Based on earlier notes by Manuel Blum/Douglas Young/Alistair Sinclair.

Lecture Note 1

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* $1, 2, \dots 6$*, 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.) $X : \mathcal{S} \to \mathbf{R}$ 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

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

(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

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

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

In Example 0 we have

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

and

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

Ex: In the case of r.v. X above, $\mathrm{E}(X) = 7$ 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:

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

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 $\mathrm{E}(X) = \frac{3}{9}\cdot 2 + \frac{6}{9}\cdot 1 = \frac{12}{9} = \frac{4}{3}$. So, *on average*, we expect 1.333… empty bins. □

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

Ex: Prove the theorem! □

Now

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

so

$$\mathrm{E}(X) = n\left(1 - \tfrac{1}{n}\right)^m. \qquad (*)$$

(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.)

<u>Let's look first at the special case $m = n$ (i.e., n balls, n bins)</u>
Then $(1 - \frac{1}{n})^m = (1 - \frac{1}{n})^n \to \frac{1}{\mathrm{e}}$ as $n \to \infty$, so $\mathrm{E}(X_i) \to \frac{1}{\mathrm{e}}$.

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

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

In the above calculation, we used the following standard fact:
Fact 1: If $t^2/n \to 0$ as $n \to \infty$, then $(1 + \frac{t}{n})^n \sim \mathrm{e}^t$ as $n \to \infty$ (and actually $(1 + \frac{t}{n})^n \le \mathrm{e}^t \quad \forall n$). □

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

Ex: What is $\mathrm{E}(X)$ in the case $m = 2n$? □

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

Now let's look at other values of m.

<u>What if $m = n \ln n$?</u>
From $(*)$, we have $\mathrm{E}(X) = n(1 - \frac{1}{n})^{n \ln n} \sim n\mathrm{e}^{-\ln n} = 1$. (Check this using Fact 1.)
And similarly, if $m = n \ln n + cn$, then $\mathrm{E}(X) \sim \mathrm{e}^{-c}$.
Note that this essentially solves the coupon collecting problem: about $n \ln n$ cereal boxes are enough. To see this, note that $\Pr[X \ge 1] \le \mathrm{E}(X)$ (why??), so $\Pr[\text{more than } n \ln n + cn \text{ boxes needed}] \le \mathrm{e}^{-c}$ for any c. (Note that we've used a bound on *expectation* to get a bound on *probability.* This is an example of <u>Markov's inequality</u> (see later).)

<u>What if $m = \sqrt{n}$?</u>
Again from $(*)$ and Fact 1 we have

$$\mathrm{E}(X) = n\left(1 - \tfrac{1}{n}\right)^{\sqrt{n}} \sim n\mathrm{e}^{-1/\sqrt{n}} = n\left(1 - \tfrac{1}{\sqrt{n}} + \tfrac{1}{2!n} - \tfrac{1}{3!n^{3/2}} + \cdots\right) = n - \sqrt{n} + \tfrac{1}{2} - \ldots$$

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 $\mathrm{E}(\#\text{ collisions}) = \mathrm{E}(X - (n - \sqrt{n})) = \frac{1}{2} - \frac{1}{3!\sqrt{n}} + \ldots \to \frac{1}{2}$ as $n \to \infty$.

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

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

$$\mathrm{E}(X) = n\left(1 - \tfrac{1}{n}\right)^{c\sqrt{n}} \sim n\mathrm{e}^{-c/\sqrt{n}} = n\left(1 - \tfrac{c}{\sqrt{n}} + \tfrac{c^2}{2!n} - \tfrac{c^3}{3!n^{3/2}} + \cdots\right) = n - c\sqrt{n} + \tfrac{c^2}{2} - \ldots$$

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 $\mathrm{E}\,(\#\text{ collisions}) \geq 1$? 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?

For simplicity we'll consider only the case $m = n$.
Define the r.v. $Y_k = \#$ bins with load $\geq k$.

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

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

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

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

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

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

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

Fact 2: $(\frac{n}{k})^k \leq \binom{n}{k} \leq (\frac{n\mathrm{e}}{k})^k$. $\square$

Now let's use the Claim to get a bound on the load in the fullest bin. The function $n\left(\frac{\mathrm{e}}{k}\right)^k$ decreases with k. How large does k have to be (as a function of n) before it approaches 0? If you plug in the value $k = \frac{(1+\epsilon)\ln n}{\ln\ln n}$ and take logs, you'll find that

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

where $f(n) \to 0$ as $n \to \infty$. (Check this!) So for any $\epsilon > 0$, we have that $n\left(\frac{\mathrm{e}}{k}\right)^k \to 0$ as $n \to \infty$. 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 $\Pr[\text{any bin contains more than } \frac{(1+\epsilon)\ln n}{\ln\ln n} \text{ balls}] \to 0$ as $n \to \infty$, for any $\epsilon > 0$. Can you say anything when $\epsilon = 0$? $\square$

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

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

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

$$X = \text{score on first die}; \quad Y = \text{score on second die}; \quad Z = \text{maximum score on both dice}.$$

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., $\Pr[E \wedge F] = 0$.