Quick Review of Probability

Anil Maheshwari

School of Computer Science
Carleton University
Canada

Outline

Sample Space & Events

Random Variable

Geometric Distribution

Coupon Collector Problem

Balls and Bins

- Collisions
- Size of Bins

Sample Space & Events

Basic Definition

Definitions

Sample Space S = Set of Outcomes.

Events $\mathcal{E}$ = 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, $Pr(\bar{A}) = 1 - Pr(A)$.

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:
   $S = \{(i, j) : 1 \leq i, j \leq 6\}$;
   $\mathcal{E} = \{\emptyset, \{1, 1\}, \{1, 2\}, \dots, S\}$

Random Variable

Expectation

Definition

A random variable X is a function from sample space S to Real numbers, $X : S \rightarrow \Re$.
Expected value of a discrete random variable X is given by $E[X] = \sum_{s \in S} X(s) * Pr(X = X(s))$.

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

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

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

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

Linearity of Expectation

Definition

Consider two random variables X, Y such that $X, Y : S \to \Re$, then $E[X+Y] = E[X] + E[Y]$.
In general, consider n random variables $X_1, X_2, \ldots, X_n$ such that $X_i : S \to \Re$, 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

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

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

Proof of Linearity of Expectation

$$
\begin{aligned}
E[X+Y] &= \sum_{\omega \in S} (X+Y)[\omega] \cdot P(\omega) \\
&= \sum_{\omega \in S} (X[\omega] + Y[\omega]) \cdot P(\omega) \\
&= \sum_{\omega \in S} (X[\omega] \cdot P(\omega) + Y[\omega] \cdot P(\omega)) \\
&= \sum_{\omega \in S} X[\omega] \cdot P(\omega) + \sum_{\omega \in S} Y[\omega] \cdot P(\omega) \\
&= E[X] + E[Y]
\end{aligned}
$$

This generalizes to the sum of n random variables:
$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n]$.

More detailed proof

$$
\begin{aligned}
E[X+Y] &= \sum_x \sum_y (x+y)P(X=x, Y=y) \\
&= \sum_x \sum_y xP(X=x, Y=y) + \sum_x \sum_y yP(X=x, Y=y) \\
&= \sum_x \sum_y xP(X=x, Y=y) + \sum_y \sum_x yP(X=x, Y=y) \\
&= \sum_x x \sum_y P(X=x, Y=y) + \sum_y y \sum_x P(X=x, Y=y) \\
&= \sum_x xP(X=x) + \sum_y yP(Y=y) \\
&= E[X] + E[Y]
\end{aligned}
$$

Geometric Distribution

Geometric Distribuition

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 $n \in N$ 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 $0 < q < 1$)
- $\frac{1}{(1-q)^2} = \sum_{k=0}^{\infty} kq^{k-1}$

$$
\begin{aligned}
E[Z] &= \sum_{k=0}^{\infty} kPr(Z = k) = \sum_{k=0}^{\infty} kq^k p \\
&= pq \sum_{k=0}^{\infty} kq^{k-1} \\
&= \frac{pq}{(1-q)^2} = \frac{1-p}{p}
\end{aligned}
$$

Examples

Examples:

1. Flipping a fair coin till we get a Head:
   $p = \frac{1}{2}$ and $E[X] = \frac{1}{p} = 2$
2. Roll a die till we see a 6:
   $p = \frac{1}{6}$ and $E[X] = \frac{1}{p} = 6$
3. Keep buying LottoMax tickets till we win (assuming we have 1 in 33294800 chance).
   $p = \frac{1}{33294800}$ 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 $N_i =$# of boxes bought till the i-th coupon is collected.

Each N_i is a geometric random variable.

Coupon's Collector Problem Contd.

Let $N = \sum_{j=1}^{n} N_i$;

Note $N_1 = 1$

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

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

$H_n = \sum_{i=1}^{n} \frac{1}{i}$ 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 $nH_n + cn$, for some constant c?

Claim

$Pr(N > n \ln n + cn) < \frac{1}{e^c}$

Proof Sketch: 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 $\leq n(\frac{1}{ne^c}) = \frac{1}{e^c}$ $\square$

E.g., For $c = 7$, $\frac{1}{e^7} < 0.001$

Balls and Bins

Balls & Bins

Model

We have m Balls and n Bins. We throw each ball in a bin uniformly at random.

What is the probability of following events:

1. Expected number of balls in a bin.
2. Balls i and j are in the same bin.
3. Expected number of empty bins.
4. Bin $\#i$ receives (a) 0 balls, (b) k balls, and (c) $\geq k$ balls.
5. All bins have $\leq \frac{c \ln n}{\ln \ln n}$ balls.

Applications: Birthday Paradox, Load Balancing, Perfect Hashing

Expected number of balls in bin b

Number of Balls = m
Number of Bins = n.

Define random variables $Y_i = \begin{cases} 1, & \text{if } i \text{ th ball is in bin } b \\ 0, & \text{otherwise} \end{cases}$

$$
\begin{aligned}
E[\#\text{ Balls in Bin } b] &= E\left[\sum_{i=1}^{n} Y_i\right] \\
&= \sum_{i=1}^{m} E[Y_i] \qquad \text{(Linearity of Expectation)} \\
&= \sum_{i=1}^{m} P(\text{Ball } i \text{ is in Bin } b) \\
&= \sum_{i=1}^{m} 1/n = m/n
\end{aligned}
$$

Probability of ball j is in the bin containing ball i

Number of Balls = m
Number of Bins = n.

$Pr[\text{Balls } i \text{ and } j \text{ in same bin}] = \frac{1}{n}$.

Define r.v. X_{ij} ($1 \leq i \leq m-1$, $i+1 \leq j \leq m$) as follows:

$$X_{ij} = \begin{cases} 1 & \text{if ball } j \text{ is in the Bin of ball } i \\ 0 & \text{Otherwise} \end{cases}$$

$Pr(X_{ij} = 1) = \frac{1}{n}$.
$E[X_{ij}] = \frac{1}{n}$.

Expected number of collisions

Number of Balls = m
Number of Bins = n.

Define $X = \sum_{i,j} X_{ij}$ = Total # of collisions

$E[X] = E[\sum_{i,j} X_{ij}]$

By Linearity of Expectation: $E[\sum_{i,j} X_{ij}] = \sum_{i,j} E[X_{ij}]$

Thus $E[X] = \frac{1}{n}\binom{m}{2}$

Birthday Paradox

Number of Balls = m = Number of Students
Number of Bins = n = Number of days in a Year.

For two students to have same Birthday:
What value of m will result in $E[X] = \frac{1}{n}\binom{m}{2} \geq 1$

For $m \geq 28$, $E[X] = \frac{1}{365}\frac{m(m-1)}{2} \geq 1$

Birthday Paradox Contd.

What is minimum value of m so that the probability that two students share the same birthday is $\geq \frac{1}{2}$?

Probability that all m students have distinct birthday's is given by: $(1-\frac{1}{n})(1-\frac{2}{n})(1-\frac{3}{n})\ldots(1-\frac{m-1}{n})$

Let us use the inequality $1-x \leq e^{-x}$.

We want:

$$\begin{aligned} e^{-\frac{1}{n}}e^{-\frac{2}{n}}e^{-\frac{3}{n}}\ldots e^{-\frac{m-1}{n}} &\leq \frac{1}{2} \\ e^{-\frac{m(m-1)}{2n}} &\leq \frac{1}{2} \end{aligned}$$

Now using $n = 365$, we have $e^{-\frac{m(m-1)}{2*365}} \leq \frac{1}{2}$ or $m \geq 23$.

Intuition: There are $m^2 \approx n$ pairs where a collision can occur.

Number of Balls in Bin i

Number of Balls = m; Number of Bins = n.

Problem I

What is the probability that Bin i receives no balls?

$$\left(1 - \frac{1}{n}\right)^m \leq e^{-\frac{m}{n}}$$

If $n = m$, $(1 - \frac{1}{n})^n \leq e^{-1} = 0.37$.

Problem II

What is the probability that Bin i receives exactly k balls?

$$\binom{m}{k} \left(\frac{1}{n}\right)^k \left(1 - \frac{1}{n}\right)^{m-k}$$

Number of Balls in Bin i contd.

Number of Balls = m; Number of Bins = n.

Problem III

What is the probability that Bin i receives $\geq k$ balls?

$$\leq \binom{m}{k}\left(\frac{1}{n}\right)^k$$

Fix a set S of k balls.

Pr(Bin i receives all balls in S)$= (\frac{1}{n})^k$.

The remaining balls can go anywhere.

Take the union bound over all possible subset of k balls.

Useful Bounds: $(\frac{n}{k})^k < \binom{n}{k} < \frac{n^k}{k!} < (\frac{en}{k})^k$.
If $n = m$, we have $\binom{n}{k}\left(\frac{1}{n}\right)^k \leq \left(\frac{e}{k}\right)^k$

Expected Number Empty Bins

Number of Balls = m; Number of Bins = n.

Problem IV

What is Expected # of Empty Bins?

Define a r.v. X_i such that

$$X_i = \begin{cases} 1 & \text{if Bin } i \text{ is empty} \\ 0 & \text{Otherwise} \end{cases}$$

From Problem I, $Pr(X_i = 1) \leq e^{-\frac{m}{n}}$ and $E[X_i] \leq e^{-\frac{m}{n}}$

Thus, $E[\text{\# of Empty Bins}] = \sum_{i=1}^{n} E[X_i] \leq ne^{-\frac{m}{n}}$

When $n = m$, $E[\text{\# of Empty Bins}] \leq \frac{n}{e}$

Max # Balls in Bins

Number of Balls = Number of Bins = n.

Max # of Balls in Bins

With probability $\geq 1 - \frac{1}{n}$ all bins receive fewer than $3\frac{\ln n}{\ln \ln n}$ balls.

$Pr(\text{Bin } i \text{ has more that } k \text{ balls}) \leq \binom{m}{k}\left(\frac{1}{n}\right)^k = \binom{n}{k}\left(\frac{1}{n}\right)^k \leq \left(\frac{e}{k}\right)^k$

Substitute $k = 3\frac{\ln n}{\ln \ln n}$ and show that

$Pr(\text{Bin } i \text{ has more that } k \text{ balls}) \leq \frac{1}{n^2}$

Thus by Union Bound,

$Pr(\text{Any bin has more that } k \text{ balls}) \leq \frac{1}{n}$

Max # Balls in Bins Contd.

Claim

For $k = 3\frac{\ln n}{\ln \ln n}$, $\left(\frac{e}{k}\right)^k \leq \frac{1}{n^2}$

Proof.

$$
\begin{aligned}
\left(\frac{e}{k}\right)^k &= \left[e\frac{\ln\ln n}{3\ln n}\right]^{\frac{3\ln n}{\ln\ln n}} \\
&= \left[e^1 e^{\ln \frac{\ln\ln n}{3\ln n}}\right]^{\frac{3\ln n}{\ln\ln n}} \\
&= e^{\frac{3\ln n}{\ln\ln n}[1+\ln\ln\ln n-\ln(3\ln n)]} \\
&= e^{\frac{3\ln n}{\ln\ln n}[1+\ln\ln\ln n-\ln 3-\ln\ln n]} \\
&\leq e^{\frac{3\ln n}{\ln\ln n}[\ln\ln\ln n-\ln\ln n]} = e^{\left[-3\ln n+\frac{3\ln n\ln\ln\ln n}{\ln\ln n}\right]}
\end{aligned}
$$

For large values of n, $-3\ln n + \frac{3\ln n\ln\ln\ln n}{\ln\ln n} \leq -2\ln n$.

Thus, $\left(\frac{e}{k}\right)^k \leq e^{-2\ln n} = \frac{1}{n^2}$ $\square$

References

1. Introduction to Probability by Blitzstein and Hwang, CRC Press 2015.
2. Courses Notes of COMP 2804 by Michiel Smid.
3. Probability and Computing by Mitzenmacher and Upfal, Cambridge Univ. Press 2005.
4. My Notes on Algorithm Design.