

CS174 Fall 98: Lecture Note 1

Alistair Sinclair, August 1998; based on earlier notes by Manuel Blum/Douglas Young.

Basic definitions

- A (statistical) experiment: *e.g.*, *toss a die*.
- A sample space \mathcal{S} is the set of possible outcomes of the experiment, known as sample points, each with its associated probability: *e.g.*, *the numbers 1, 2, ..., 6, each with probability $\frac{1}{6}$* .
- An event $E \subseteq \mathcal{S}$ is a set of sample points: *e.g.*, *the event that the roll is even is the set $E = \{2, 4, 6\}$* .
- A random variable (r.v.) $X : \mathcal{S} \rightarrow \mathbf{R}$ is a function from the sample space to the reals: *e.g.*, *$X = 0$ if the roll of the die is 5 or 6, and $X = 1$ otherwise*.

Example 0

Experiment: toss two (labeled) dice

Sample space: $6^2 = 36$ sample points, each with probability $\frac{1}{36}$

An event: $E =$ the sum of the pips is bigger than 10 (i.e, the set of sample points $\{(5, 6), (6, 5), (6, 6)\}$)

Another event: $F =$ the second number is larger than the first

A random variable: $X =$ the sum of the pips on the dice

Another r.v.: $Y = \begin{cases} 1 & \text{if sum of pips} = 7; \\ 0 & \text{otherwise.} \end{cases}$

Note: the 0-1-valued r.v. Y is called the indicator r.v. of the event “sum of pips = 7.”

We can calculate the probabilities of events: e.g., in Example 0 we have

$$\Pr[E] = \frac{3}{6^2} = \frac{1}{12} \quad \text{and} \quad \Pr[X = 2] = \frac{1}{6^2} = \frac{1}{36}.$$

(Note that for any r.v. X and any real value x , “ $X = x$ ” is always an event: namely, the set of sample points s for which $X(s) = x$.)

Usually it is much easier to calculate *expectations* than probabilities.

The expected value (or average value or mean) of a r.v. X is defined to be

$$E(X) = \sum_k \Pr[X = k] \cdot k.$$

(In this sum, k need not be an integer; any discrete set of values is allowed. Note also that $E(X) = \infty$ is possible.)

In Example 0 we have

$$\begin{aligned} E(X) &= \sum_{k=2}^{12} \Pr[X = k] \cdot k = \sum_{k=2}^{12} \left(\sum_{i+j=k} \Pr[(i, j)] \right) \cdot k \\ &= \frac{1}{6^2} \cdot 2 + \frac{2}{6^2} \cdot 3 + \frac{3}{6^2} \cdot 4 + \cdots + \frac{3}{6^2} \cdot 10 + \frac{2}{6^2} \cdot 11 + \frac{1}{6^2} \cdot 12 \\ &= \frac{252}{36} = 7, \end{aligned}$$

and

$$E(Y) = \Pr[Y = 1] \cdot 1 = \frac{6}{6^2} = \frac{1}{6}.$$

Ex: In the case of r.v. X above, $E(X) = 7$ happens to be the most probable value of X . Give examples of simple r.v.'s U, V on the sample space of Example 0 such that:

- $E(U)$ is not a possible value of U (i.e., $\Pr[U = E(U)] = 0$);
- $E(V)$ is a possible but improbable value of V (i.e., $\Pr[V = E(V)]$ is positive but small). \square

Example 1: balls and bins (see Note 0)

Experiment: toss m (labeled) balls sequentially into n (labeled) bins.

Sample space: n^m sample points (n possible bins for ball 1, n for ball 2 etc.), each with prob. $\frac{1}{n^m}$.

Random variable: $X =$ number of empty bins.

Ex: In the case $m = 2, n = 3$, list all $3^2 = 9$ sample points. Check that $E(X) = \frac{3}{9} \cdot 2 + \frac{6}{9} \cdot 1 = \frac{12}{9} = \frac{4}{3}$. So, *on average*, we expect 1.333... empty bins. \square

More generally, define n indicator random variables $X_i = \begin{cases} 1 & \text{if bin } i \text{ is empty;} \\ 0 & \text{otherwise.} \end{cases}$

Then $X = \sum_{i=1}^n X_i$.

Theorem: $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$ for any family of r.v.'s $\{X_i\}$. \square

Ex: Prove the theorem! \square

Now

$$E(X_i) = \Pr[X_i = 1] \cdot 1 = \Pr[X_i = 1] = (n-1)^m \times \frac{1}{n^m} = \left(1 - \frac{1}{n}\right)^m,$$

so

$$E(X) = n \left(1 - \frac{1}{n}\right)^m. \quad (*)$$

(In the above we used the fact that $\Pr[\text{bin } i \text{ is empty}] = (n-1)^m \times \frac{1}{n^m}$. Check you understand this by counting how many sample points have bin i empty.)

Let's look first at the special case $m = n$ (i.e., n balls, n bins)

Then $\left(1 - \frac{1}{n}\right)^n = \left(1 - \frac{1}{n}\right)^n \rightarrow \frac{1}{e}$ as $n \rightarrow \infty$, so $E(X_i) \rightarrow \frac{1}{e}$.

Hence $E(X) = nE(X_i) \sim \frac{n}{e}$ as $n \rightarrow \infty$. I.e., *the expected number of empty bins approaches $\frac{n}{e}$ as n gets large.*

Notation: For functions f, g , we write $f(n) \sim g(n)$ to denote $\frac{f(n)}{g(n)} \rightarrow 1$ as $n \rightarrow \infty$.

In the above calculation, we used the following standard fact:

Fact 1: If $t^2/n \rightarrow 0$ as $n \rightarrow \infty$, then $\left(1 + \frac{t}{n}\right)^n \sim e^t$ as $n \rightarrow \infty$ (and actually $\left(1 + \frac{t}{n}\right)^n \leq e^t \quad \forall n$). \square

Ex: Compute $E(X)$ exactly for $m = n = 1, 2, 4, 8, 16, \dots$. When does the value $\frac{n}{e}$ become a good estimate of $E(X)$? \square

Ex: What is $E(X)$ in the case $m = 2n$? \square

Ex: What is the expected number of bins containing *exactly one* ball in the case $m = n$? \square

Now let's look at other values of m .

What if $m = n \ln n$?

From (*), we have $E(X) = n(1 - \frac{1}{n})^{n \ln n} \sim ne^{-\ln n} = 1$. (Check this using Fact 1.)

And similarly, if $m = n \ln n + cn$, then $E(X) \sim e^{-c}$.

Note that this essentially solves the coupon collecting problem: about $n \ln n$ cereal boxes are enough. To see this, note that $\Pr[X \geq 1] \leq E(X)$ (why??), so $\Pr[\text{more than } n \ln n + cn \text{ boxes needed}] \leq e^{-c}$ for any c . (Note that we've used a bound on *expectation* to get a bound on *probability*. This is an example of Markov's inequality (see later).)

What if $m = \sqrt{n}$?

Again from (*) and Fact 1 we have

$$E(X) = n \left(1 - \frac{1}{n}\right)^{\sqrt{n}} \sim ne^{-1/\sqrt{n}} = n \left(1 - \frac{1}{\sqrt{n}} + \frac{1}{2!n} - \frac{1}{3!n^{3/2}} + \dots\right) = n - \sqrt{n} + \frac{1}{2} - \dots$$

But $n - \sqrt{n} = \#$ empty bins if no collisions. (A "collision" is a ball that is not the first to arrive in its bin.) Hence $E(\# \text{ collisions}) = E(X - (n - \sqrt{n})) = \frac{1}{2} - \frac{1}{3!\sqrt{n}} + \dots \rightarrow \frac{1}{2}$ as $n \rightarrow \infty$.

So, when we throw $m = \sqrt{n}$ balls into n bins, the probability of a collision is $\leq \frac{1}{2}$ (as $n \rightarrow \infty$).

More generally, if we take $m = c\sqrt{n}$ then

$$E(X) = n \left(1 - \frac{1}{n}\right)^{c\sqrt{n}} \sim ne^{-c/\sqrt{n}} = n \left(1 - \frac{c}{\sqrt{n}} + \frac{c^2}{2!n} - \frac{c^3}{3!n^{3/2}} + \dots\right) = n - c\sqrt{n} + \frac{c^2}{2} - \dots$$

So in this case we expect about $\frac{c^2}{2}$ collisions; and consequently, the probability of *any* collision occurring is at most $\frac{c^2}{2}$. (Why?)

Ex: The above is a generalization of the famous *birthday problem*: "if there are m people at a party, what is the probability that at least two of them have the same birthday?" For $n = 365$, how large does m have to be before $E(\# \text{ collisions}) \geq 1$? How large does it have to be before $\Pr[\text{a collision occurs}] \geq \frac{1}{2}$? \square

How many balls in the fullest bin?

Define the r.v. $Y_k = \#$ bins with load $\geq k$.

Claim: $E(Y_k) \leq n \left(\frac{e}{k}\right)^k$ as $k, n \rightarrow \infty$.

Proof of Claim: Let's write $Y_k = \sum_{i=1}^n X_{k,i}$, where $X_{k,i} = \begin{cases} 1 & \text{if bin } i \text{ contains } \geq k \text{ balls;} \\ 0 & \text{otherwise.} \end{cases}$

Then $E(Y_k) = \sum_{i=1}^n E(X_{k,i}) = nE(X_{k,i})$ (since clearly $E(X_{k,i})$ is the same for all i).

But we have

$$\begin{aligned} E(X_{k,i}) &= \Pr[\text{bin } i \text{ has } \geq k \text{ balls}] = \sum_{j=k}^n \Pr[\text{bin } i \text{ has } j \text{ balls}] \\ &= \sum_j \binom{n}{j} \left(\frac{1}{n}\right)^j \left(1 - \frac{1}{n}\right)^{n-j} \leq \sum_j \binom{n}{j} \left(\frac{1}{n}\right)^j \\ &\leq \sum_j \left(\frac{ne}{j}\right)^j \left(\frac{1}{n}\right)^j = \sum_j \left(\frac{e}{j}\right)^j \\ &\leq \left(\frac{e}{k}\right)^k \left(1 + \frac{e}{k} + \left(\frac{e}{k}\right)^2 + \dots\right) = \left(\frac{e}{k}\right)^k \frac{1}{1-e/k} \sim \left(\frac{e}{k}\right)^k, \end{aligned}$$

Since $E(Y_k) = nE(X_{k,i})$, this completes the proof. \square

Ex: Verify the following fact, which we used in line 2 of the above calculation:

$$\Pr[\text{bin } i \text{ has exactly } j \text{ balls}] = \binom{n}{j} (n-1)^{n-j} \times \frac{1}{n^n} = \binom{n}{j} \left(\frac{1}{n}\right)^j \left(1 - \frac{1}{n}\right)^{n-j}. \quad \square$$

Ex: In line 3 of the above calculation, we used another standard fact which you might like to check:

Fact 2: $\left(\frac{n}{k}\right)^k \leq \binom{n}{k} \leq \left(\frac{ne}{k}\right)^k$. \square

Now let's use the Claim to get a bound on the load in the fullest bin. The function $n \left(\frac{e}{k}\right)^k$ decreases with k . How large does k have to be (as a function of n) before it drops below 1?

If you plug in the value $k = \frac{(1+\epsilon)\ln n}{\ln \ln n}$ and take logs, you'll find that

$$\ln \left(n \left(\frac{e}{k} \right)^k \right) = \ln n + k(1 - \ln k) = (-\epsilon + f(n)) \ln n,$$

where $f(n) \rightarrow 0$ as $n \rightarrow \infty$. (Check this!) So for any $\epsilon > 0$, we have that $n \left(\frac{e}{k}\right)^k \rightarrow 0$ as $n \rightarrow \infty$. This means that we should not expect any bin to have more than $\frac{(1+\epsilon)\ln n}{\ln \ln n}$ balls, for any $\epsilon > 0$.

Ex: Show that $\Pr[\text{any bin contains more than } \frac{(1+\epsilon)\ln n}{\ln \ln n} \text{ balls}] \rightarrow 0$ as $n \rightarrow \infty$, for any $\epsilon > 0$. Can you say anything when $\epsilon = 0$? \square

We should go back and prove the fundamental Theorem stated earlier, namely

Theorem: Let $\{X_i\}_{i=1}^n$ be any collection of r.v.'s on sample spaces \mathcal{S}_i . Then $E(\sum_{i=1}^n X_i) = \sum_{i=1}^n E(X_i)$.

Proof: We will show, for two r.v.'s X, Y , that $E(X + Y) = E(X) + E(Y)$. This suffices to prove the theorem. Now

$$\begin{aligned} E(X + Y) &= \sum_i \sum_j \Pr[(X = i) \wedge (Y = j)] \cdot (i + j) \\ &= \sum_i \sum_j \left\{ \Pr[(X = i) \wedge (Y = j)] \cdot i + \Pr[(X = i) \wedge (Y = j)] \cdot j \right\} \\ &= \left\{ \sum_i i \cdot \sum_j \Pr[(X = i) \wedge (Y = j)] \right\} + \left\{ \sum_j j \cdot \sum_i \Pr[(X = i) \wedge (Y = j)] \right\} \\ &= \sum_i i \cdot \Pr[X = i] + \sum_j j \cdot \Pr[Y = j] \\ &= E(X) + E(Y). \quad \square \end{aligned}$$

Note: This result is surprising because it holds even if X and Y are *dependent*. By contrast, it is **not** generally true that $E(XY) = E(X)E(Y)$, unless X and Y are *independent*.

Definition: Two events E, F are independent if $\Pr[E \wedge F] = \Pr[E]\Pr[F]$. Two r.v.'s X, Y are independent if the events $X = i$ and $Y = j$ are independent for all i and j .

Ex: Prove that $E(XY) = E(X)E(Y)$ if X, Y are independent. \square

Ex: Consider again the sample space of Example 0 (tossing two dice). Define the r.v.'s

X = score on first die; Y = score on second die; Z = maximum score on both dice.

Are X, Y independent? Are X, Z independent? Justify your answers carefully. \square

If events E, F are not independent, we can't say much about $\Pr[E \wedge F]$, except that $\Pr[E \wedge F] \leq \min\{\Pr[E], \Pr[F]\}$.

What about $\Pr[E \vee F]$? Evidently

$$\Pr[E \vee F] = \Pr[E] + \Pr[F] - \Pr[E \wedge F] \leq \Pr[E] + \Pr[F].$$

This inequality is often very useful. Note that the inequality is an equality if and only if E, F are mutually exclusive, i.e., $\Pr[E \wedge F] = 0$.