A NOTE ON POISSON THINNING

SOUMENDU SUNDAR MUKHERJEE

Note: If you have comments or questions, feel free to email me at soumendu@berkeley.edu. Thanks to Professor Jim Pitman for his careful reading of this note and several suggestions that improved the exposition substantially.

One of the students in one of the discussion sessions of Stat 201A asked me about a more conceptual explanation of the independence of S_N and $N - S_N$ in Problem 2(b) of Worksheet #1, and also about why the assumption $N \sim \text{Poisson}(\mu)$ is important. In this note I will attempt to explain these via the "Poisson approximation of Binomial" heuristic. We will also see that Poisson is actually characterized by this property.

Roughly speaking, Poisson approximation of Binomial says that *count of many unlikely events follows an approximate Poisson distribution.* To be more concrete, suppose that we have n i.i.d. Bernoulli trials X_i with success probability μ/n. Then if $N = \sum_{i=1}^n X_i$ is the total number of successes in n trials, we know that N has a Binomial distribution, namely $\text{Bin}(n, \mu/n)$.

Exercise 1 (Poisson approximation of Binomial)**.** *Prove that*

$$\mathbb{P}(N = k) = \binom{n}{k} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{n-k} \xrightarrow{n\to\infty} \frac{e^{-\mu}\mu^k}{k!}.$$

So, roughly speaking we can think of N as approximately a Poisson random variable with mean μ.

Coming back to our problem, let us interpret N and S_N in this light. Suppose, for a large number n, we have n i.i.d. Bernoulli trials $X_i \sim \text{Ber}(\mu/n)$. Independent of the X_i's, suppose we have n i.i.d. Bernoulli trials $Y_i \sim \text{Ber}(p)$. Set $N = \sum_{i=1}^n X_i$. Then, as discussed above, N is approximately a $\text{Poisson}(\mu)$ random variable. Now define

$$S_N := \sum_{i=1}^n X_i Y_i.$$

Note that this is essentially a sum of N i.i.d. $\text{Ber}(p)$ random variables (to see this note that, by definition of N, exactly N of the X_i's are 1, and the rest are 0). So, indeed S_N is "approximately" a sum of $\text{Poisson}(\mu)$ many i.i.d. $\text{Ber}(p)$ trials. Note also that

$$N - S_N = \sum_{i=1}^n X_i(1 - Y_i).$$

Let us now calculate the joint distribution of S_N and $N - S_N$. It is obvious that $X_iY_i, X_i(1 - Y_i)$ and $(1 - X_i)$ are jointly $\text{Multinomial}(1; \frac{\mu p}{n}, \frac{\mu(1-p)}{n}, (1 - \frac{\mu}{n}))$ so that $S_N, N - S_N$ and $n - N$ are jointly

Date: September 20, 2015.

Multinomial$(n; \frac{\mu p}{n}, \frac{\mu(1-p)}{n}, (1-\frac{\mu}{n}))$. Let us do an alternative and more elaborate proof of this fact here. We have

$$\begin{aligned}&\mathbb{P}(S_N = k, N - S_N = l)\\ &= \sum_{y_1,\ldots,y_n} \mathbb{P}\left(S_N = k, N - S_N = l \mid Y_1 = y_1, \ldots, Y_n = y_n\right) \times \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n).\end{aligned}$$

Now comes the crucial observation: once we know the Y_i's, S_N is basically the sum of X_i's corresponding to those i's for which $Y_i = 1$, and $N - S_N$ is the sum of X_i's corresponding to those i's for which $Y_i = 0$. Thus given the values of Y_i's S_N and $N - S_N$ depend on disjoint sets of the X_i's, and therefore are independent. Moreover, being sums of i.i.d. Bernoulli's both of them are Binomial:

$$\begin{aligned}\mathbb{P}\left(S_N = k, N - S_N = l \mid Y_1 = y_1, \ldots, Y_n = y_n\right) &= \binom{\sum y_i}{k} \left(\frac{\mu}{n}\right)^k \left(1 - \frac{\mu}{n}\right)^{\sum y_i - k}\\ &\quad \times \binom{n - \sum y_i}{l} \left(\frac{\mu}{n}\right)^l \left(1 - \frac{\mu}{n}\right)^{n - \sum y_i - k}\\ &= \binom{\sum y_i}{k}\binom{n - \sum y_i}{l} \left(\frac{\mu}{n}\right)^{k+l} \left(1 - \frac{\mu}{n}\right)^{n-k-l}.\end{aligned}$$

Using this we get,

$$\mathbb{P}(S_N = k, N - S_N = l) = \sum_{y_1,\ldots,y_n} \binom{\sum y_i}{k}\binom{n - \sum y_i}{l} \left(\frac{\mu}{n}\right)^{k+l} \left(1 - \frac{\mu}{n}\right)^{n-k-l} \times \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n).$$

Exercise 2. *Justify the following equality:*

$$\begin{aligned}&\sum_{y_1,\ldots,y_n} \binom{\sum y_i}{k}\binom{n - \sum y_i}{l} \left(\frac{\mu}{n}\right)^{k+l} \left(1 - \frac{\mu}{n}\right)^{n-k-l} \times \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n)\\ &\qquad = \sum_{m=k}^{n-l} \binom{m}{k}\binom{n - m}{l} \left(\frac{\mu}{n}\right)^{k+l} \left(1 - \frac{\mu}{n}\right)^{n-k-l} \times \mathbb{P}(\sum Y_i = m).\end{aligned}$$

Exercise 3. *Noting that $\sum Y_i$ is a Bin(n, p) random variable, and using Exercise 2 show that*

$$\mathbb{P}(S_N = k, N - S_N = l) = \frac{n!}{k! \times l! \times (n-k-l)!} \left(\frac{\mu p}{n}\right)^k \left(\frac{\mu(1-p)}{n}\right)^l \left(1 - \frac{\mu}{n}\right)^{n-k-l}.$$

Note that this just says that $S_N, N - S_N$ and $n - N$ are jointly Multinomial$(n; \frac{\mu p}{n}, \frac{\mu(1-p)}{n}, (1-\frac{\mu}{n}))$ as observed before.

Exercise 4. *Conclude that*

$$\mathbb{P}(S_N = k, N - S_N = l) \xrightarrow{n\to\infty} \frac{(\mu p)^k (\mu(1-p))^l}{k! \times l!} e^{-\mu} = \mathbb{P}(N_1 = k) \times \mathbb{P}(N_2 = l),$$

where $N_1 \sim$ Poisson(μp) and $N_2 \sim$ Poisson$(\mu(1-p))$.

This shows that S_N and $N - S_N$ are approximately independent, at least for large n, and have approximate Poisson distributions with means μp and $\mu(1-p)$ respectively. This gives us a heuristic justification of why the Poisson distribution is important: it is really the underlying "rare Bernoulli

trials X_i" which cause the (approximate) independence of S_N and $N-S_N$. You may find it interesting to rephrase all these in the coin tossing language.

Now it turns out that Poisson distribution is the only (non-degenerate) distribution having this property. To be precise we have the following theorem.

Theorem 0.1. *If N is a non-negative integer valued random variable with $\mathbb{P}(N=0)<1$ and X_i's are i.i.d. Bernoulli(p) random variables independent of N, then $S_N := \sum_{i=1}^{N} X_i$ is independent of $N-S_N$ if and only if N is Poisson.*

Proof. The *if* part is straightforward and was exactly the content of Problem 2(b) of Worksheet #1. The proof of the *only if* part is a couple of simple exercises involving the probability generating function (PGF). Let $\phi(z)=\mathbb{E}(z^N)$ be the PGF of N.

Exercise 5. *Show that the joint PGF of S_N and $N-S_N$ is given by*

$$G(z_1,z_2)=\phi(pz_1+(1-p)z_2).$$

Now since N and $N-S_N$ are supposed to be independent, we must have that

$$G(z_1,z_2)=G(z_1,1)G(1,z_2),$$

i.e. ϕ must satisfy the following functional equation

$$\phi(pz_1+(1-p)z_2)=\phi(pz_1+(1-p))\phi(p+(1-p)z_2),$$

which looks suspiciously similar to Cauchy's functional equation.

Exercise 6. *Show that $g(z)=\log\phi(z+1)$ satisfies Cauchy's functional equation:*

$$g(z_1+z_2)=g(z_1)+g(z_2)$$

and conclude that

$$g(z)=\lambda z,$$

for some constant λ. Argue why $\lambda>0$, necessarily.

This tells us that $\phi(z)=e^{\lambda(z-1)}$, i.e. N follows a Poisson distribution with mean λ. □

The story does not end here. We may ask if the following "converse" is true: if we knew that N is Poisson, then for independence of S_N and $N-S_N$ to hold, we ought to have that X_i's are Bernoulli. The answer turns out to be yes! Before we embark on a proof of this fact, let's derive a general functional equation for the PGF's, characterizing independence of S_N and $N-S_N$.

Exercise 7. *Suppose that X_i's are non-negative integer valued random variables, i.i.d. with PGF ψ and N is an independent non-negative integer valued random variable with PGF ϕ. Then S_N and $N-S_N$ are independent if and only if their joint PGF factorizes, i.e.*

$$\phi(z_2\psi(z_1/z_2))=\phi(\psi(z_1))\phi(z_2\psi(z_2^{-1})). \tag{1}$$

Theorem 0.2. *Suppose that N is Poisson(λ). Then in order for S_N and $N-S_N$ to be independent we must have that X_1 is Bernoulli.*

Proof. This will again be a series of simple exercises.

Exercise 8. *Show using (1) that if S_N and $N-S_N$ are independent, then ψ must satisfy the following functional equation*

$$z_2\psi(z_1/z_2)+1=\psi(z_1)+z_2\psi(z_2^{-1})$$

Write $z_2^{-1}=y$ and $z_1=x$, then show that the above equation can be rewritten as

$$\psi(xy)-\psi(y)=y(\psi(x)-1) \tag{2}$$

We have to solve this functional equation. Since ψ is differentiable, we can instead look at the derivatives of this equation. (A standard trick, that, in fact, can be used to solve Cauchy's equation too. Try it!) We have, after differentiating both sides with respect to y, that

$$x\psi'(xy)-\psi'(y)=\psi(x)-1.$$

Plug in $y=1$ to conclude that

$$x\psi'(x)-\mu=\psi(x)-1, \tag{3}$$

where $\mu:=\psi'(1)$. Now that we have a linear ODE for ψ, we are in business.

Exercise 9. *Solve (3) to conclude that $X_1\sim$ Bernoulli(μ). (You must argue why $\mu\in[0,1]$.)*

This completes our proof. □

Remark 0.1. *One can work with the MGF to allow for a more general X_1 (rather than just non-negative integer-valued X_1). The analysis will be exactly similar.*