Randomness and Computation

or, “Randomized Algorithms”

Mary Cryan

School of Informatics
University of Edinburgh

Coupon Collector Problem

> *“Coupon collecting" is the activity of buying cereal-packets, each of which will have a coupon inside. There are be n different types of “coupon" (eg cards with a photo of a foot-baller) and the goal is to collect one copy of each ... then stop buying.*

On Tuesday we showed that the expected number of purchases needed $\mathrm{E}[X]$ to collect all cards is $\sim n\ln(n)$.

Today we examine how likely a example “run" of the purchasing process is to come close to that expectation.

Results like *Markov's Inequality*, *Chebyshev's Inequality* and (Friday) *Chernoff/Hoeffding Bounds* help us show *concentration* about the mean.

Markov's Inequality

The simplest one.

Theorem (3.1, Markov's Inequality)

Let X be any random variable that takes only non-negative values. Then for any $a > 0$,

$$\Pr[X \geq a] \;\leq\; \frac{\mathrm{E}[X]}{a}.$$

Proof.

Define the indicator function $I = I(X)$ by

$$I(x) \;=\; \begin{cases} 0 & x < a \\ 1 & x \geq a \end{cases}$$

Then $X \geq a \cdot I(X)$, and hence $I(X) \leq \frac{X}{a}$.
Taking expectation of both sides, and using $\mathrm{E}[I] = \Pr[X \geq a]$, we have

$$\Pr[X \geq a] \;=\; \mathrm{E}[I] \;\leq\; \frac{1}{a}\mathrm{E}[X].$$

Variance, Moments of a Random Variable

Definition (3.1)

The *kth moment* of a random variable X is defined to be $\mathrm{E}[X^k]$.

Definition (3.2)

The *variance* of a random variable is defined to be

$$\mathrm{Var}[X] \;=_{def}\; \mathrm{E}[(X - \mathrm{E}[X])^2] \;=\; \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

The *standard deviation* of a random variable X is defined as

$$\sigma[X] \;=\; \sqrt{\mathrm{Var}[X]}.$$

(we saw why $\mathrm{E}[(X - \mathrm{E}[X])^2]$ and $\mathrm{E}[X^2] - \mathrm{E}[X]^2$ were equal on Tuesday)

Covariance of two Random Variables

Definition (3.3)

The *covariance* of two random variables X and Y is defined as

$$\mathrm{Cov}[X, Y] = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right].$$

Theorem (3.2)

For any two random variables X, Y, we have

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y].$$

Proof.

The definition of Var gives $\mathrm{Var}[X + Y] = \mathrm{E}[(X + Y)^2] - \mathrm{E}[X + Y]^2$.
By squaring, and linearity of exp., this is
$\mathrm{E}[X^2] + \mathrm{E}[Y^2] + \mathrm{E}[2XY] - (\mathrm{E}[X]^2 + \mathrm{E}[Y]^2 + 2\mathrm{E}[X]\mathrm{E}[Y])$.
This is $\mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{E}[XY] - 2\mathrm{E}[X]\mathrm{E}[Y]$.
Expanding $\mathrm{Cov}[X, Y]$, linearity of Exp., gives $2E[XY] - 2\mathrm{E}[X]\mathrm{E}[Y]$,so
$\mathrm{Var}[X + Y]$ equals $\mathrm{Var}[X] + \mathrm{Var}[Y] + 2\mathrm{Cov}[X, Y]$. □

(pairwise) Independent Random Variables

Theorem (3.3)

If X, Y are a pair of independent random variables, then

$$\mathrm{E}[XY] = \mathrm{E}[X] \cdot \mathrm{E}[Y].$$

Proof is in the book, reasoning wrt Definition 2.2 (may do on visualiser).

Corollary (3.4)

If X, Y are a pair of independent random variables, then

$$\mathrm{Cov}[X, Y] = 0$$

and

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Proof is straightforward application of Thm 3.3.

Chebyshev's Inequality

Theorem (3.2, Chebyshev's Inequality)

For every $a > 0$,

$$\Pr[|X - \mathrm{E}[X]| \geq a] \leq \frac{\mathrm{Var}[X]}{a^2}.$$

Proof.

Because the probability is of the *absolute value* of $X - \mathrm{E}[X]$, we know that for any $b > 0$, $|X - \mathrm{E}[X]| = b$ happens $\Leftrightarrow (X - \mathrm{E}[X])^2 = b^2$ happens.
So $\Pr[|X - \mathrm{E}[X]| \geq a] = \Pr[(X - \mathrm{E}[X])^2 \geq a^2]$.
Applying Markov's Ineq. to the random variable $(X - \mathrm{E}[X])^2$, we know

$$\Pr[(X - \mathrm{E}[X])^2 \geq a^2] \leq \frac{\mathrm{E}[(X - \mathrm{E}[X])^2]}{a^2}.$$

and by definition of $\mathrm{Var}(\cdot)$, this gives

$$\Pr[|X - \mathrm{E}[X]| \geq a] = \Pr[(X - \mathrm{E}[X])^2 \geq a^2] \leq \frac{\mathrm{Var}[X]}{a^2}.$$

Bounding Coupon Collector purchases - Markov

Remember X are the number of packets bought until we have all n different cards, $\mathrm{E}[X] = n\ln(n) + \Theta(n)$ is the expected number.

Consider how likely we are to need *twice* the expected number of purchases ($2\mathrm{E}[X]$).By Markov's Ineq.,

$$\Pr[X \geq 2\mathrm{E}[X]] \leq \frac{\mathrm{E}[X]}{2\mathrm{E}[X]} = \frac{1}{2}.$$

Or, if we are willing to spend $10\mathrm{E}[X]$ (ie, $10n(\ln(n) + 1)$), there is at most $1/10$ probability we fail to get all cards.

Very boring! (Markov's Ineq)

We can do much better with Chebyshev's Ineq. ...

Bounding Coupon Collector purchases - Chebyshev

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2}.$$

- Need to evaluate $\mathrm{Var}[X]$, which is $\mathrm{Var}[X_1 + \ldots X_n]$.
- Looking back at Corollary 3.4, see that for independent Y, Z, $\mathrm{Var}[Y + Z] \;=\; \mathrm{Var}[Y] + \mathrm{Var}[Z]$.
- Recall that X_i, the *number of packets* bought to get the i-th new card, is independent of the value of X_{i-1} or any of the earlier X_h values. X_i only depends on the values n and i.
- Hence the random variables $X_1, \ldots, X_n$ are all mutually independent.
- So
$$\mathrm{Var}[X] \;=\; \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + \ldots + \mathrm{Var}[X_n].$$

Bounding Coupon Collector purchases - Chebyshev

Each X_i is a geometric random variable with parameter $\frac{n-(i-1)}{n}$.

Lemma (3.8)

For any geometric random variable X with parameter p, $\mathrm{E}[X] = p^{-1}$ and $\mathrm{Var}[X] = \frac{1-p}{p^2}$.

Proof.

We have $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$. For geometric variable, $\mathrm{E}[X]^2 = p^{-2}$. Working with $\mathrm{E}[X^2] = \sum_{j=1}^{\infty} j^2 \cdot \Pr[X = j]$, we have

$$\begin{aligned}
\mathrm{E}[X^2] &= \sum_{j=1}^{\infty} j^2 \cdot (1-p)^{j-1} p && \text{a geom. variable} \\
&= \frac{p}{1-p} \sum_{j=1}^{\infty} j^2 \cdot (1-p)^j \\
&= \frac{p}{1-p} \frac{(1-p)^2 + (1-p)}{p^3} && \text{formula for } \textstyle\sum_{j=1}^{\infty} i^2 x^i.
\end{aligned}$$

$\square$

Bounding Coupon Collector purchases - Chebyshev

Lemma (3.8)

For any geometric random variable X with parameter p, $\mathrm{E}[X] = p^{-1}$ and $\mathrm{Var}[X] = \frac{1-p}{p^2}$.

Proof (cont'd).

So $\mathrm{E}[X^2] = \frac{p}{1-p} \frac{(1-p)(2-p)}{p^3} \;=\; \frac{2-p}{p^2}$, hence

$$\begin{aligned}
\mathrm{E}[X^2] - \mathrm{E}[X]^2 &= \frac{2-p}{p^2} - \frac{1}{p^2} \\
&= \frac{(2-p)-1}{p^2} \\
&= \frac{1-p}{p^2},
\end{aligned}$$

as claimed. $\square$

Bounding Coupon Collector purchases - Chebyshev

$$\Pr[|X - \mathrm{E}[X]| \geq a] \;\leq\; \frac{\mathrm{Var}[X]}{a^2} \;=\; \frac{\sum_{j=1}^{n} \mathrm{Var}[X_j]}{a^2}.$$

Each individual X_j is geometric with parameter $\frac{n-(j-1)}{n}$. So each X_j has

$$\mathrm{Var}[X_j] \;=\; \frac{j-1}{n}\left(\frac{n}{(n+1-j)}\right)^2 \;\leq\; \left(\frac{n}{n+1-j}\right)^2.$$

Hence

$$\mathrm{Var}[X] \;\leq\; n^2 \sum_{j=1}^{n} \left(\frac{1}{n+1-j}\right)^2 = n^2 \sum_{j=n}^{1} \left(\frac{1}{j}\right)^2 \;=\; \frac{\pi^2 n^2}{6}.$$

(using Euler's series for the $\frac{\pi}{6}$, see page 5 of "TCS cheat sheet").

Bounding Coupon Collector purchases - Chebyshev

We know $\mathrm{Var}[X] \leq \frac{\pi^2 n^2}{6}$ for our coupon collector process.

Suppose we are willing to make $2\mathrm{E}[X]$ (about $2n\ln(n)$) purchases.

Buying this number of packets, the probability we fail to get all cards is

$$\begin{aligned} & \Pr[X > 2\mathrm{E}[X]] \\ = & \Pr[X - \mathrm{E}[X] > \mathrm{E}[X]] \\ \leq & \Pr[|X - \mathrm{E}[X]| > \mathrm{E}[X]] \end{aligned}$$

We can upper bound the probability of the bad event $|X - \mathrm{E}[X]| > \mathrm{E}[X]$ (aka "didn't get all cards") using Chebyshev's Inequality with $a = \mathrm{E}[X]$:

$$\Pr[|X - \mathrm{E}[X]| \geq \mathrm{E}[X]] \leq \frac{\mathrm{Var}[X]}{\mathrm{E}[X]^2} \leq \frac{\pi^2 n^2}{6n^2 H(n)^2}.$$

This value simplifies to $\frac{\pi^2}{6H(n)^2}$, which is less than $\frac{2}{\ln(n)^2}$, much better probability than $1/2$ (given by Markov for $2\mathrm{E}[X]$ purchases).

Bounding Coupon Collector purchases - Chebyshev

If we are willing to make as many as $10\mathrm{E}[X]$ purchases, then we will upper-bound (probability of the "bad" scenario) $\Pr[|X - \mathrm{E}[X]| \geq 9\mathrm{E}[X]]$ by setting $a = 9\mathrm{E}[X]$ in Chebyshev's Inequality:

$$\Pr[|X - \mathrm{E}[X]| \geq 9\mathrm{E}[X]] \leq \frac{\mathrm{Var}[X]}{(9\mathrm{E}[X])^2} \leq \frac{\pi^2 n^2}{6 \cdot 81 \cdot n^2 H(n)^2}.$$

Working details with π, the right-hand side is at most $\frac{1}{49 \cdot H(n)^2}$, much much less/better than the $\frac{1}{10}$ we got with Markov's Ineq.

Wrapping up today

Next week we will continue the theme of "bounding deviation from the mean" by introducing some stronger concentration inequalities called Chernoff/Hoeffding bounds (which hold for iterations of independent Poisson trials, and related distributions).

First, on Friday (to give a break) we will look at a simple random algorithm to approximately calculate **Max** Cut, and show how to *derandomize* it.

- I have distributed the spec for coursework 1, deadline is 4pm, Thursday, 14th Feb, 2018.
- Tutorials are starting next week. I have distributed the first tutorial sheet today.