Poisson Approximation via Steins Method

1. on $\mathbb{N} = \{0, 1, 2, 3, \cdots\}$, the Poisson distribution is the probability measure with $\text{Poi}_\lambda(j) = \frac{e^{-\lambda}\lambda^j}{j!}, 0 \leqslant j < \infty, \lambda > 0$ is called the parameter. Familiar properties of the Poisson distribution include: if $X \sim \text{Poi}_\lambda, \mathbb{E}(X) = \lambda, \text{Var}(X) = \lambda, \mathbb{E}\left(X(X-1)\cdots(X-k+1)\right) = \lambda^k$. If $Y \sim \text{Poi}_\eta$ is independent of $X$, then $X + Y \sim \text{Poi}_{\lambda+\eta}$.

2. The Poisson heuristic, or law of "small numbers" says that if $\{X_i\}_{i\in I}$ is a collection of binary random variables with $X_i \in \{0, 1\}$, with $I$ a finite index set, and, if $P(X_i = 1) = p_i$ is "small" for all i and the $X_i$ are "not too dependent", then $\sum_i X_i = W$ has an "approximate" Poisson distribution with parameter $\lambda = \sum_i p_i$. To make this precise, we must carefully quantify "small", "not too dependent" and "approximate". We treat these in inverse order below.

3. Total variation distance
Our theorems are proved in a metric on the space of all probabilities. We introduce this abstractly. Let $(\Omega, \mathcal{F})$ be a measurable space, with $\mu$ and $\nu$ probabilities on $\mathcal{F}$. Define

$$d_{\text{TV}}(\mu, \nu) = \|\mu - \nu\|_{\text{TV}} = \sup_{A\in\mathcal{F}} |\mu(A) - \nu(A)|$$

The total variation distance is a very strong metric on measures. If $\mu$ and $\nu$ are close in total variation, they are close for most practical purposes. There are many other useful properties. When you learn what a Banach space is, you will learn that the finite signed measures on $(\Omega, \mathcal{F})$ are the duals of the Banach space of bounded measurable functions, thus $\|\mu - \nu\|_{\text{TV}}$ is just the dual norm; this is the content of b) above. One other useful fact:

$$\|\mu - \nu\|_{\text{TV}} = \inf_P P\{X \neq Y\}$$

with inf over $P$s on $\Omega \times \Omega$ with margins $\mu, \nu$.
For finite spaces this follows from the duality in the assignment problem.

4. Dependency graphs
One way to make "almost independent" precise is to use dependency graphs. Let $I$ be a finite set, $E \subseteq I \times I$ a collection of unordered edges. Write $i \sim j$

if $\{i,j\} \in E$. Let $N_i = \{i\} \cup \{j : j \sim i\}$ be the neighborhood of $i$ in $E$.
**Definition** Let $\{X_i\}_{i\in I}$ be a collection of random variables. We say $(I, E)$ is a dependency graph on $X_i$, if $I_1, I_2$ are disjoint subsets of I with no edges between them, then $\{X_i\}_{i\in I_1}$ and $\{X_i\}_{i\in I_2}$ are independent of each other. Of course, if $\{X_i\}_{i\in I}$ are independent of each other, then the empty graph works. We will see more interesting examples.

5. The main theorem
Suppose $\{X_i\}_{i\in I}$ is a finite collection of binary random variables with dependency graph $(I, E)$. Let $p_i = P(X_i = 1), p_{ij} = P(X_i = 1, X_j = 1)$. Let $\lambda = \sum_i p_i, W = \sum_i X_i$. then

$$d_{\mathrm{TV}}(W, \mathrm{Poi}_\lambda) \leqslant \min(3, \lambda^{-1})\{\sum_{i\in I} \sum_{j\in N_i\setminus\{i\}} p_{ij} + \sum_{i\in I} \sum_{j\in N_i} p_i p_j\}$$

6. Example
a) Suppose $I = \{1, 2, 3, \cdots, n\}, \lambda_i = \begin{cases} 1 & \text{with probability } \frac{1}{i} \\ 0 & \text{with probability } 1 - \frac{1}{i} \end{cases}$ The random variable $W = X_1 + \cdots + X_n$ is well known as the distribution of the number of cycles in a random permutation. We have $\lambda = 1 + \frac{1}{2} + \cdots + \frac{1}{n} = \log n - \gamma + O(\frac{1}{n}) \sim \log n$. The $X_i$ are independent, so again the first term above vanishes. The second term is $\sum_{i=1}^n \frac{1}{i^2} \sim \frac{\pi^2}{6} < \infty$, thus

$$d_{\mathrm{TV}}(W, \mathrm{Poi}(\lambda)) \leqslant \frac{\pi^2 16}{\log n}$$

One point of example b), here the lead term $\min(3, \lambda^{-1})$ saves us (gives us an error term which is small)

7. The birthday problem (with bells and whistles)
Let $[n] = \{1, 2, \cdots, n\}$ index a group of people. Let $I = \{\alpha \subset [n] : |\alpha| = k\}$ be the $k-$element subsets, so $|I| = \begin{pmatrix} n \\ k \end{pmatrix}$. Here $k \geqslant 2$. Suppose each person is given one of c colors independently with probability $\frac{1}{c}$. Let

$$X_\alpha = \begin{cases} 1 & \text{if all in } \alpha \text{ have same color} \\ 0 & \text{otherwise} \end{cases}$$

thus $p_\alpha = c^{1-k}$. Let $W = \sum_{\alpha \in I} X_\alpha$, thus $W > 0 \Leftrightarrow$ some group of $k$ people all have the same color. Thus if $k = 2, c = 365$, we have the classical birthday problem $P(W > 0) =$ chance two or more people have the same birthday. We show in a moment that the Poisson approximation is accurate if $k$ is small and $n, c$ are large. Here $\lambda = \left( \begin{array}{c} n \\ k \end{array} \right) c^{1-k}$. Thus

$$P(W = 0) \sim e^{-\lambda} = e^{-\left( \begin{array}{c} n \\ k \end{array} \right) c^{1-k}}$$

Given $k$ and $c$, we can solve this for $n$ to get $P(W = 0) = \frac{1}{2}$ or $P(W = 0) =$ .05 or whatever we want. When $k = 2, \lambda = \left( \begin{array}{c} 23 \\ 2 \end{array} \right) /365$ equals $\log(2)$ to four decimal places, so $P(\text{Birthday match out of } 23) = \frac{1}{2}$.
To use the theorem, we must choose a dependency graph. The easiest choice is to notice that $X_\alpha, X_\beta$ are independent iff $\alpha \cap \beta = \emptyset$, then $N_\alpha = \{\beta : \beta \cap \alpha \neq \emptyset\}$. Let us state the bound formally.

**Proposition** For positive integers $c, k, n$, let $W$ be the number of monochromatic $k-$tuples if an $n-$element set is colored with $c$ colors in an i.i.d. fashion. Let $\lambda = \left( \begin{array}{c} n \\ k \end{array} \right) c^{1-k}$, then

$$d_{\mathrm{TV}}(W, \mathrm{Poi}(\lambda)) \leqslant \min(3, \lambda^{-1}) \{ \left( \begin{array}{c} n \\ k \end{array} \right) \sum_{a=1}^{k-1} \left( \begin{array}{c} k \\ a \end{array} \right) \left( \begin{array}{c} n-k \\ k-a \end{array} \right) c^{1-(2k-a)} + \left( \begin{array}{c} n \\ k \end{array} \right) c^{2-2k} \sum_{a=1}^{k} \left( \begin{array}{c} k \\ a \end{array} \right) \left( \begin{array}{c} n-k \\ k-a \end{array} \right) \}$$

*proof* The first double sum in the main theorems bound is over $k-$sets $\alpha \neq \beta$ with $\alpha \cap \beta \neq \emptyset$. By symmetricity, we may fix $\alpha$ as $\{1, 2, \cdots, k\}$. An intersection with cardinality $a$ can be chosen in $\left( \begin{array}{c} k \\ a \end{array} \right) \left( \begin{array}{c} n-k \\ k-a \end{array} \right)$ ways. Then the quantity $p_{\alpha\beta} = c^{1-(k+k-a)}$ with $1 \leqslant a \leqslant k-1$. This gives the first term in the bound in the proposition. The second term is similar.

Remarks When $k = 2, \lambda = \left( \begin{array}{c} n \\ 2 \end{array} \right) /c$. This is "a number" when $n$ is of order

$\sqrt{c}$. For $n \gg \sqrt{c}$, a birthday match is near certain. For $n \ll \sqrt{c}$, a birthday match is near impossible. The chance of a match is approximately $e^{-\lambda}$. The error term in the proposition simplifies to $\min(3, \lambda^{-1}) \left[ \binom{n}{2} c^{-2}(4n-7) \right]$. If $n = \sqrt{c}$, this is of order $\frac{1}{\sqrt{c}}$. For $c = 365, n = 23, \lambda = \log 2$, to four significant figures. The error term becomes $.233\cdots$ (Using $\frac{1-e^{-\lambda}}{\lambda}$ get .055). This is one of the laments of a theoretician. Ths Poisson approximation is actually terrific, but after all our hard work we only get an error estimate of .233. When $k = 3, \lambda = \binom{\lambda}{3}/c^2$. Thus $n$ of order $c^{\frac{2}{3}}$ suffice. The bound becomes

$$\min(3, \lambda^{-1})\Big\{\Big[3\binom{n-3}{2}c^{-4} + 3\binom{n-3}{1}c^{-3}\Big] + \binom{n}{3}c^{-4}\Big[3\binom{n-3}{2} + 3\binom{n-3}{1} + 1\Big]\Big\}$$

For $c$ large and $n$ of order $c^{\frac{2}{3}}$, the bound is of order $c^{-\frac{1}{3}}$. When $n = 84, c = 365, \lambda = .7152$ and $P(\text{failure}) = e^{-\lambda} = .4891$.

There are many variations of the birthday problem — one may have a general graph, and paint vertices color $i$ with probability $c_i$ and so on. All of these may be done as the present example. See my paper with Fred Mosteller (1989 Journal of the American Statistical Association), for many further examples. we will give further Poisson examples after proving the main theorem.

8. An introduction to Steins method

Let $Z$ be a random variable with mean $\lambda$. Our proof of the main result uses a characterizing property of $Z$ often called A Stein equation: $Z \sim \text{Poi}_\lambda$ iff $\forall$ bounded $f : \mathbb{N} \to \mathbb{R}$

$$(*)\mathbb{E}(\lambda f(Z+1) - Zf(Z)) = 0$$

It is easy to show that if $Z \sim \text{Poi}_\lambda$, then $*$ is satisfied. It will follow from what we prove below, that if $*$ is satisfied for many functions $f$, then $Z \sim \text{Poi}_\lambda$. The essence of Steins method for showing that $W$ has appproximate $\text{Poi}_\lambda$ distribution is to show for every bounded $f, |\mathbb{E}(\lambda f(W+1) - Wf(W))|$ is small.

To see this more sharply, we will prove

Let $A \subseteq \mathbb{N}$, for $\lambda > 0$, let $Z \sim \text{Poi}_\lambda, \exists! f : \mathbb{N} \to \mathbb{R}$ with $f(0) = 0, f$ bounded such that $\forall \omega$ :

$$\lambda f(\omega + 1) - \omega f(\omega) = \delta_A(\omega) - \text{Poi}_\lambda(Z \in A) \ast \ast$$

Further $|f(k+1) - f(k)| \leqslant \min(3, \lambda^{-1})$ for all $k \in \mathbb{N}$.
Of course, $\ast\ast$ makes the converse of $\ast$ transparent: if $W$ is a random variable with $\mathbb{E}(\lambda f(W+1) - Wf(W)) = 0$ for all bounded $f$, then choose $f$ as in $\ast\ast$ we have

$$0 = \mathbb{E}(\lambda f(W+1) - Wf(W)) = P\{W \in A\} - \text{Poi}_\lambda(Z \in A)$$

So $W$ is Poisson. In the next section, we show that $\ast\ast$ implies the main theorem. In the following section we prove $\ast\ast$.

9. Proof of main theorem using $\ast\ast$. Let $W = \sum_i X_i$ be as in section 5, based on a dependency graph. For $A \subseteq \mathbb{N}$, let $f$ be chosen as in $\ast\ast$, then $P\{Z \in A\} - P\{W \in A\} = \mathbb{E}[Wf(W) - \lambda f(W+1)] = \sum_i \mathbb{E}[X_i f(W) - p_i f(W+1)] = \Delta$.
Let $W_i = W - X_i, V_i = \sum_{j \notin N_i} X_j$. Thus $X_i f(W) = X_i f(W_i + 1)$. Thus, by independence of $X_i$ and $V_i$,

$$\begin{aligned}\Delta &= \sum_i \mathbb{E}\left[(X_i - p_i) f(W_i+1)\right] + p_i \mathbb{E}\left[f(W_i+1) - f(W+1)\right] \\ &= \sum_i \mathbb{E}\left[(X_i - p_i)(f(W_i+1) - f(V_i+1)\right] + p_i \mathbb{E}\left[f(W_i+1) - f(W+1)\right]\end{aligned}$$

By $\ast\ast$, $|f(W_i+1) - f(W+1)| \leqslant \min(3, \lambda^{-1}) X_i$, so $|p_i \mathbb{E}\left[f(W_i+1) - f(W+1)\right]| \leqslant \min(3, \lambda^{-1}) p^2$.
Further, $f(W_i+1) - f(V_i+1)$ can be written as a telescoping sum of terms of form $f(u + X_i) - f(u)$, each bounded in modulus by $\min(3, \lambda^{-1}) X_i$. Thus,

$$\begin{aligned}|\mathbb{E}\left[(X_i - p_i)(f(W_i+1) - f(V_i+1))\right]| &\leqslant \mathbb{E}\left[|X_i - p_i| \sum_{j \in N_i \setminus i} \min(3, \lambda^{-1}) X_j\right] \\ &\leqslant \min(3, \lambda^{-1}) \sum_{j \in N_i \setminus i} (p_{ij} + p_i p_j)\end{aligned}$$

Combining estimates gives the conclusion of the main theorem.

proof of $**$: Let $A \subseteq \{0,1,2,3,\cdots\}$, we prove that $\exists! f : \{0,1,2,\cdots\} \to \mathbb{R}$ with $f(0) = 0$ and such that for all $\omega, \lambda f(\omega+1) - \omega f(\omega) = \delta_A(\omega) - \text{Poi}_\lambda(A)$. Further we show $|f(\omega)| \leqslant 1.25$ and $|f(\omega+1) - f(\omega)| \leqslant \min(3, \lambda^{-1}$.
Note first that if $f(0)$ is set to any fixed value, then $f(\omega)$ is uniquely determined inductively from $\lambda f(\omega+1) = \omega f(\omega) + \delta_A(\omega) - \text{Poi}_\lambda(A)$. We take $f(0) = 0$, but this choice doesn't enter any of our uses of $f$. The harder part is to bound $f$ and its first difference. We do this by explilcitly writing out $f$, showing that $|f| \leqslant 1.25$ and so $|f(\omega+1) - f(\omega)| \leqslant 3$. Finally we show that the first difference is bounded by $\lambda^{-1}$. The proof is slightly tedious but easy with more work, barbour and eagleson (1983). Poisson approximation for some statistics based on exchangable trials, advanced applied probability. 15 585-600, have shown $|f(\omega)| \leqslant \min(1, 1.4\lambda^{-\frac{1}{2}}), |f(\omega+1) - f(\omega)| \leqslant \frac{1-e^{-\lambda}}{\lambda}$.

a) Representing $f$. Since $f(0) = 0, \lambda f(1) = \delta_A(\omega) - \text{Poi}_\lambda(A)$, multiply the basic recurrence by $\frac{\lambda^\omega}{\omega!}$ to get $\frac{\lambda^{\omega+1} f(\omega+1)}{\omega!} - \frac{\lambda^\omega f(\omega)}{(\omega-1)!} = \frac{\lambda^\omega}{\omega!}(\delta_A(\omega) - \text{Poi}_\lambda(A))$. Sum from 1 to $k-1$:

$$f(k) = \frac{(k-1)!}{\lambda^k} \sum_{\omega=0}^{k-1} \frac{\lambda^\omega}{\omega!}(\delta_A(\omega) - \text{Poi}_\lambda(A))$$

Since $\sum_{\omega=0}^{\infty} \frac{\lambda^\omega}{\omega!}(\delta_A(\omega) - \text{Poi}_\lambda(A)) = 0$. This is can also be written

$$f(k) = -\frac{(k-1)!}{\lambda^k} \sum_{\omega=k}^{\infty} \frac{\lambda^\omega}{\omega!}(\delta_A(\omega) - \text{Poi}_\lambda(A))$$

b) First bounds on $f$. We have $|\delta_A(\omega) - \text{Poi}_\lambda(A)| \leqslant 1$. The two forms of $f(k)$ in a) give

$$|f(k)| \leqslant \frac{1}{\lambda} \sum_{m=0}^{k-1} \frac{(k-1)!}{\lambda^m (k-1-m)!} \leqslant \frac{1}{\lambda} \sum_{m=0}^{\infty} (\frac{k-1}{\lambda})^m = (\lambda - (k-1))^{-1}$$

$$|f(k)| \leqslant \sum_{m=0}^{\infty} \frac{\lambda^m (k-1)!}{(k+m)!} \leqslant \sum_{m=0}^{\infty} \frac{\lambda^m}{k(k+1)^m} = \frac{k+1}{k(k+1-\lambda)}$$

Use the first bound for $k \leqslant \lambda + \frac{1}{5}$, and the second for $k \geqslant \lambda + \frac{1}{5}, |f(k)| \leqslant 1.25$. For $k = 2$. Finally, for $k = 1, f(1) = \lambda^{-2}(\delta_A(0) - \text{Poi}_\lambda(A))$. This is maximized when $A = \emptyset$, and minimized when $A = \{1,2,3,\cdots\}$. Thus

$|f(1)| \leqslant \lambda^{-1}(1-e^{-\lambda}) < 1$. This proves $|f(k)| \leqslant 1.25$ and $|f(k+1)-f(k)| \leqslant 3$ for all $k$.

Added note
For an up to date review of Steins method for Poisson approximation, see S. Chatterjee, P. Diaconis, E. Meckes (2005) Exchangeable pairs and Poisson approximation. Probability surveys 2, 64-106.
Notes on literature
Steins method is part of a general approach for deriving approximations for sums of dependent random variables. It is mostly used for Poisson and normal approximation. But multivariate and compound Poisson variants are also available. One hallmark of the method: it comes with an error term.
For Poisson approximation, the method is sometimes called the chen-stein method. There are actually three "applications'. The one we have done in class using dependency graphs is developed and illustrated in Arratia, Goldstein, Gordon (1990) "Poisson approximation and the Chen-Stein method" Statistical Science 5 403-434. This is a very readable article which has many examples; to longest head runs, DNA string matcheing, permutations with restricted position and much else. It has useful extensions of the dependency graph approach not covered in class. Mathew Penrose "Random Geometric Graphs" makes nice use of these tools.
The second approach (to Steins method for Poisson approximation) is Barbours coupling approach. A book-length treatment appears in Barbour, Holst and Janson (1992) "Poisson Approximation" Oxford. This is more technical but has hundreds of worked examples.
The third approach is Steins own method of exchangeable pairs. This is a very general approach which you can find in Steins book "Approximate Computation of Expectations (1986);;. I find this the easiest to use but it has not been packaged as neatly as the dependency graph approach and is harder to teach. Steins book gives the general version of "Steins method" for any distribution. It is one of the great achievements of $20^{\text{th}}$ century probability. All three variations use the characterization of Poissons by $\mathbb{E}(\lambda f(W+1) - Wf(W)) = 0$. All three depend on $**$.

Assignment

**Problem 1**
a) Show that $d_{\text{TV}}(\mu,\nu) = d_{\text{TV}}(\nu,\mu), d_{\text{TV}}(\mu,\nu) \leqslant d_{\text{TV}}(\mu,\eta) + d_{\text{TV}}(\eta,\nu)$ and

$d_{\mathrm{TV}}(\mu,\nu) = 0 \Leftrightarrow \mu = \nu$, so $d_{\mathrm{TV}}$ is a metric.
b) Show that $d_{\mathrm{TV}}(\mu,\nu) = \frac{1}{2}\sup_{\|f\|_\infty \leqslant 1} |\mathbb{E}_\mu(f) - \mathbb{E}_\nu(f)|$ with $\|f\|_\infty = \sup|f(\omega)|$, and $\mathbb{E}_\mu(f) = \mathbb{E}_\mu(f(X))$ with $X \sim \mu$.
c) Let $\eta$ be a $\sigma$-finite measure on $(\Omega, \mathcal{F})$. Let $\mu$ and $\nu$ have densities $f_\mu$ and $f_\nu$ with respect to $\eta$ (So e.g., $\mu(A) = \int_A f_\mu(\omega)\lambda(\mathrm{d}\omega)$. Show that

$$\|\mu - \nu\|_{\mathrm{TV}} = \frac{1}{2}\int |f_\mu(\omega) - f_\nu(\omega)|\lambda(\mathrm{d}\omega)$$

**Problem 2** Picture $n$ boys and $n$ girls connected in a complete bipartite graph. "Color" the boys and girls with $c$ colors in an i.i.d. fashion. Assuming $c$ is large (e.g. c = 365). How large does $n$ have to be to have probability $\frac{1}{2}$ that some boy and some girl are given the same color? Do this by proving $n$ careful Poisson approximations.

**Problem 3** Show $|f(k+1) - f(k)| \leqslant \lambda^{-1}$ for $\lambda \geqslant \frac{1}{3}$.

**Problem 3.5** Let $X_0, X_1, \cdots, X_n$ be positive random variables. A test for "clumping" will be based on $Y_i = \begin{cases} 1 & \text{if } |X_{i+1} - X_i| > \varepsilon \\ 0 & \text{otherwise} \end{cases}$ $1 \leqslant i \leqslant n$. Let $W = \sum_1^n Y_i$. Suppose, in fact, that $X_i$ are i.i.d. with $P(X_i > t) = e^{-t}$. For $n$ large, find $\varepsilon = \varepsilon(n)$ so $W$ has a non trivial limit distribution.

**Problem 4** Pick $\omega \in S_n$ uniformly. Let

$$W(\omega) = \#\{i : \omega(i+1) = \omega(i) + 1\}$$

Find good approximation for $P(W = j)$ (and prove your answer)
For proof, look in paper by A?-Gorden-Goldstein (Statistical Science). Find a version of Steins method that fits and check condition.