Balls and Bins II : Poisson Random Variables and Poisson Approximation[1]

- Recall the balls-and-bins setting: m balls are independently thrown into n bins. $\mathsf{L}_i^{(m)}$ is the random variable indicating the number of balls in the ith bin. These are identical but not independent random variables, whose expectation is $\frac{m}{n}$.

  In this lecture, we connect these random loads with *Poisson* random variables which are a powerful class of discrete random variables. In some sense, they form the discrete analog of the famous Gaussian random variables. Of note will be the following "approximation theorem": to argue about events involving the random load vector $\mathsf{L}^{\vec{(m)}} := (\mathsf{L}_1^{(m)}, \mathsf{L}_2^{(m)}, \ldots, \mathsf{L}_n^{(m)})$, it suffices to argue about a vector of ***independent*** Poissons, which is a much easier thing to do.

- To show the connection, let us figure out the probability $\mathsf{L}_i^{(m)}$ is exactly r for some non-negative integer r. We see that

$$\mathbf{Pr}[\mathsf{L}_i^{(m)} = r] = \underbrace{\binom{m}{r}}_{\text{ways to select } r \text{ balls}} \cdot \underbrace{\left(\frac{1}{n}\right)^r}_{\text{which all fall in bin } i} \cdot \underbrace{\left(1 - \frac{1}{n}\right)^{m-r}}_{\text{and the rest don't.}} \tag{1}$$

$$\underbrace{\approx}_{\text{when } r \ll n} \frac{m^r}{r!} \cdot \left(\frac{1}{n}\right)^r \cdot e^{-\frac{m}{n}} \tag{2}$$

  Let's list out the approximations: we have approximated $m(m-1)\ldots(m-r+1) \approx m^r$, we have approximated $\left(1 - \frac{1}{n}\right) \approx e^{-\frac{1}{n}}$, and $m - r \approx m$. All of these are "ok", when $n \gg 1$ and $r \ll n$. But the point is actually to show the connection with Poisson random variables which we describe next.

- ***Poisson Random Variables.*** A Poisson random variable Z with *parameter* μ, denoted as $Z \sim \mathrm{Pois}(\mu)$, is a non-negative integer valued random variable with pdf defined as

$$\text{For non-negative integer } r, \quad \mathbf{Pr}[Z = r] = \frac{e^{-\mu}\mu^r}{r!} \tag{Poisson Random Variable}$$

  Note that (2) is exactly the RHS of (Poisson Random Variable) when $\mu = \frac{m}{n}$. Let's verify a couple of things, and then look at some magical properties of these variables.

  Claim 1. Z defined in (Poisson Random Variable) is a valid probability distribution.

  Proof. The RHS in (Poisson Random Variable) is indeed > 0 for any r. We need to check that it sums to 1. Indeed,

$$\sum_{r=0}^{\infty} \mathbf{Pr}[Z = r] = e^{-\mu} \cdot \underbrace{\sum_{r=0}^{\infty} \frac{\mu^r}{r!}}_{\text{This is } e^{\mu}} = 1$$

  □

[1] *Lecture notes by Deeparnab Chakrabarty. Last modified : 7th April, 2021*
These have not gone through scrutiny and may contain errors. If you find any, or have any other comments, please email me at deeparnab@dartmouth.edu. Highly appreciated!

Claim 2. The expectation of $Z \sim \mathrm{Pois}(\mu)$ is μ.

Proof.

$$\mathbf{Exp}[Z] \;=\; e^{-\mu}\sum_{r=1}^{\infty}\frac{r\cdot\mu^r}{r!} = \mu e^{-\mu}\sum_{r=1}^{\infty}\frac{\mu^{r-1}}{(r-1)!} = \mu\cdot\underbrace{e^{-\mu}\sum_{s=0}^{\infty}\frac{\mu^s}{s!}}_{=1 \text{ Claim 1}} = \mu$$

□

Exercise: *Calculate the variance of $Z \sim \mathrm{Pois}(\mu)$. Surprised?*

Before we dive into the deeper connection with balls and bins, let's cover a powerful fact about Poisson random variables.

Theorem 1 (Sum of independent Poissons is Poisson)**.** Let $Z_1, \ldots, Z_n$ be n *independent* Poisson random variables with $Z_i \sim \mathrm{Pois}(\mu_i)$. Then, $Z := \sum_{i=1}^n Z_i$ is $\sim \mathrm{Pois}(\mu)$ where $\mu := \sum_{i=1}^n \mu_i$.

Proof. Let's prove this for $n = 2$ and the rest follows inductively. Let $Z = Z_1 + Z_2$. Then,

$$\begin{aligned}
\mathbf{Pr}[Z = r] \;&=\; \sum_{s=0}^{r}\mathbf{Pr}[Z_1 = s \;\wedge\; Z_2 = r-s] \underbrace{=}_{\text{independence}} \sum_{s=0}^{r}\mathbf{Pr}[Z_1 = s]\cdot\mathbf{Pr}[Z_2 = r-s] \\
&= \sum_{s=0}^{r}\left(\frac{e^{-\mu_1}\mu_1^s}{s!}\right)\cdot\left(\frac{e^{-\mu_2}\mu_2^{r-s}}{(r-s)!}\right) \\
&= e^{-(\mu_1+\mu_2)}\sum_{s=0}^{r}\frac{\mu_1^s\mu_2^{r-s}}{s!(r-s)!} \;=\; \frac{e^{-\mu}}{r!}\sum_{s=0}^{r}\underbrace{\frac{r!}{s!(r-s)!}}_{\text{observe this is}\binom{r}{s}}\mu_1^s\mu_2^{r-s} \\
&= \frac{e^{-\mu}\mu^r}{r!} \quad \text{by the Binomial Theorem}
\end{aligned}$$

□

This above facts allow us to prove exactly the same Chernoff bounds for sums of Poisson variables (which, recall, are very different from Bernoulli variables; in particular, these Poisson random variables are unbounded.)

Theorem 2 (Chernoff Bounds for Sums of Independent Poissons.)**.** Let X be a Poisson random variable with parameter μ. Then for any $t > 0$, we have

$$\mathbf{Pr}[X \geq (1+t)\mu] \;\leq e^{-\mu\cdot g(t)} \quad \text{and} \quad \mathbf{Pr}[X \leq (1-t)\mu] \;\leq e^{-\mu\cdot h(t)} \tag{3}$$

where $g(t) := (1+t)\ln(1+t) - t$ and $h(t) := (1-t)\ln(1-t) + t$.

Remark: *Consequently, using Theorem 1 one gets the following. Suppose $X_1, \ldots, X_n$ are **independent** Poisson random variables and $X = \sum_{i=1}^n X_i$. Then for any $\varepsilon \in (0,1)$,*

$$\mathbf{Pr}[X \le (1-\varepsilon)\,\mathbf{Exp}[X]] \le e^{-\frac{\varepsilon^2\,\mathbf{Exp}[X]}{2}} \tag{LT}$$

and

$$\mathbf{Pr}[X \ge (1+\varepsilon)\,\mathbf{Exp}[X]] \le e^{-\frac{\varepsilon^2\,\mathbf{Exp}[X]}{3}} \tag{UT1}$$

For the "upper tail", that is for "larger" deviations, we have when $1 \le t \le 4$, we have the following (changing ε to t so as to underscore that the deviation is big)

$$\mathbf{Pr}[X \ge (1+t)\,\mathbf{Exp}[X]] \le e^{-\frac{t^2\,\mathbf{Exp}[X]}{4}} \tag{UT2}$$

and for $t > 4$ (really large), we have

$$\mathbf{Pr}[X \ge (1+t)\,\mathbf{Exp}[X]] \le e^{-\frac{t\ln t\,\mathbf{Exp}[X]}{2}} \tag{UT3}$$

- ***The Poisson Approximation : Connection with Balls and Bins.*** Till now, the connection between balls-and-bins and Poisson random variables seems a bit tenuous: (2) is after all an approximation. Is thinking of the $\mathsf{L}_i^{(m)}$'s as Poisson random variables correct? Is it useful? The following theorem captures this connection rigorously, and is called the *Poisson Approximation*.

Theorem 3 (Poisson Approximation for Balls and Bins.)**.**

Suppose you throw m balls into n bins, each ball independently landing on a bin uniformly at random. Let $\mathcal{E}$ be an event of interest whose indicator random variable is a function of $f(\mathsf{L}_1^{(m)}, \ldots, \mathsf{L}_n^{(m)})$. Consider a second experiment where we choose n ***independent and identical*** Poisson random variables $(Z_1, \ldots, Z_n)$ where each $Z_i \sim \mathrm{Pois}(\frac{m}{n})$. Then,

$$\mathbf{Pr}[f(\mathsf{L}_1^{(m)}, \ldots, \mathsf{L}_n^{(m)}) = 1] \le e\sqrt{m} \cdot \mathbf{Pr}[f(Z_1, \ldots, Z_n) = 1] \tag{Gen-PA}$$

and if f is a *monotonically non-decreasing or non-increasing* function, then in fact

$$\mathbf{Pr}[f(\mathsf{L}_1^{(m)}, \ldots, \mathsf{L}_n^{(m)}) = 1] \le 2 \cdot \mathbf{Pr}[f(Z_1, \ldots, Z_n) = 1] \tag{Mon-PA}$$

In plain English, the probability the event $\mathcal{E}$ occurs in the balls-and-bins setting can be approximated by the probability that the same event occurs when the "loads" are independent Poisson random variables.

- *Lower bound on the maximum load.* It should be clear how Theorem 3 can be useful : we now have *independence* over the various bins which was missing in the normal balls-and-bins setting. Let us illustrate this by showing a converse to a theorem we showed in a previous lecture : when we throw n balls independently into n different bins, the maximum load is in fact $\Omega(\frac{\ln n}{\ln\ln n})$ with high probability.

Theorem 4. For large enough n, if we throw n balls into n bins, then the probability the maximum load is $\leq \frac{\ln n}{2\ln\ln n}$ is at most $2e^{-\sqrt{n}}$.

Proof. Define $f(x_1,\ldots,x_n) = 1$ if all $x_i \leq \frac{\ln n}{2\ln\ln n}$, and 0 otherwise. Note that f is a monotonically decreasing function. We are interested in upper bounding $\mathbf{Pr}[f(\mathsf{L}_1^{(n)},\ldots,\mathsf{L}_n^{(n)}) = 1]$. Instead, we will upper bound the probability $\mathbf{Pr}[f(Z_1,\ldots,Z_n) = 1]$, where $Z_i \sim \text{Pois}(1)$ (note that $m = n$ and therefore, $m/n = 1$).

First, fix an $Z_i \sim \text{Pois}(1)$ and let us calculate the probability this is less than $L := \left\lfloor \frac{\ln n}{2\ln\ln n} \right\rfloor$.

$$\mathbf{Pr}[Z_i \leq L] \;=\; e^{-1}\sum_{j\leq L}\frac{1}{j!} \;=\; e^{-1}\cdot(e - \underbrace{\sum_{j>L}\frac{1}{j!}}_{\geq \frac{1}{(L+1)!}}) \;\leq\; 1 - \frac{1}{e(L+1)!}$$

Now, since the Z_i's are *independent*, we get that $\mathbf{Pr}[f(Z_1,\ldots,Z_n) = 1] = \mathbf{Pr}[\wedge_{i=1}^n\{Z_i \leq L\}] = (\mathbf{Pr}[Z_i \leq L])^n$. Using (Mon-PA), we get

$$\mathbf{Pr}[f(\mathsf{L}_1^{(n)},\ldots,\mathsf{L}_n^{(n)}) = 1] \leq 2\cdot\left(1 - \frac{1}{e(L+1)!}\right)^n \tag{4}$$

What remains is a calculation similar to the upper bound proof. We get that for large enough n,

$$\ln\left(e(L+1)!\right) \leq \ln L^L \;=\; L\ln L \;\leq\; \frac{\ln n}{2\ln\ln n}\cdot(\ln\ln n) = \frac{\ln n}{2} \;\;\Rightarrow\;\; e(L+1)! \leq \sqrt{n}$$

Substituting in (4), we get

$$\mathbf{Pr}[f(\mathsf{L}_1^{(n)},\ldots,\mathsf{L}_n^{(n)}) = 1] \leq 2\left(1 - \frac{1}{\sqrt{n}}\right)^n \underbrace{\leq}_{\text{Use : } (1-t)\leq e^{-t} \text{ to see this}} 2e^{-\sqrt{n}}$$ □

- ***The Proof of the Poisson Approximation Theorem.*** The main observation is the following elementary lemma which states that if we throw m balls into n bins, then the ***distribution*** of the load vector is *precisely* the same as the distribution of n ***independent*** Poisson random variables with parameter $\mu := \frac{m}{n}$ *conditioned* on the event that their sum is m. That is, if we sample n independent Poisson random variables with parameter $\frac{m}{n}$ and reject anything whose sum is not m, then the resulting distribution of vectors is the same as the distribution of the loads on the n bins when m balls are thrown.

Lemma 1. For any tuple of non-negative integers $(m_1, m_2, \ldots, m_n)$ such that $\sum_{i=1}^n m_i = m$,

$$\mathbf{Pr}[(\mathsf{L}_1^{(m)}, \mathsf{L}_2^{(m)}, \ldots, \mathsf{L}_n^{(m)}) = (m_1, \ldots, m_n)] = \mathbf{Pr}\left[(Z_1, Z_2, \ldots, Z_n) = (m_1, \ldots, m_n) \mid \sum_{i=1}^n Z_i = m\right]$$

where each $Z_i \sim \mathrm{Pois}(\frac{m}{n})$ and are mutually independent.

Proof. There is not much to this lemma rather than a calculation. Let us calculate the LHS. How many ways can we split m balls into n sets such that set i has m_i balls? This is precisely the *multinomial* coefficient, and equals

$$\binom{m}{m_1, m_2, \ldots, m_n} = \frac{m!}{m_1! m_2! \cdots m_n!}$$

Given such a split, what is the probability that the first specified m_1 balls go into bin 1? The answer is $\left(\frac{1}{n}\right)^{m_1}$. Similarly for the other bins. And therefore,

$$\mathbf{Pr}[(\mathsf{L}_1^{(m)}, \mathsf{L}_2^{(m)}, \ldots, \mathsf{L}_n^{(m)}) = (m_1, \ldots, m_n)] = \frac{m!}{m_1! m_2! \cdots m_n!} \cdot \left(\frac{1}{n}\right)^m \tag{LHS}$$

Now let's compute the RHS. We get,

$$\mathbf{Pr}\left[(Z_1, Z_2, \ldots, Z_n) = (m_1, \ldots, m_n) \mid \sum_{i=1}^n Z_i = m\right] = \frac{\mathbf{Pr}[(Z_1, Z_2, \ldots, Z_n) = (m_1, \ldots, m_n)]}{\sum_{i=1}^n Z_i = m} \tag{5}$$

Note that the numerator event implies the denominator event and therefore we don't include it as an "and" in the numerator. Now, the $\mathbf{Pr}[Z_i = m_i] = \frac{e^{-\mu}\mu^{m_i}}{m_i!}$, and the Z_i's are independent. Therefore,

$$\mathbf{Pr}[(Z_1, Z_2, \ldots, Z_n) = (m_1, \ldots, m_n)] = \frac{e^{-n\mu}\mu^m}{m_1! m_2! \cdots m_n!}$$

Finally, by Theorem 1, $\sum_{i=1}^n Z_i$ is also a Poisson random variable with parameter $n\mu$. Therefore, $\mathbf{Pr}[\sum_{i=1}^n Z_i = m] = \frac{e^{-n\mu}(n\mu)^m}{m!}$. Plugging these into (5), we get

$$\begin{aligned}\mathbf{Pr}\left[(Z_1, Z_2, \ldots, Z_n) = (m_1, \ldots, m_n) \mid \sum_{i=1}^n Z_i = m\right] &= \frac{e^{-\mu}\mu^{m_i} \cdot m!}{e^{-n\mu}(n\mu)^m \cdot m_1! m_2! \cdots m_n!} \\ &= \frac{m!}{m_1! m_2! \cdots m_n!} \cdot \left(\frac{1}{n}\right)^m \\ &\underbrace{=}_{\text{(LHS)}} \mathbf{Pr}[(\mathsf{L}_1^{(m)}, \mathsf{L}_2^{(m)}, \ldots, \mathsf{L}_n^{(m)}) = (m_1, \ldots, m_n)]\end{aligned}$$ □

- *Completing the proof.* Now we can prove Theorem 3. In fact, one can establish more general statements than in (Gen-PA) and (Mon-PA). One can show that for *non-negative* function $f : \mathbb{Z}^n \to \mathbb{R}_{\geq 0}$, one has

$$\mathbf{Exp}[f(\mathsf{L}_1^{(m)}, \mathsf{L}_2^{(m)}, \ldots, \mathsf{L}_n^{(m)})] \quad \leq e\sqrt{m} \cdot \mathbf{Exp}[f(Z_1, Z_2, \ldots, Z_n)]$$

and if f is monotone, the $e\sqrt{m}$ can be replaced by 2. This implies the theorem since the expectation is the same as probability of occurrence for an indicator random variable. We start with the RHS:

$$\begin{aligned}
\mathbf{Exp}[f(Z_1, \ldots, Z_n)] \quad &= \sum_{k=0}^{\infty} \mathbf{Exp}[f(Z_1, \ldots, Z_n) | \sum_i Z_i = k] \cdot \mathbf{Pr}[\sum_{i=1}^{n} Z_i = k] \\
&\geq \mathbf{Exp}[f(Z_1, \ldots, Z_n) | \sum_{i=1}^{n} Z_i = m] \cdot \mathbf{Pr}[\sum_{i=1}^{n} Z_i = m] \quad \text{This uses non-negativity of } f. \\
&\underbrace{=}_{\text{Lemma 1}} \mathbf{Exp}[f(\mathsf{L}_1^{(m)}, \mathsf{L}_2^{(m)}, \ldots, \mathsf{L}_n^{(m)})] \cdot \frac{e^{-m} m^m}{m!}
\end{aligned}$$

The proof follows since $m! < e\sqrt{m}(m/e)^m$.

To replace the $e\sqrt{m}$ by 2 for *monotone* functions, one is a bit more careful with the inequality. Suppose f was monotonically increasing (non-decreasing). Then,

$$\sum_{k=0}^{\infty} \mathbf{Exp}[f(Z_1, \ldots, Z_n) | \sum_i Z_i = k] \cdot \mathbf{Pr}[\sum_{i=1}^{n} Z_i = k] \geq \mathbf{Exp}[f(Z_1, \ldots, Z_n) | \sum_{i=1}^{n} Z_i = m] \cdot \mathbf{Pr}[\sum_{i=1}^{n} Z_i \geq m]$$

because if the $\sum_i Z_i$ is larger, f is only larger. And now uses another pretty fact about Poisson random variables.

Fact 1. Let $Z \sim \text{Pois}(m)$ where m is an integer. Then $\mathbf{Med}(Z) = m$. That is, $\mathbf{Pr}[Z \geq m] \geq \frac{1}{2}$ and $\mathbf{Pr}[Z \leq m] \geq \frac{1}{2}$.

Plugging this fact into above gives the 2. Do you see how to get the 2 when f is monotonically decreasing (non-increasing)?