Notes for Recitation 12 (5/1/98)

1 Estimating Binomial Distributions

Let $f_{n,p}(k)$ be the probability of throwing k out of n heads with a coin whose probability of heads is p, i.e., $f_{n,p}$ is the *binomial distribution function*:

$$f_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

We have from lecture that

$$f_{n,p}(\alpha n) = \frac{2^{\left(\alpha \log_2\left(\frac{p}{\alpha}\right) + (1-\alpha)\log_2\left(\frac{1-p}{1-\alpha}\right)\right)\cdot n} \overbrace{e^{a_n - a_{\alpha n} - a_{n-\alpha n}}}^{\sim 1}}{\sqrt{2\pi\alpha(1-\alpha)n}}$$

where, the a_i symbols arise from the error in Stirling's approximation, namely, a_i denotes a value between $1/(12i+1)$ and $1/12i$. This expression looks nasty, but is really very useful. It is proved by routine manipulation from Sterling's formula:

$$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{a_n}$$

The proof is given in Figure 1.

We also have

$$F_{n,p}(\alpha n) \le \frac{1-\alpha}{1-\alpha/p} f_{n,p}(\alpha n) \qquad \text{for } \alpha < p$$

where $F_{n,p}$ is the *cumulative distribution function* for $f_{n,p}$

$$F_{n,p}(x) = \sum_{y \le x} f_{n,p}(y).$$

$$
\begin{aligned}
f_{n,p}(\alpha n) &= \binom{n}{\alpha n} p^{\alpha n}(1-p)^{(1-\alpha)n} \\
&= \frac{n!}{(\alpha n)!((1-\alpha)n)!}P, \qquad \text{where } P ::= p^{\alpha n}(1-p)^{(1-\alpha)n}, \\
&= P\frac{\sqrt{2\pi n}(n/e)^n e^{a_n}}{(\sqrt{2\pi\alpha n}((\alpha n)/e)^{\alpha n}e^{a_{\alpha n}})(\sqrt{2\pi(1-\alpha)n}((1-\alpha)n/e)^{(1-\alpha)n}e^{a_{(1-\alpha)n}})} \\
&= P\frac{\sqrt{2\pi n}}{\sqrt{2\pi\alpha n}\sqrt{2\pi(1-\alpha)n}}\frac{(n/e)^n}{(\alpha n/e)^{\alpha n}((1-\alpha)n/e)^{(1-\alpha)n}}\frac{e^{a_n}}{e^{a_{\alpha n}}e^{a_{(1-\alpha)n}}} \\
&= PS\frac{n^n/e^n}{((\alpha n)^{\alpha n}/e^{\alpha n})((1-\alpha)n)^{(1-\alpha)n}/e^{(1-\alpha)n}}E, \\
&\qquad \text{where } S ::= 1/\sqrt{2\pi\alpha(1-\alpha)n} \text{ and } E ::= e^{a_n - a_{\alpha n} - a_{(1-\alpha)n}}, \\
&= PSE\frac{n^n/e^n}{(\alpha n)^{\alpha n}((1-\alpha)n)^{(1-\alpha)n}/e^n} \\
&= PSE\frac{n^n}{(\alpha n)^{\alpha n}((1-\alpha)n)^{(1-\alpha)n}} \\
&= PSE\frac{n^n}{\alpha^{\alpha n}(1-\alpha)^{(1-\alpha)n}n^{\alpha n}n^{(1-\alpha)n}} \\
&= PSE\frac{1}{\alpha^{\alpha n}(1-\alpha)^{(1-\alpha)n}} \\
&= SEP\alpha^{-\alpha n}(1-\alpha)^{-(1-\alpha)n} \\
&= SE2^{\log_2(P\alpha^{-\alpha n}(1-\alpha)^{-(1-\alpha)n})} \\
&= SE2^{(\alpha n\log_2 p + (1-\alpha)n\log_2(1-p) - \alpha n\log_2\alpha - (1-\alpha)n\log_2(1-\alpha))} \\
&= SE2^{n(\alpha\log_2 p + (1-\alpha)\log_2(1-p) - \alpha\log_2\alpha - (1-\alpha)\log_2(1-\alpha))} \\
&= SE2^{n(\alpha\log_2(p/\alpha) + (1-\alpha)\log_2((1-p)/(1-\alpha)))} \\
&= \frac{2^{(\alpha\log_2\frac{p}{\alpha} + (1-\alpha)\log_2\frac{1-p}{1-\alpha})n}e^{a_n - a_{\alpha n} - a_{(1-\alpha)n}}}{\sqrt{2\pi\alpha(1-\alpha)n}}
\end{aligned}
$$

Figure 1: Proving the Approximation Formula for $f_{n,p}(\alpha n)$

1.1 Transmission Across a Noisy Channel

Suppose that we are transmitting bits across a noisy channel. (For example, say your modem uses a phone line that faintly picks up a local radio station.) Suppose we transmit $10,000$ bits, and each arriving bit is incorrect with probability 0.01. Assume that these errors occur independently (this is not realistic). What is the probability that more than 2% of the bits are erroneous?

We can solve this problem using the bound given above on $F_{n,p}$. However, one trick is required. The bound only applies to $\alpha < p$; therefore, we have to work in terms of *correct* bits instead of *erroneous* bits.

$$\begin{aligned}
\Pr(> \text{ than } 2\% \text{ errors}) &= \Pr(\leq \ 98\% \text{ correct}) \\
&= F_{n,0.99}(0.98n) \\
&\leq 1.98 \frac{2^{-0.005646 \cdot 10,000}}{0.3509\sqrt{10,000}} \\
&\leq 2^{-60}
\end{aligned}$$

The probability that more than 2% of the bits are erroneous is incredibly small! This again demonstrates the extreme improbability of outcomes on the tails of the binomial distribution.

1.2 Polling

Another good example of binomial distributions comes up in polling. The Gallup polling service reported that on November 10, only 37% of American adults thought Louise Woodward was guilty. Furthermore, Gallup asserted that since they asked the opinions of 623 adults, "one can say with 95 percent confidence that the error attributable to sampling and other random effects is within plus or minus 4 percentage points". Can we confirm this claim?

We want to determine p, the fraction of Americans that think Louise is guilty. Our plan is to sample the opinion of n people chosen uniformly and at random, *with replacement*. That is, we might poll the same person twice! This may seem inefficient, but it simplifies the analysis. If G is the number of people in our sample that say Louise is guilty, then we will claim that p is about G/n.

The problem is that our sample might not be representative. For example, maybe everyone in the country outside of North Dakota thinks Louise is guilty, but by bad luck the sample contained mostly North Dakotans. Then our poll would give the wrong answer.

Let ϵ be the margin of error we can tolerate, and let δ be the probability that our result lies outside this margin. How many people must we poll so that our result is within ϵ percent of national opinion with probability at least $1 - \delta$? For example, Gallup claims that for $\epsilon = 0.04$ and $\delta = 0.05$, polling 623 people is sufficient.

We can define δ, the probability that our poll is off by more than the margin of error ϵ, as follows:

$$\begin{aligned}
\delta &= \underbrace{\Pr\left(\frac{G}{n} < p - \epsilon\right)}_{\substack{\text{too many in sample}\\ \text{say ``not guilty''}}} + \underbrace{\Pr\left(\frac{G}{n} > p + \epsilon\right)}_{\substack{\text{too many in sample}\\ \text{say ``guilty''}}} \\
&= \Pr(G < (p-\epsilon)n) + \Pr(G > (p+\epsilon)n)
\end{aligned}$$

Each term in the definition of δ can be evaluated using the bound given above on $F_{n,p}$. In the second term, we must use the same trick as in the Noisy Channel problem to ensure that $\alpha < p$. We observe that $\Pr(G/n > p+\epsilon) = \Pr((n-G)/n < 1-p-\epsilon)$, where $(n-G)/n$ is the fraction of people polled who say that Louise is not guilty, and $1-p$ is the fraction of all Americans who say that she is not guilty. This gives:

$$\delta \leq F_{n,p}((p-\epsilon)n) + F_{n,1-p}((1-p-\epsilon)n)$$

This is an expression for the probability that our poll is off by more than the margin of error. The problem is that the expression contains p, the fraction of Americans that think Louise is guilty. This is the number we are trying to determine by polling!

Fortunately, we can upper bound δ by using the following:

Fact 1 For all ϵ, the maximum value of δ occurs when $p = \frac{1}{2}$.

The fact implies that to get an upper bound on δ, we can pretend that half of the people think Louise is guilty. This gives:

$$\begin{aligned}
\delta &\leq F_{n,\frac{1}{2}}((\frac{1}{2}-\epsilon)n) + F_{n,1-\frac{1}{2}}((1-\frac{1}{2}-\epsilon)n) \\
&= 2F_{n,\frac{1}{2}}((\frac{1}{2}-\epsilon)n)
\end{aligned}$$

Now suppose that we want a margin of error of 4% as Gallup claimed. Plugging in $\epsilon = 0.04$ gives:

$$\begin{aligned}\delta &\leq 2F_n(0.46n) \\ &\leq 2 \cdot 6.75 \cdot \frac{2^{-0.004622\ldots\cdot n}}{1.2492\ldots\sqrt{n}}\end{aligned}$$

We want to poll enough people so that δ is less than 0.05. The easiest way is to plug in values for n, the number of people polled:

n = people polled	upper bound on probability poll is wrong		
500	9.7%		
600	6.4%		
623	5.9%	$\leftarrow$	Gallup's poll size
650	5.3%		
662	5.0%	$\leftarrow$	our poll size
700	4.3%		

Gallup's poll size is just about right. By our calculation, polling 662 people is sufficient to determine public opinion to within 4% with confidence of 95%. We can be certain that Gallup's poll of 623 people gives 95% confidence with at most a 4.13% margin of error. But the real margin of error may be less, since we made several approximations. Still, our approximations must be quite good, since 4.13% is quite close to Gallup's claim of 4%.

The remarkable point is that the population of the country has no effect on the poll size! Whether there are a thousand people or a billion in the country, polling only a few hundred is sufficient!

2 Sums of random variables

Sums of random variables arise in many applications. The basic question is: given a pair (or a set) of random variables, what is the distribution of the random variable defined as their sum?

For two random variables, the distribution function for their sum can be obtained by a special formula. Let the random variables be R_1 and R_2 with distribution fuctions f_1 and f_2. Then the distribution function for their sum is

$$g(x) = \sum_{x_1, x_2 : x_1 + x_2 = x} f(x_1, x_2).$$

Here f is the *joint* distribution function of f_1 and f_2, i.e., $f(x)$ is the probability that $R_1 = x_1$ and $R_2 = x_2$.

If R_1 and R_2 are independent, then $f(x_1, x_2) = f_1(x_1)f_2(x_2)$, so

$$g(x) = \sum_{x_1, x_2 : x_1 + x_2 = x} f_1(x_1)f_2(x_2).$$

Do you recognize this? The function g is the *convolution* of f_1 and f_2.

We can represent this nicely in terms of generating functions in the special case when the range of R_1 and R_2 is $\mathbb{N}$. For this we define generating functions of probabilities.

Definition 2.1 Let R be a random variable $R : S \to \mathbb{N}$, then the *probability generating function (PGF)* for R is

$$G(z) = \sum_{i=0}^{\infty} \Pr(R = i) z^i$$

Examples:

- R is Bernoulli: $\Pr(R = 1) = p$, $\Pr(R = 0) = 1 - p$.

$$G(z) = (1 - p) + pz$$

- R is uniform on $[0, N - 1]$.

$$G(z) = \frac{1}{N} + \frac{1}{N}z + \frac{1}{N}z^2 + \ldots + \frac{1}{N}z^{N-1} = \frac{1}{N}\left(\frac{1 - z^N}{1 - z}\right)$$

- R is binomial with parameters p and n.

$$G(z) = \sum_{k=0}^{n} \binom{n}{k} p^k (1 - p)^{n-k} z^k = [(1 - p) + pz]^n$$

Now we can state the product rule for sums of independent random variables. Notice that this is analogous to the product rule when we discussed counting.

Theorem 2.2 If R_1 and R_2 are independent random variables from S to $\mathbb{N}$ with PGF's $G_1(z)$ and $G_2(z)$, then the PGF of $R = R_1 + R_2$ is

$$G(z) = G_1(z)G_2(z)$$

Proof.

$$
\begin{aligned}
G(z) &= \sum_{n=0}^{\infty} \Pr(R = n) z^n \\
\Pr(R = n) &= \Pr(R_1 + R_2 = n) \\
&= \sum_{i,j:i+j=n} \Pr(R_1 = i \wedge R_2 = j) \\
&= \sum_{i,j:i+j=n} \Pr(R_1 = i) \Pr(R_2 = j)
\end{aligned}
$$

The last step uses the fact that R_1 and R_2 are independent. But then, the coefficient of z^n in $G(z)$ is

$$
\Pr(R = n) = \text{coefficient of } z^n \text{ in } \left(\sum_{i=0}^{\infty} \Pr(R_1 = i) z^i\right) \left(\sum_{j=0}^{\infty} \Pr(R_2 = j) z^j\right)
$$

Hence $G(z) = G_1(z)G_2(z)$. ■

As an example here is a problem we've seen before, but we'll view it differently. There are n devices in a circuit we are building. Each of the n devices could be faulty. The probability that any one fails is p, and the failures are independent. From this we would like to get the distribution for the number of failures — for example, what is the probability that there are 2 failures?.

Define the random variables $R_1, \ldots, R_n$ as

$$
R_i = \begin{cases} 1 & i^{th} \text{ device fails} \\ 0 & \text{otherwise} \end{cases}
$$

Then,

$$
\forall i, \Pr(R_i = 1) = p
$$

Let R be the number of failures overall.

$$
R = R_1 + R_2 \ldots + R_n
$$

The PGF for each R_i is $(1 - p) + pz$. Therefore, by the product rule, the PGF for R is $[(1 - p) + pz]^n$. Hence, the probability of k errors is

$$
\begin{aligned}
\Pr(R = k) &= \text{coefficient of } z^k \\
&= \binom{n}{k} p^k (1 - p)^{n-k}
\end{aligned}
$$

This is also an alternative derivation of the Binomial distribution.

The expectation of a random variable can be expressed compactly via its PGF.

Theorem 2.3 Let $R : S \to \mathbb{N}$ be a random variable with PGF $G(z)$.

$$E(R) = G'(1)$$

Proof.

$$\begin{aligned} G(z) &= \sum_{i=0}^{\infty} \Pr(R = i) z^i \\ G'(z) &= \sum_{i=0}^{\infty} \Pr(R = i) i z^{i-1} \\ G'(1) &= \sum_{i=0}^{\infty} \Pr(R = i) i \\ &= E(R) \end{aligned}$$

For example, the expectation of a binomial distribution with parameters p and n:

$$G'(z) = n[(1-p) + pz]^{n-1} p$$

$$G'(1) = np$$

3 The Poisson distribution (Optional)

Next we study the Poisson distribution which occurs in a surprising variety of real-life applications.

One setting for a Poisson distribution is random events occurring in time, such as radioactive disintegrations, incoming telephone calls or people coming to a cash register. Each event can be a point on the time axis and we are interested in the distribution of points, namely, what is the probability of finding exactly k points within a fixed interval of specified length? So for example this could be the number of α-particles emitted by a radioactive substance that reach a given portion of space during time t. This particular example is actually known to correspond very closely to the Poisson distribution.

The *Poisson* distribution is

$$\Pr(R = k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

for the random variable $R : S \to \mathbb{N}$. Here $\lambda > 0$ is a parameter of the distribution.

Let us quickly check that this indeed a valid distribution.

$$\begin{aligned}\sum_{k=0}^{\infty} \Pr(R=k) &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= e^{-\lambda} e^{\lambda} \\ &= 1\end{aligned}$$

In the final exam a few years ago, students were asked to calculate the expectation of R. For this let us write the PGF for R.

$$\begin{aligned}G(z) &= \sum_{k=0}^{\infty} \Pr(R=k) z^k \\ &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} z^k \\ &= e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda z)^k}{k!} \\ &= e^{-\lambda} e^{\lambda z} \\ &= e^{-\lambda+\lambda z}\end{aligned}$$

Then $G'(z) = \lambda e^{-\lambda+\lambda z}$ and $E(R) = G'(1) = \lambda$.

They were also asked to show that the sum of independent Poisson distributed random variables is itself Poisson distributed.

Let R_1, R_2 be random variables with distributions

$$\Pr(R_1 = k) = \frac{\lambda_1^k e^{-\lambda_1}}{k!}$$
$$\Pr(R_2 = k) = \frac{\lambda_2^k e^{-\lambda_2}}{k!}$$

Let $R = R_1 + R_2$ with PGF G. To find the distribution of R let us evaluate its PGF.

$$G(z) = G_1(z) G_2(z)$$

Where G_1 and G_2 are the PGF's for R_1 and R_2 respectively.

$$\begin{aligned}G(z) &= e^{-\lambda_1+\lambda_1 z} e^{-\lambda_2+\lambda_2 z} \\ &= e^{-(\lambda_1+\lambda_2)+(\lambda_1+\lambda_2)z}\end{aligned}$$

But this is the PGF for a random variable with a Poisson distribution with parameter $\lambda_1 + \lambda_2$. Hence the sum of two Poisson distributions is Poisson.

4 Team Problems

Problem 1. This is a real problem that came from a research project in computational biology last term. The CS PhD student hadn't taken 6.042 and so he needs our help. Can you solve the problem?

It is known that atoms occur in different isotopes. For example, Carbon atoms usually have 6 protons and 6 neutrons, but occasionally they have 7 neutrons. In fact 1.11% of all Carbon atoms have 7 neutrons and 98.89% have 6. From what we know, the appearance of the different isotopes seems to be a random event, unrelated to the functionality of the atom. We can tell which Carbon atoms have an extra neutron only by their atomic weight. The atomic weight of the generic Carbon atom is 12 and of its isotope is 13.

The research team is studying proteins composed of Carbon, Oxygen, Nitrogen and Hydrogen (the four building blocks of life). Their atomic weights and isotopes are listed below. The second column is the atomic weight and the third column is the frequency of the isotope.

Hydrogen	1	$P_{H,1}$
	2	$P_{H,2}$
Carbon	12	$P_{C,12}$
	13	$P_{C,13}$
Nitrogen	14	$P_{N,14}$
	15	$P_{N,15}$
Oxygen	15	$P_{O,15}$
	16	$P_{O,16}$
	17	$P_{O,17}$

The research team wants to write software that can predict the distribution of atomic weights for any protein. Describe a simple algorithm for computing the weight distribution of any protein that contains n_C carbon atoms, n_O oxygen atoms, n_N nitrogen atoms and n_H hydrogen atoms. [Hint: Use generating functions!]

Solution: Let R_H be the random variable whose value is the atomic weight of a random Hydrogen atom. Similarly, let R_C, R_N and R_O be the random variables for the atomic weight of Carbon, Nitrogen and Oxygen atoms.

We can write down the generating functions for their distributions:

$$\begin{aligned} G_H(z) &= P_{H,1}z + P_{H,2}z^2 \\ G_C(z) &= P_{C,12}z^{12} + P_{C,13}z^{13} \\ G_N(z) &= P_{N,14}z^{14} + P_{N,15}z^{15} \\ G_O(z) &= P_{0,15}z^{15} + P_{O,16}z^{16} + P_{O,17}z^{17} \end{aligned}$$

Then, let R be the random variable representing the atomic weight of a protein with n_C carbon atoms, n_O oxygen atoms, n_N nitrogen atoms and n_H hydrogen atoms. We can write:

$$R = \sum_{i=1}^{n_H} R_{H,i} + \sum_{i=1}^{n_C} R_{C,i} + \sum_{i=1}^{n_N} R_{N,i} + \sum_{i=}^{n_O} R_{O,i}$$

Here the distribution of each hydrogen atom in the protein, $R_{H,i}$, is identical to the distribution of R_H. Similarly, the $R_{C,i}$'s have the same distribution as R_C, the $R_{N,i}$ have the same distribution as R_N and the $R_{O,i}$'s have the same distribution as R_O.

By the product rule, the PGF for R is

$$\begin{aligned} G(z) &= (P_{H,1}z + P_{H,2}z^2)^{n_H}(P_{C,12}z^{12} + P_{C,13}z^{13})^{n_C} \\ &\quad (P_{N,14}z^{14} + P_{N,15}z^{15})^{n_N}(P_{0,15}z^{15} + P_{O,16}z^{16} + P_{O,17}z^{17})^{n_O} \end{aligned}$$

Hence the probability that a protein molecule with n_H, n_C, n_N, n_O atoms of H, C, N, O respectively has atomic weight k is the coefficient of z^k in the above polynomial.

In fact, the student has now written a program to multiply the polynomials and so the problem is solved!

Problem 2. Coupon Collector problem: We are trying to collect baseball cards. Assume there are n different cards. We can buy them one at a time. When we buy a card, it is equally likely that we get any one of the n cards, and purchases are mutually independent. We want to study how many cards we need to buy before we get one each of the n different kinds.

(a) Suppose we already have k different cards. What is the expected number of cards we need to buy to get a new card?

Solution: Define X_k to be the random variable denoting the number of purchases to get the $(k+1)$st new card, after getting k different cards. After getting k different cards, a new purchase has probability $(n-k)/n$ of getting a new card. We can view each new purchase as Bernoulli trial with probability of success $p = (n-k)/n$. Hence the expected number of trials to success is $1/p = n/(n-k)$. That is,

$$E(X_k) = \frac{n}{n-k}.$$

(b) Using the above, what is the expected number of cards we need to buy to get one of each kind?

Solution: Since $X_0 = 1$, the number of purchases we need is $1+\sum_{i=1}^{n-1} X_i$. By linearity of expectation,

$$\begin{aligned} E(1+\sum_{i=1}^{n-1} X_i) &= 1+\sum_{i=1}^{n-1} E(X_i) \\ &= 1+\sum_{i=1}^{n-1} \frac{n}{n-i} \\ &= n\sum_{i=1}^{n} \frac{1}{i} \\ &= n\mathrm{H}_n \sim n\ln n + \gamma n + \Theta(1) \end{aligned}$$

Where γ is Euler's constant.

(c) Suppose we buy $n\ln n + cn$ cards, where $c > 0$ is a real number. What is the probability that we don't have the first card?

Solution: Let A_1 be the event that we don't have the first card in $n\ln n+cn$ trials.

In any single trial the probability that we don't get A_1 is $1 - 1/n$. Since the trials are independent,

$$\begin{aligned} \Pr(A_1) &= (1-\frac{1}{n})^{n\ln n+cn} \\ &\leq e^{-\frac{1}{n}(n\ln n+cn)} \\ &= e^{-\ln n+c} = \frac{e^{-c}}{n} \end{aligned}$$

(d) Suppose we buy $n\ln n + cn$ cards, what is an upper bound on the probability that we don't have all n cards?

Solution: Let A_i be the event that we don't have the i^{th} card after $n\ln n + cn$ trials.

Then the probability that we don't have all the cards is

$$\Pr(A_1 \cup \ldots \cup A_n) \leq \sum_{i=1}^{n} \Pr(A_i)$$

by Boole's inequality. But this is

$$n\frac{e^{-c}}{n} = e^{-c}.$$