{j=1}^{n} N_i$; Note $N_1 = 1$

$$E[N_j] = \frac{1}{\text{Pr of success in finding the } j^{th} \text{ coupon}} = \frac{1}{n - j + 1}$$

$$E[N] = \sum_{j=1}^{n} \frac{n}{n - j + 1} = nH_n, \text{ where } H_n = n\text{-th Harmonic Number.}$$

$$H_n = \sum_{i=1}^{n} \frac{1}{i} \text{ and } \ln n \leq H_n \leq \ln n + 1.$$ 

Thus, $E[N] = nH_n \approx n\ln n,$
Is \( E[N] = nH_n = n \ln n \) a good estimate?

What is the probability that \( E[N] \) exceeds \( 2nH_n \)? Applying Markov’s Inequality:

\[
Pr(X > s) \leq \frac{E[X]}{s} \quad Pr(N > 2nH_n) < \frac{E[N]}{2nH_n} = \frac{nH_n}{2nH_n} = \frac{1}{2}
\]

Can we have a better bound?

Next: We show \( Pr(N > n \ln n + cn) < \frac{1}{e^c} \)

Pr. of missing a coupon after \( n \ln n + cn \) boxes have been bought

\[
= (1 - \frac{1}{n})^{n \ln n + nc} \leq e^{-\frac{1}{n} (n \ln n + cn)} = \frac{1}{ne^c}.
\]

Pr. of missing at least one coupon \( \leq n(\frac{1}{ne^c}) = \frac{1}{e^c} \).
1. Introduction to Probability by Blitzstein and Hwang, CRC Press 2015.

2. Courses Notes of COMP 2804 by Michiel Smid.
