Statistical Principles[*]

Overview

We will study three phenomenon of random processes that describe three very important, and possibly un-intuitive, phenomenon. The goal will be to explore, formalize, and hopefully make intuitive these phenomenon. They are

Birthday Paradox: To measure the expected collision of random events.

A random group of 23 people has about a 50% chance of having two with the same birthday.

Coupon Collectors: To measure the expectation for seeing all possible outcomes of a discrete random variable.

Consider a lottery where on each trial you receive one of n possible coupons at random. It takes in expectation about $1.577n \ln n$ trials to collect them all.

ε-Samples: To bound the number of events needed to evenly distribute random samples.

Consider the same random lottery after k trials. We expect to have k/n of each coupon, but some more, some less. For any one coupon it takes $k = (1/\varepsilon^2)\ln(1/\delta)$ trials (for $0 < \varepsilon < 1$ and $0 < \delta < 1$) so that there are no more than $n/k + \varepsilon k$ of that coupon with probability at least $1 - \delta$.

From another perspective, these all describe the effects of random variation. The first describes collision events, the second *covering* events, and the third smoothing events – how long it takes for random variation to evenly distribute. I know the last one seems a bit mysterious now, but I hope it will become more natural by the time we get there.

Model. For all settings, there is a common model of random elements drawn from a discrete universe. The universe has n possible objects; we represent this as $[n]$ and let $i \in [n]$ represent one element (indexed by i) in this universe. The n objects may be IP addresses, days of the year, words in a dictionary, but we can always have each element (IP address, day, word) map to a distinct integer i where $0 < i \leq n$. Then we study the properties of drawing k items uniformly at random from $[n]$ with replacement.

1 Birthday Paradox

First, let us consider the famous situation of birthdays. Lets make formal the setting. Consider a room of k people, chosen at random from the population, and assume each person is equally likely to have any birthday (excluding February 29th), so there are $n = 365$ possible birthdays.

The probability that any two (i.e. $k = 2$) people (Alice and Bob) have the same birthday is $1/n = 1/365 \approx 0.003$. The birthday of Alice could be anything, but once it is known by Alice, then Bob has probability $1/365$ of matching it.

To measure that at least one pair of people have the same birthday, it is easier to measure the probability that no pair is the same. For $k = 2$ the answer is $1 - 1/n$ and for $n = 365$ that is about 0.997.

[*] CS 6955 Data Mining; Spring 2012 Instructor: Jeff M. Phillips, University of Utah

For a general number k (say $k = 23$) there are $\binom{k}{2} = k \cdot (k-1)/2$ (read as k *choose* 2) pairs. For $k = 23$, then $\binom{23}{2} = 253$. Note that $\binom{k}{2} = \Theta(k^2)$.

We need for each of these events that the birthdays do not match. Assuming independence we have

$$(1 - 1/n)^{\binom{k}{2}} \quad \text{or} \quad 0.997^{253} = 0.467.$$

And the probability there is a match is thus 1 minus this number

$$1 - (1 - 1/n)^{\binom{k}{2}} \quad \text{or} \quad 1 - 0.997^{253} = 0.532,$$

just over 50%.

What are the problems with this?

- First, the birthdays may not be independently distributed. More people are born in spring. There may be non-negligible occurrence of twins.

  Sometimes this is really a problem, but often it is negligible. Other times this analysis will describe an algorithm we create, and we can control independence.

- Second, what happens when $k = n+1$, then we should always have some pair with the same birthday. But for $k = 366$ and $n = 365$ then

  $$1 - (1 - 1/n)^{\binom{k}{2}} = 1 - (364/365)^{\binom{366}{2}} = 1 - (0.997)^{66795} = 1 - 7 \times 10^{-88} < 1.$$

  Yes, it is very small, but it is less than 1, and hence must be wrong.

  Really, the probability should be

  $$1 - \left(\frac{n-1}{n}\right)^{k-1} \cdot \left(\frac{n-2}{n-1}\right)^{k-2} \cdot \left(\frac{n-3}{n-2}\right)^{k-3} \cdot \ldots = 1 - \prod_{i=1}^{k-1} \left(\frac{n-i-1}{n-i}\right)^{k-i},$$

  where in the $(n-1)$st term $(n-(n-1)-1)/(n-(n-1)) = 0/1 = 0$. We can think of checking each person against all others. In a series of epochs, each person checks against all others, that have not yet been checked. When that person is done checking, the next epoch begins. Thus, the first epoch has $k-1$ checks, each with probability $364/365$ of no match. Only if all are no matches do we continue. Then the second epoch only has 364 possible birthdays remaining, since the first person was already checked, and everyone else is different. So each of the $k-2$ checks in the second epoch has probability $363/364$ of no match, and so on.

Take away message.

- There are collisions in random data!

- More precisely, if you have n equi-probability random events, then expect after about $k = \sqrt{2n}$ events to get a collision. Note $\sqrt{2 \cdot 365} \approx 27$, a bit more than 23.

  Note that $(1 + \frac{\alpha}{t})^t \approx e^{\alpha}$ for large enough t. So setting $k = \sqrt{2n}$ then

  $$1 - (1 - 1/n)^{\binom{k}{2}} \approx 1 - (1 - 1/n)^n \approx 1 - e^{-1} \approx .63$$

  This is not exactly $1/2$, and we used a bunch of $\approx$ tricks, but it shows *roughly* what happens.

- This is pretty accurate. Note for $n = 365$ and $k = 18$ then

  $$1 - (1 - 1/n)^{\binom{k}{2}} = 1 - (364/365)^{153} \approx .34$$

  and when $k = 28$ then

  $$1 - (1 - 1/n)^{\binom{k}{2}} = 1 - (364/365)^{378} \approx .64.$$

  This means that if you keep adding (random) people to the room, the first matching of birthdays happens 28% of the time between the 18th and 28th person. When $k = 50$ people are in the room, then

  $$1 - (1 - 1/n)^{\binom{k}{2}} = 1 - (364/365)^{1225} \approx .965,$$

  and so only about 3.5% percent of the time are there no pair with the same birthday.

2 Coupon Collectors

Lets now formalize the famous coupon lottery. There are n types of coupons, and we participate in a series of independent trials, and on each trial we have equal probability $(1/n)$ of getting each coupon. *We want to collect all toys available in a McDonald's Happy Meal.* How many trials (k) should we expect to partake in before we collect all coupons?

Let r_i be the expected number of trials we need to take before receiving exactly i distinct coupons. Let $r_0 = 0$, and set $t_i = r_i - r_{i-1}$ to measure the expected number of trials between getting $i - 1$ distinct coupons and i distinct coupons.

Clearly, $r_1 = t_1 = 1$, and it has no variance. Our first trials always yields a new coupon.

Then the expected number of trials to get all coupons is $T = \sum_{i=1}^{n} t_i$.

To measure t_i we will define p_i as the probability that we get a new coupon after already having $i - 1$ distinct coupons. Thus $t_i = 1/p_i$. And $p_i = (n - i)/n$.

We are now ready for some algebra:

$$T = \sum_{i=1}^{n} t_i = \sum_{i=0}^{n-1} \frac{n}{n-i} = n \sum_{i=1}^{n} \frac{1}{i}.$$

Now we just need to bound the quantity $\sum_{i=1}^{n}(1/i)$. This is known at the nth *Harmonic Number H_n*. It is known that $H_n = \gamma + \ln n + o(1/n)$ where $\ln(\cdot)$ is the natural log (that is $\ln e = 1$) and $\gamma \approx 0.577$ is the *Euler-Masheroni constant*. Thus we need, in expectation,

$$k = T = nH_n = n(\gamma + \ln n)$$

trials to obtain all distinct coupons.

Extensions.

- What if some coupons are more likely than others. *McDonalds offers three toys: **A**lvin, **S**imon, and **T**heodore, and for every 10 toys, there are 6 Alvins, 3 Simons, and 1 Theodore.* How many trials do we expect before we collect them all?

  In this case, there are $n = 3$ probabilities $\{p_1 = 6/10, p_2 = 3/10, p_3 = 1/10\}$ so that $\sum_{i=1}^{n} = 1$.

  The analysis and tight bounds here is a bit more complicated, but the key insight is that it is dominated by the smallest probability event. Let $p^* = \min_i p_i$. Then we need about

  $$k \approx \left(\frac{1}{p^*}\right)(\gamma + \ln n)$$

  random trials to obtain all coupons.

- These properties can be generalized to a family of events from a continuous domain. Here there can be events with arbitrarily small probability of occurring, and so the number of trials we need to get all events becomes arbitrarily large (following the above non-uniform analysis). So typically we set some probability $\varepsilon \in [0, 1]$. (Typically we consider ε as something like $\{0.01, .001\}$ so $1/\varepsilon$ something like $\{100, 1000\}$. Now we want to consider any set of events with combined probability greater than ε. (We can't consider all such subsets, but we can restrict to all, say, contiguous sets – intervals if the events have a natural ordering). Then we need

  $$k \approx \frac{1}{\varepsilon} \log \frac{1}{\varepsilon}$$

  random trials to have at least one random trial in any subset with probability at least ε. Such a set is called an *ε-net*.

Take away message.

- It takes about $n \ln n$ trials to get all items at random from a set of size n, not n. That is we need an extra about $\ln n$ factor to guarantee we hit all events.
- When probability are not equal, then it is the smallest probability item that dominates everything!
- To hit all regions of size εn we need about $(1/\varepsilon) \log(1/\varepsilon)$ samples, even if they can be covered by $1/\varepsilon$ items.

3 ε-Samples

The goal here is to obtain a random sample large enough from $[n]$ so that all elements have about the same number of occurrences. This is a bit trickier to formalize since what does "about the same number" mean?

Let S_k be the set of k random samples from $[n]$. Let f_i represent the number of trials in S_k that have value i. Clearly after k trials, the expected value $\mathbf{E}[f_i] = k/n$ for each i.

For instance, let

$$W_k = \max_i \left| f_i - \frac{k}{n} \right|.$$

It turns out, that as k increases, in expectation, W_k grows. This suggests we are best setting $k = 0$. But then this sample is useless, it tells us nothing!

A better error notion is

$$Z_k = \max_i \left| \tilde{f}_i - \frac{1}{n} \right|,$$

where $\tilde{f}_i = f_i/k$, and $1/n$ represents the expected value of $\tilde{f}_i$, that is, fraction of elements expected to have value i. Now Z_k decreases as k increases. And for some small parameter $\varepsilon \in [0, 1]$, if $Z_k \leq \varepsilon$, then we say S_k is an *ε-sample* (we will introduce a more general definition later).

So, how large does k need to be for S_k to be an ε-sample? The answer:

$$k \approx 1/\varepsilon^2$$

Note, this is independent of n. We can set $\varepsilon = c/n$, so $Z_k < c/n$ and $W_k < c$, and then we need $k \approx n^2/c^2$. Note for $c \geq 1$ this is a bit silly again, since this holds for $k = 0$. But when $c = .1$ or so, this make sense, and we need about $k \approx 100n^2$ samples to achieve this bound.

Extensions.

- This naturally extends to when elements $i \in [n]$ have non-uniform probabilities of being sampled. Then
$$Z_k = \max_i \left|\tilde{f}_i - p_i\right|,$$
where p_i is the probability that i is sampled in a random trial. Again, we need $k \approx 1/\varepsilon^2$ for S_k to be expected to be an ε-sample.

- Like with the coupon collector extension to ε-nets, this generalizes directly to continuous domains. Now for a continuous ordered domain, we can consider the fraction of samples that fall in any interval versus the expected number (do we approximate the density accurately?). For the error to be within ε for *all* intervals, we again need $k \approx 1/\varepsilon^2$ samples, in expectation.

  This generalizes again to more abstract domains through the concept of VC-dimension ν, where we need about $k \approx (\nu/2)/\varepsilon^2$ samples. This includes higher-dimensional domains (say $\mathbb{R}^d$), where balls, axis-aligned rectangles, and half spaces all require about $k \approx (d/2)/\varepsilon^2$ samples.

  Note that the analysis for this is easy for showing any single index $i \in [n]$ has at most ε error after $1/\varepsilon^2$ samples (as we will see next). It requires fair bit of extra analysis to show this for all indices, or all intervals.

- Other error metrics can also be considered such as the average error, or average squared error (as opposed to the worst case error we consider). But about the same number of samples ($k \approx 1/\varepsilon^2$) are still needed to achieve ε error bounds.

Take away message.

- Requires sample of size about $k \approx 1/\varepsilon^2$ to achieve an ε-sample (or about $k \approx n^2$ samples if we require constant amount of absolute error).

- This generalizes to any simple enough range (e.g. interval) of values, even in a continuous setting.

- Need to be careful in how to define "smoothness" of a distribution, but most reasonable measures require about the same number of samples to achieve "smoothness."

4 Tail Bounds

Up until now, we have looked at expected values on these phenomenon. But often we want to ensure that the probability of these phenomenon occurring approaches 1. We will use two tools: the Markov Inequality, and the Chernoff-Hoeffding inequality.

Markov inequality. Consider a random variable X such that all possible values of X are non-negative, then
$$\mathbf{Pr}[X > \alpha] \leq \frac{\mathbf{E}[X]}{\alpha}.$$
This is not too hard to see. Consider if this was not true, and $\mathbf{Pr}[X > \alpha] > \mathbf{E}[X]/\alpha$. Let $\gamma = \mathbf{Pr}[X > \alpha]$. Then, since $X > 0$, we need to make sure the expected value of X does not get too large. So, let the instances of X from the probability distribution of its values which are less than $\mathbf{E}[X]/\alpha$ be as small as possible, namely 0. Then we can still reach a contradiction:
$$\mathbf{E}[X] \geq (1-\gamma)0 + (\gamma)\alpha = \gamma\alpha > \frac{\mathbf{E}[X]}{\alpha}\alpha = \mathbf{E}[X].$$

Chernoff-Hoeffding inequalities. This is a much stronger inequality, but is basically, just an extension of the Markov inequality. There are many different forms that are easily translated between. I will state several of them (from general to specific), so hopefully it is easy to find the one that fits best.

[CH1] Consider a set of r independent random variables $\{X_1, \ldots, X_r\}$ such that $a_i \leq X_i \leq b_i$ for each $i \in [r]$. Let $\Delta_i = b_i - a_i$. Let $M = \sum_{i=1}^r X_i$ (a sum of X_is). Then

$$\mathbf{Pr}[|M - \mathbf{E}[M]| > \alpha] \leq 2 \exp\left(\frac{-2\alpha^2}{\sum_{i=1}^r \Delta_i^2}\right).$$

[CH2] Consider a set of r independent random variables $\{X_1, \ldots, X_r\}$ such that $a_i \leq X_i \leq b_i$ for each $i \in [r]$. Let $\Delta_i = b_i - a_i$. Let $A = \frac{1}{r}\sum_{i=1}^r X_i$ (an average of X_is). Then

$$\mathbf{Pr}[|A - \mathbf{E}[A]| > \alpha] \leq 2 \exp\left(\frac{-2\alpha^2 r^2}{\sum_{i=1}^r \Delta_i^2}\right).$$

[CH3] Consider a set of r independent random variables $\{X_1, \ldots, X_r\}$ such that $-\Delta_i \leq X_i \leq \Delta_i$ for each $i \in [r]$. Let $A = \frac{1}{r}\sum_{i=1}^r X_i$ (an average of X_is) and $\mathbf{E}[A] = 0$. Then

$$\mathbf{Pr}[|A| > \alpha] \leq 2 \exp\left(\frac{-\alpha^2 r^2}{2\sum_{i=1}^r \Delta_i^2}\right).$$

[CH4] Consider a set of r independent identically distributed (iid) random variables $\{X_1, \ldots, X_r\}$ such that $-\Delta \leq X_i \leq \Delta$ and $\mathbf{E}[X_i] = 0$ for each $i \in [r]$. Let $A = \frac{1}{r}\sum_{i=1}^r X_i$ (an average of X_is) and $\mathbf{E}[A] = 0$. Then

$$\mathbf{Pr}[|A| > \alpha] \leq 2 \exp\left(\frac{-\alpha^2 r}{2\Delta^2}\right).$$

[CH5] Consider a set of r independent identically distributed (iid) random variables $\{X_1, \ldots, X_r\}$ such that $-\Delta \leq X_i \leq \Delta$ for each $i \in [r]$. Let $M = \sum_{i=1}^r X_i$ (a sum of X_is). Then

$$\mathbf{Pr}[|M - \mathbf{E}[M]| > \alpha] \leq 2 \exp\left(\frac{-\alpha^2}{2r\Delta^2}\right).$$

To prove these bounds we adapt the Markov bound to see for any $t > 0$

$$\mathbf{Pr}[M > \alpha] = \mathbf{Pr}[\exp(tM) > \exp(t\alpha)] \leq \frac{\mathbf{E}[\exp(tM)]}{\exp(t\alpha)} = \frac{\prod_{i=1}^r \mathbf{E}[\exp(tX_i)]}{\exp(t\alpha)}.$$

The setting t appropriately (separately for two cases: when $M > 0$ and when $M < 0$), yields the desired bounds. We will omit these long tedious details here.

4.1 Proving ε-Sample Bounds

One use of the Chernoff-Hoeffding inequality is to prove a weaker form of the main result for ε-samples. We consider the number of random trials we expect to see with index $i \in [n]$.

Here we have k random variables $\{X_1, \ldots, X_k\}$, one for each random trial. We have $X_j = 1$ if the jth trial chooses an element $i \in [n]$, and $Y_j = 0$ otherwise. We note that $\Delta_i = 1$ since each X_i is in $[0, 1]$ with $a_i = 0$ and $b_i = 1$.

It follows that $M = \sum_{j=1}^k X_j$ is the number of random trials with index i, and $A = M/k = \frac{1}{k}\sum_{j=1}^k X_j$ is the fraction of random trials with index i. That is $\tilde{f}_i = A$ and $f_i = M$ in this context; $\mathsf{E}[A] = k/n$. So setting $\alpha = \varepsilon$ we can apply the Chernoff-Hoeffding bound ([CH2]), for some parameter $\delta \in (0,1)$, to say:

$$\mathsf{Pr}\left[\left|\tilde{f}_i - \frac{k}{n}\right| > \varepsilon\right] = \mathsf{Pr}[|A - E[A]| > \varepsilon] \leq 2\exp\left(\frac{-2\alpha^2 k^2}{\sum_{j=1}^k \Delta_i^2}\right) = 2\exp\left(-2\varepsilon^2 k\right) \leq \delta.$$

Solving for k in terms of ε and δ yields:

$$k \geq \frac{2}{\varepsilon^2}\ln\frac{2}{\delta}.$$

Note that this applies only for a single index i. We want this to be true for all indices at the same time with the same probability of failure δ. For this we need another probabilistic trick: the *union bound*.

The Union Bound. Consider t random variables $\{Z_1, \ldots, Z_t\}$ where each random variable Z_i is 1 with probability p_i, and is 0 with probability $q_i = 1 - p_i$. Then *all* random variables are 1 with probability $p \geq 1 - \sum_{i=1}^t q_i$.

The key is to add the probability of failures, and subtract this from 1 to lower bound the probability they are all 1.

Lets apply this to our ε-sample result for one index i to see how the bound on k changes to get $1 - \delta$ probability of having at most ε error on *all* indices. There are n indices and lets say we select $k = (2/\varepsilon^2)\ln(2/\delta')$ samples total, so each index is correct with probability at least $1 - \delta'$. Thus, each index has more than ε error with probability at most δ'. Thus no index has probability of failure more than $n\delta'$. Solving for all indices having at most ε error with probability at least $1 - \delta$ requires setting $1 - \delta = 1 - n\delta'$, and yields $\delta' = \delta/n$ and

$$k = \frac{2}{\varepsilon^2}\ln\frac{2n}{\delta}.$$

Note we now have a factor of n in the number of samples we need, but at least it is wrapped inside of a $\ln(\cdot)$ term, so its effect is pretty small.

It is possible to completely remove the $\ln(n)$ term from this bound (yielding the ε-sample result stated above), but it is a considerably more complicated. The most common approach uses a clever form a *negative dependence*. Note that two indices are not independent, but if one is close to expected value, then all others are just as likely (and in fact) more likely to be close to their expected value.