Math/Stat 394
F.W. Scholz

Poisson-Binomial Approximation

Theorem 1: Let X_1 and X_2 be independent Poisson random variables with respective parameters $\lambda_1 > 0$ and $\lambda_2 > 0$. Then $S = X_1 + X_2$ is a Poisson random variable with parameter $\lambda_1 + \lambda_2$.
Proof:

$$\begin{aligned}
P(X_1 + X_2 = z) &= \sum_{i=0}^{\infty} P(X_1 + X_2 = z, X_2 = i) = \sum_{i=0}^{\infty} P(X_1 + i = z, X_2 = i) \\
&= \sum_{i=0}^{z} P(X_1 = z - i, X_2 = i) = \sum_{i=0}^{z} P(X_1 = z - i)\, P(X_2 = i) = \sum_{i=0}^{z} \frac{e^{-\lambda_1}\lambda_1^{z-i}}{(z-i)!}\frac{e^{-\lambda_2}\lambda_2^{i}}{i!} \\
&= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^z}{z!} \sum_{i=0}^{z} \frac{z!}{i!(z-i)!} \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^i \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{z-i} = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^z}{z!}
\end{aligned}$$

Corollary: If $X_1, \ldots, X_n$ are independent Poisson random variables with respective parameters $\lambda_1, \ldots, \lambda_n$ then $S = X_1 + \ldots + X_n$ is a Poisson random variable with parameter $\lambda_1 + \ldots + \lambda_n$.
Proof: By induction over n.

Lemma 1: Let S and T be two random variables with some joint distribution for (S, T). Then

$$|P(S \in A) - P(T \in A)| \le P(S \ne T).$$

Proof:

$$\begin{aligned}
P(S \in A) &= P(S \in A, S = T) + P(S \in A, S \ne T) = P(T \in A, S = T) + P(S \in A, S \ne T) \\
&= P(T \in A, S = T) + P(T \in A, S \ne T) - P(T \in A, S \ne T) + P(S \in A, S \ne T) \\
&= P(T \in A) - P(T \in A, S \ne T) + P(S \in A, S \ne T)
\end{aligned}$$

Thus

$$P(S \in A) - P(T \in A) = P(S \in A, S \ne T) - P(T \in A, S \ne T) \le P(S \in A, S \ne T) \le P(S \ne T)$$

and similarly $P(T \in A) - P(S \in A) \le P(S \ne T)$, thus $|P(T \in A) - P(S \in A)| \le P(S \ne T)$
Lemma 2: Let $S = X_1 + \ldots + X_n$ and $T = Y_1 + \ldots + Y_n$ then

$$P(S \ne T) \le \sum_{i=1}^{n} P(X_i \ne Y_i).$$

Proof: First note that $S \ne T$ implies that at least one pair of summands (X_i, Y_i) must differ, i.e., $X_i \ne Y_i$ (otherwise the sums would agree). Thus

$$\{S \ne T\} \subset \bigcup_{i=1}^{n} \{X_i \ne Y_i\}.$$

The inclusion inequality followed by Boole's inequality yields

$$P(S \ne T) \le P\left(\bigcup_{i=1}^{n} \{X_i \ne Y_i\}\right) \le \sum_{i=1}^{n} P(X_i \ne Y_i)$$

Lemma 3: For $p \leq .8$ let (X, Y) have a joint probability distribution given by

$y \setminus x$	0	1	2	3	$\ldots$	$P(Y = y)$
0	$e^{-p}(1+p) - p$	0	$p^2 e^{-p}/2!$	$p^3 e^{-p}/3!$	$\ldots$	$1 - p$
1	$p - pe^{-p}$	pe^{-p}	0	0	$\ldots$	p
$P(X = x)$	e^{-p}	$pe^{-p}/1!$	$p^2 e^{-p}/2!$	$p^3 e^{-p}/3!$	$\ldots$	1

Note that $g(g) = e^{-p}(1+p) - p$ has derivative $g'(p) = -e^{-p}p - 1 < 0$ and $g(.8) = .00879 > 0$, thus all tabled probabilities are ≥ 0 when $p \geq .8$.

Then Y is a Bernoulli random variable with parameter p, i.e., $P(Y = 1) = 1 - P(Y = 0) = p$ and X is a Poisson random variable with parameter $\lambda = p$. Further, $P(X \neq Y) \leq 2p^2$.

Proof: The marginal distributions of X and Y are evident from the table and

$$\begin{aligned} P(X \neq Y) &= 1 - P(X = Y) = 1 - P(X = Y = 0) - P(X = Y = 1) \\ &= 1 - e^{-p}(1+p) + p - pe^{-p} = 1 + p - (1+2p)e^{-p} \\ &\leq 1 + p - (1+2p)(1-p) = 2p^2 \quad \text{using } e^{-p} \geq 1 - p \text{ for all } p \end{aligned}$$

Theorem 2: Let $Y_1, \ldots, Y_n$ be independent Bernoulli RVs with $P(Y_i = 1) = 1 - P(Y_i = 0) = p_i$, respectively. Let $T = Y_1 + \ldots + Y_n$ and let S be a Poisson random variable with parameter $\lambda = p_1 + \ldots + p_n$. Then

$$|P(S \in A) - P(T \in A)| \leq 2 \sum_{i=1}^{n} p_i^2 \qquad \text{for all sets } A$$

Proof: For each i let (X_i', Y_i') have the joint distribution stipulated in Lemma 3 with $p = p_i$. Do this independently for $i = 1, \ldots, n$, i.e., (X_1', Y_1'), $\ldots$, (X_n', Y_n') are independent pairs. Then Y_i' has the same distribution as Y_i and thus $T' = Y_1' + \ldots + Y_n'$ has the same distribution as $T = Y_1 + \ldots + Y_n$. According to the corollary to Theorem 1 we have that $S = X_1' + \ldots + X_n'$ has a Poisson distribution with parameter $\lambda = p_1 + \ldots + p_n$. Now chaining Lemmas 1, 2 and 3 we get

$$|P(S \in A) - P(T \in A)| = |P(S' \in A) - P(T' \in A)| \leq 2 \sum_{i=1}^{n} p_i^2$$

Comment 1: The above proof is due to Hodges and Le Cam (1960). With more work, the factor 2 in the error bound can be dropped. In fact, by fairly elementary steps it is possible to replace 2 by $\min(1, \lambda^{-1})$, where $\lambda = p_1 + \ldots + p_n$, see Barbour et al. (1992), p. 8. The advantage in the factor $\min(1, \lambda^{-1})$ is that the bound gets smaller the more of the p_i are added up. In some sense we do not just account for the error made within each pair (X_i, Y_i), i.e., for $P(X_i \neq Y_i)$, but we also take advantage that there is error cancellation across the sums, i.e., some of the $X_i \neq Y_i$ cancel to some extent when assessing $S \neq T$.

Another type of error bound is as follows (without proof):

$$|P(S \in A) - P(T \in A)| \leq 9 \max(p_1, \ldots, p_n) \, .$$

Further improvements (Arratia et al. (1990)) relax the independence conditions on the Bernoulli trials.

Comment 2: The distribution of T, often called the Poisson-Binomial distribution, depends on the parameter vector $(p_1, \ldots, p_n)$ and is rather complicated. It can be approximated by a simple Poisson distribution, depending only on the single parameter $\lambda = p_1 + \ldots + p_n$. This approximation is accurate provided all the p_i are quite small so that the bound $2(p_1^2 + \ldots + p_n^2)$ on the approximation error is sufficiently small.
Comment 3: In the special case, when $p_1 = \ldots = p_n = p$, T has the well known binomial distribution and page 144 of Anderson et al (2018) gives a limiting argument for the Poisson approximation to a binomial distribution under the assumption that $p = p_n \to 0$ as $n \to \infty$ so that $np_n \approx \lambda > 0$. This approximation falls out easily from Theorem 2, since under these assumptions

$$2\sum_{i=1}^{n} p_i^2 = 2np^2 \approx 2\frac{\lambda^2}{n} \to 0$$

as $n \to \infty$. The limiting argument does not tell us how good the approximation is when used for a finite n. Theorem 2 gives us a crude but nevertheless useful bound on the approximation error. See the following example for a concrete application.

Example: Suppose during a particular minute of the day the $n = 2000000$ people serviced in a particular telephone service area decide independently of each other whether to place an emergency call to 911 or not. Each person has his/her own probability p_i of doing so. Suppose that the average probability for all n persons is about $\overline{p} = (p_1 + \ldots + p_n)/n = .000005$, i.e., on average about 10 persons makes such a call in that minute. Suppose further that these p_i never exceed .00001. Let T be the actual random number of 911 callers in that minute. Its exact distribution is extremely complicated, depending on $n = 2000000$ parameters $p_1, \ldots, p_n$. Using the approximation error bound of Theorem 2 and $\lambda = n\overline{p} = 10$ we find

$$2\sum_{i=1}^{n} p_i^2 \leq 2\max(p_1, \ldots, p_n)\sum_{i=1}^{n} p_i \leq 2 \cdot .00001 \cdot \lambda = .0002.$$

Actually this bound on the error of computing $P(T \in A)$ instead of $P(S \in A)$ is quite exaggerated as can be seen by the many inequalities that were employed in proving Theorem 2. Thus the Poisson approximation is much more trustworthy than it looks. When using $\min(1, \lambda^{-1})$ instead of the factor 2 our error bound for the approximation becomes .00001.

The Poisson Process Derivation

Poisson distribution for counts of random incidents in time:

Sometimes we observe random incidents occurring in time, e.g. arrival of customers, meteoroids, lightning etc. (I use the word *incident*, instead of *event*, to avoid confusion with our other usage of the term event in probability theory.) Quite often these random phenomena appear to satisfy the following basic assumptions for some positive constant λ:

1. The probability that exactly one incident occurs during a short time interval of length h is approximately proportional to h (with proportionality factor λ) , i.e., that probability is $\lambda h + o(h)$ where $o(h)$ is a function of h which goes to 0 faster than h, i.e. $o(h)/h \to 0$ as $h \to 0$ (e.g. $o(h) = h^2$).

2. The probability that two or more events occur in a short time interval of length h is negligible, i.e., equal to $o(h)$.

3. For any integers n and nonoverlapping time intervals $I_1, \ldots, I_n$ any events $E_1, \ldots, E_n$ pertaining to the separate counts of incidents occurring in these intervals are independent.

If $N(I)$ denotes the random number of incidents in the time interval I. Then we can rephrase the above postulates as follows:

1.

$$P(N((t, t+h]) = 1) = \lambda h + o(h) \quad \text{for any} \;\; h > 0 \,, \;\; t \geq 0 \,.$$

2.

$$P(N((t, t+h]) \geq 2) = o(h) \quad \text{for any} \;\; h > 0 \,, \;\; t \geq 0 \,.$$

3. For any integers $n \geq 1$, $k_1 \geq 0, \ldots, k_n \geq 0$ and $s_1 \leq t_1 \leq s_2 \leq t_2 \leq \ldots \leq s_n \leq t_n$ we have

$$P(N((s_1, t_1]) = k_1, \; N((s_2, t_2]) = k_2, \; \ldots, \; N((s_n, t_n]) = k_n)$$

$$= P(N((s_1, t_1]) = k_1) \cdot P(N((s_2, t_2]) = k_2) \cdot \ldots \cdot P(N((s_n, t_n]) = k_n) \,,$$

i.e., independence of the incident counts.

Under these postulates it follows that $N((s, s+t])$ is a Poisson random variable with parameter λt, i.e.,

$$P(N((s, s+t]) = k) = \frac{\exp(-\lambda t)(\lambda t)^k}{k!} \,. \tag{1}$$

Proof: Without loss of generality assume $s = 0$ and hence take as time interval $(0, t]$ and divide it into n equal length subintervals $I_1, \ldots, I_n$. Then the event $C = \{N((0, t]) = k\}$ can be written as the disjoint union of the following two events:

$$A = \{k \text{ of the intervals contain exactly one incident and } n-k \text{ contain 0 incidents}\}$$

and

$$B = \{N((0, t]) = k \text{ and at least one subinterval contains two or more incidents}\}$$

Hence

$$P(C) = P(A) + P(B) \,. \tag{2}$$

If we denote $E_i = \{N(I_i) \leq 1\}$ then the second postulate above says $P(E_i^c) = o(t/n)$.
Since the event B implies $\cup_{i=1}^n E_i^c$, i.e., $B \subset \cup_{i=1}^n E_i^c$, it follows that

$$P(B) \leq P\left(\cup_{i=1}^n E_i^c\right) \leq \sum_{i=1}^n P(E_i^c) \leq n \cdot o(t/n) \to 0 \;\; \text{as} \;\; n \to \infty \,.$$

using Boole's inequality for the second $\leq$ above. It remains to show that

$$P(A) \;\to\; \frac{\exp(-\lambda t)(\lambda t)^k}{k!} \quad \text{as} \;\; n \to \infty \,.$$

Since the left side of (2) does not depend on n it does not matter into how many subintervals we divide $(0, t]$. Hence we can freely let n become arbitrarily large and this would prove the claim (1).

Before doing this final step we prove some preliminaries using the independence assumptions from postulate 3:

$$\begin{aligned} P\left(N(I_1)=k_1,\ \ldots,\ N(I_n)=k_n \mid \cap_{i=1}^n E_i\right) &= \frac{P\left(\{N(I_1)=k_1\}\cap E_1\cap\ \ldots\ \cap\{N(I_n)=k_n\}\cap E_n\right)}{P\left(\cap_{i=1}^n E_i\right)} \\ &= \frac{P\left(\{N(I_1)=k_1\}\cap E_1\right)\cdot\ldots\cdot P\left(\{N(I_n)=k_n\}\cap E_n\right)}{P\left(E_1\right)\cdot\ldots\cdot P\left(E_n\right)} \\ &= P\left(\{N(I_1)=k_1\}|E_1\right)\cdot\ \ldots\ \cdot P\left(\{N(I_n)=k_n\}|E_n\right)\ . \end{aligned}$$

Further, again using the independence assumptions from postulate 3:

$$\begin{aligned} P\left(\{N(I_1)=k_1\}|E_1\right) &= \frac{P\left(\{N(I_1)=k_1\}\cap E_1\right)}{P\left(E_1\right)} = \frac{P\left(\{N(I_1)=k_1\}\cap E_1\right)P(E_2)\ldots P(E_n)}{P(E_1)\cdot P(E_2)\cdot\ldots\cdot P(E_n)} \\ &= \frac{P\left(\{N(I_1)=k_1\}\cap E_1\cap\ldots\cap E_n\right)}{P\left(E_1\cap\ldots\cap E_n\right)} = P\left(\{N(I_1)=k_1\}|\cap_{j=1}^n E_j\right) \end{aligned}$$

and similarly for $i>1$:

$$P\left(\{N(I_i)=k_i\}|E_i\right) = P\left(\{N(I_i)=k_i\}|\cap_{j=1}^n E_j\right)\ .$$

Note also that given $\cap_{j=1}^n E_j$ the random variables $N(I_i)$ can only take on the values 0 and 1 and are conditionally independent, which follows from putting the above two preliminaries together. Hence conditionally we deal with independent Bernoulli trials with success probabilities

$$\begin{aligned} p_{i,n} &= P\left(N(I_i)=1|\cap_{j=1}^n E_j\right) = P(N(I_i)=1|E_i) = \frac{P(N(I_i)=1\cap E_i)}{P(E_i)} \\ &= \frac{P\left(N(I_i)=1\right)}{P(E_i)} = \frac{\lambda t/n+o(t/n)}{1-P(E_i^c)} = \frac{\lambda t/n+o(t/n)}{1-o(t/n)} \quad \text{with} \quad \sum_{i=1}^n p_{i,n}\to\lambda t \quad \text{as } n\to\infty \end{aligned}$$

Next note that with $F=\cap_{i=1}^n E_i$ we have

$$P(A) = P(A|F)P(F) + P(A|F^c)P(F^c)\ .$$

We pointed out above that $P(F^c) = P(\cup_{i=1}^n E_i^c)\to 0$ as $n\to\infty$ and thus also $P(F)\to 1$ as $n\to\infty$. Therefore $P(A)$ and $P(A|F)$ have the same limit. But with the preliminaries shown above the conditional probability $P(A|F)$ is the Poisson-binomial probability for exactly k successes in n independent trials with success probabilities $p_{i,n}, i=1,\ldots,n$, which converges to the desired Poisson probability (with mean λt) by the Poisson-binomial approximation result. This concludes the proof.

References:

Arratia, R., Goldstein, L., and Gordon, L. (1990), "Poisson Approximation," *Statistical Science*, Vol. 5, No. 4, 403-424.

Barbour, A.D., Holst, L. and Janson, S. (1992), *Poisson Approximation*, Clarendon Press, Oxford.

Hodges, J.L., Jr. and Le Cam, L. (1960), "The Poisson Approximation to the Poisson Binomial Distribution," *Ann. Math. Statist.* 31, 737-740.

Anderson, D.F., Seppäläinen, T., and Valkó, B. (2018), *Introduction to Probability* , Cambridge University Press, Cambridge, UK.

Serfling, R.J. (1978), "Some elementary results on Poisson approximation in a sequence of Bernoulli trials," *Siam Review*, Vol. 20, No. 3, 567-579.