The Poisson distribution

We have delayed talking about this distribution because its definition involves mention of expectation, which was therefore covered first.

Example 19 At a certain traffic intersection, two accidents occur each year, on average.

Find the probability that 4 accidents will occur at this intersection next year, supposing that either 0 or 1 accidents can occur in each period of:

(a) a month

(b) a week

(c) a day.

Assume that accidents occur independently from period to period.

? $n \; p = \mu$

(a) Let Y = number of accidents per year.

Then $Y \sim \text{Bin}(12, 2/12)$. (Then, $EY = 12(2/12) = 2$.)

So $P(Y=4)=\binom{12}{4}\left(\frac{2}{12}\right)^4\left(\frac{10}{12}\right)^8=0.0888$.

$n \; p = \mu$

(b) In this case, $Y \sim \text{Bin}(52, 2/52)$. (Then, $EY = 52(2/52) = 2$.)

So $P(Y=4)=\binom{52}{4}\left(\frac{2}{52}\right)^4\left(\frac{50}{52}\right)^{48}=0.0902$.

$n \quad p$

(c) Now, $Y \sim \text{Bin}(365, 2/365)$. (Then, $EY = 365(2/365) = 2$.)

So $P(Y=4)=\binom{365}{4}\left(\frac{2}{365}\right)^4\left(\frac{363}{365}\right)^{361}=0.0902$ (same, to 4 decimals).

A limit appears to have been reached.

For example, we would get the same answer (to 4 decimals) if we divided the year up into $365 \times 24 = 8760$ hours and worked on the assumption that $Y \sim \text{Bin}(8760, 2/8760)$. In that case, $P(Y=4)=\binom{8760}{4}\left(\frac{2}{8760}\right)^4\left(\frac{8758}{8760}\right)^{8756}=0.0902.$

More generally, suppose that 0 or 1 accidents can occur in each of n time periods, and that the total expected number of accidents is fixed at 2. Then in the limit as n tends to infinity, we find that the required probability can be written

$$P(Y=4)=\boxed{\frac{e^{-2}2^4}{4!}}=0.0902.$$

fixed

Proof: Let $Y \sim \text{Bin}(n,2/n)$. $\quad \mu = np = n\times\left(\frac{2}{n}\right) = \boxed{2}$

$$\text{Then } P(Y=4)=\boxed{\binom{n}{4}\left(\frac{2}{n}\right)^4\left(1-\frac{2}{n}\right)^{n-4}}$$

$$=\boxed{\frac{n!}{4!(n-4)!}}\frac{2^4}{n^4}\left(1-\frac{2}{n}\right)^n\left(1-\frac{2}{n}\right)^{-4}$$

$\to 1$ as $n \to \infty$

$$=\left\{\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\left(\frac{n-3}{n}\right)\frac{(n-4)!}{(n-4)!}\right\}\boxed{\frac{2^4}{4!}}\boxed{\left(1-\frac{2}{n}\right)^n}\boxed{\left(1-\frac{2}{n}\right)^{-4}}$$

$$\to 1\times\boxed{\frac{2^4}{4!}}\times\boxed{e^{-2}}\times 1 \quad \text{as } n\to\infty. \qquad \left(\text{Note that } e^x=\lim_{m\to\infty}\left(1+\frac{x}{m}\right)^m.\right)$$

$$\text{Similarly, } P(Y=3)=\boxed{\frac{e^{-2}2^3}{3!}}, \text{ etc.}$$

We say that Y has a *Poisson* distribution (with parameter 2).

When it makes sense to take a limit as here, we say that a *Poisson process* is involved.

A random variable Y has the *Poisson distribution* with parameter λ if its pdf is of the form

pmf

$$f(y)=\boxed{\frac{e^{-\lambda}\lambda^y}{y!}},\quad y=0,1,2,\dots \quad (\lambda\ge 0).$$

We write $Y\sim\text{Poi}(\lambda)$ and $f(y)=f_{\text{Poi}(\lambda)}(y)$.

Summarising, $f_{\text{Poi}(\lambda)}(y)=\lim\limits_{\substack{n\to\infty\\ \lambda=np}} f_{\text{Bin}(n,p)}(y)$.

Let's verify that the mean of the Poisson distribution with parameter λ is λ:

$$EY=\sum_{y=0}^{\infty}y\frac{e^{-\lambda}\lambda^y}{y!}=\lambda e^{-\lambda}\sum_{y=1}^{\infty}\frac{\lambda^{y-1}}{(y-1)!}=\lambda e^{-\lambda}\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}=\lambda e^{-\lambda}e^{\lambda}=\boxed{\lambda}.$$

(✓)

$EY = VY = \lambda$

Exercise: Show that the variance of the Poisson distribution with parameter λ is also λ. (Hint: Either use the mgf technique, or first find $E\{Y(Y-1)\}$.)

$= \sum_{y=0}^{\infty} y(y-1)\frac{e^{-\lambda}\lambda^y}{y!} = \dots = E(Y^2 - Y)$

$EY^2 = E(Y^2 - Y) + EY$

$VY = EY^2 - (EY)^2$

$= \dots - \dots$

$= \boxed{} = \lambda$

The Poisson distribution is useful for modelling all sorts of count data, not just traffic accident frequencies.

Example 20 A total of 27 defects were found in a 1 km coil of thread.

Find the probability (approximately) that a 10 m coil of the same type of thread (taken randomly from another coil) will have at least one defect.

Let Y be the number of defects in the 10 m coil.
Then (approximately), $Y \sim \text{Poi}(\lambda)$, where $\lambda = 0.27$.

"model"

(Each 1 metre length of the 1 km coil has 27/1000 = 0.027 defects on average.
So the expected number of defects in the 10 m coil is (about) $0.027 \times 10 = 0.27$.)

Thus $P(Y \geq 1) = 1 - P(Y = 0) = 1 - \boxed{\frac{e^{-0.27}0.27^0}{0!}} = 1 - e^{-0.27} = 1 - 0.763 = 0.237$.

Note: We have assumed that the number of defects in the thread follows a *Poisson process*. This means that it is reasonable to consider the rope as being comprised of a very large number of tiny lengths, such that:

(a) each length can have either 0 or 1 defects (but not 2 or 3, say)
(b) each length has the same chance of containing a defect
(c) whether a certain length has a defect is independent of whether any other lengths have a defect.

We could test the Poisson assumption by counting the number of defects in each of the one hundred 10 m lengths that make up the 1 km coil. This would result in a sequence of 100 numbers like 0,0,1,0,2,0,0,0,0,1,....,0,1,0,0.

If the Poisson assumption is justified, about 24 (23.7%) of these numbers will be nonzero. Also, the number of 2's will be about $\frac{e^{-0.27}0.27^2}{2!} \times 100 = 2.78$ (ie, 3 or so), etc. If we find large discrepancies between the expected and observed numbers, then there will be good reason to doubt the Poisson assumption.

The Poisson approximation to the binomial

As a consequence of the definition of the Poisson dsn in terms of the binomial dsn, the $Bin(n,p)$ can be approximated by the $Poi(\lambda = np)$ when n is 'large' and p is 'small'.

Example: We roll 2 dice together 30 times. Find the pr. that one double six comes up.
Solution: Number of double sixes = $Y \sim \text{Bin}(30,1/36) \approx \text{Poi}(5/6)$
(since $30(1/36) = 5/6$).

So $P(Y=1) \approx \boxed{\dfrac{e^{-5/6}(5/6)^1}{1!}} = 0.3622$. (The exact prob. is $\boxed{\binom{30}{1}\dfrac{1}{36}\left(\dfrac{35}{36}\right)^{29}} = 0.3681$.)

Example 2: We roll 3 dice together 30 times. Find the pr. that one triple six comes up.
Solution: Y = number of triple sixes $\sim \text{Bin}(30,1/216) \approx \text{Poi}(5/36)$
(since $30(1/216) = 5/36$).

So $P(Y=1) \approx \dfrac{e^{-5/36}(5/36)^1}{1!} = 0.1209$. (The exact pr is $\binom{30}{1}\dfrac{1}{216}\left(\dfrac{215}{216}\right)^{29} = 0.1214$.)

We see that the approximation here improves as p gets smaller. Generally, the Poisson approximation should be considered only when the exact binomial probability is hard or impossible to calculate. As a rule of thumb, the Poisson approximation is 'good' if n is at least 20 and p is at most 0.05, or if n is at least 100 and np is at most 10.

Tchebysheff's theorem (or Chebyshev's theorem)

Let Y be a rv with mean μ and variance σ^2 (assumed to be finite).
Also let k be a positive constant. Then

$$\boxed{P(|Y-\mu| < k\sigma) \geq 1 - \frac{1}{k^2}}\,.$$

(Equivalently, $P(|Y-\mu| \geq k\sigma) \leq \dfrac{1}{k^2}$.)

Example 21 The number of customers, Y, who visit a certain store per day has been observed over a long period of time and found to have mean and variance both equal to about 50.

$n = 2000$

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 49.93 \approx 50$

Find a lower bound for the probability that from 40 to 60 customers will visit the store tomorrow (inclusive).

$\frac{1}{n-1}\sum(y_i - \bar{y})^2 = 50.18 \approx 50$

$$
\begin{aligned}
P(Y \in \{40,41,\dots,60\}) &= P(39 < Y < 61) \qquad\qquad (*)\\
&= P(|Y-50| < 11)\\
&= P(|Y-\mu| < k\sigma) \quad \text{where: } \mu = 50,\ \sigma = \sqrt{50} \text{ so that } k = 11/\sqrt{50}\\
&\geq 1 - \frac{1}{k^2}\\
&= 1 - \frac{1}{(11/\sqrt{50})^2}\\
&= 1 - 50/121\\
&= 0.5868.
\end{aligned}
$$

So the prob. that 40 to 60 people will visit the store tomorrow is *at least* 58.68%.

Note 1: At (*) above we could also have written

$$
\begin{aligned}
P(Y \in \{40,41,\dots,60\}) &= P(39.9 < Y < 60.1)\\
&= P(|Y-50| < 10.1)\\
&= P(|Y-\mu| < k\sigma) \quad \text{where: } \mu = 50,\ \sigma = \sqrt{50} \text{ so that } k = 10.1/\sqrt{50}\\
&\geq 1 - \frac{1}{k^2}\\
&= 0.5099.
\end{aligned}
$$

So the probability in question is at least 50.99%. This is equally true, but not so useful, as saying that the probability is at least 58.68%. It is for this reason that at (*) we used the *widest* possible interval we could have, i.e. (39,61), before proceeding.

Note 2: Recall that count data often has a Poisson distribution. So it is of interest to work out the probability *exactly*, under the assumption that Y has this distribution.

Suppose that $Y \sim \text{Poi}(50)$. Then

$$P(40 \leq Y \leq 60) = \frac{e^{-50}50^{40}}{40!} + \frac{e^{-50}50^{41}}{41!} + \dots + \frac{e^{-50}50^{60}}{60!} = 0.8633$$

(which is indeed at least as big as 0.5868, the lower bound worked out earlier).

Note that if Y's mean and variance were *different*, it would be inappropriate to model Y's dsn as Poisson. But we could still apply Chebyshev's thm to get a lower bound.

Proof of Chebyshev's theorem

$$\overset{1}{\cancel{\sigma^2}} = E(Y-\mu)^2 = \sum_y (y-\mu)^2 p(y)$$

by defn

$$= \sum_{y:|y-\mu|<k\sigma} (y-\mu)^2 p(y) + \sum_{y:|y-\mu|\ge k\sigma} (y-\mu)^2 p(y)$$

$$\ge \cancel{\sum_{y:|y-\mu|<k\sigma} 0\, p(y)} + \sum_{y:|y-\mu|\ge k\sigma} (k\sigma)^2 p(y)$$

since $(y-\mu)^2 \ge 0$ for all y,

and $(y-\mu)^2 \ge (k\sigma)^2$ for all y such that $|y-\mu| \ge k\sigma$

$$= 0 + k^2\sigma^2 \sum_{y:|y-\mu|\ge k\sigma} p(y)$$

$$= k^2\cancel{\sigma^2} P(|Y-\mu| \ge k\sigma).$$

So $\boxed{P(|Y-\mu| \ge k\sigma) \le \frac{1}{k^2}}$, and hence also $P(|Y-\mu| < k\sigma) \ge 1 - \frac{1}{k^2}$.

QED

Two other measures of central tendency (other than the mean) EY

The *mode* of a rv Y is any value y at which Y's pdf, $p(y)$ (or $f(y)$) is a maximum.

(There may be more than one mode, and the mode may then also be defined as the set of all such modes.)

The *median* of a rv Y is any value y such that ER

$$\boxed{P(Y \le y) \ge 1/2 \quad \text{and} \quad P(Y \ge y) \ge 1/2\,.}$$

(There may be more than one median, and the median may then also be defined as the set of all such medians.)

Example 22 Find the mode and median of the geometric distribution with parameter 1/4.

F(y) cdf (see Ch4)

y	$p(y) = \left(\frac{3}{4}\right)^{y-1}\frac{1}{4}$	$P(Y \le y)$		$P(Y \ge y) = 1 - P(Y < y) = 1 - P(Y \le y-1)$	
1	1/4 = 16/64	16/64	(< 1/2)	1	(≥ 1/2)
2	(3/4)1/4 = 12/64	28/64	(< 1/2)	1 − 16/64 = 48/64	(≥ 1/2)
3	(9/16)1/4 = 9/64	37/64	(≥ 1/2)	1 − 28/64 = 36/64	(≥ 1/2) ✓
4	(27/64)1/4 = 6.75/64	43.75/64	(≥ 1/2)	1 − 37/64 = 27/64	(< 1/2)
etc.					

We see that *Mode*(Y) = 1 and *Median*(Y) = 3.

Note that $P(Y \le 3) = 37/64 \ge 1/2$ *and* $P(Y \ge 3) = 1 - 28/64 = 36/64 \ge 1/2$, and that no value other than 3 could be substituted here. So 3 is the only median.

Also, Y's mean is *Mean*(Y) = EY = 1/(1/4) = 4 (see Tutorial 5).

$$\sum_{y=1}^{\infty} y\,f(y) = \ldots.$$

$$\approx \sum_{y=1}^{10} y\,f(y) = \ldots\ \boxed{} \cong 4$$

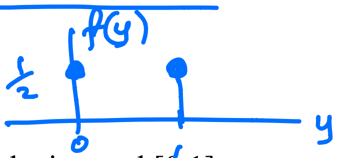
Some general interpretations

1. *mean* = average of a very large number of independent realisations of Y
2. *mode* = most likely value of Y
3. *median* = 'middle' value, such that Y is at least 50% likely to be above or at that ~~or~~ value and at least 50% likely to be at or below it.

An example of a rv with multiple medians and modes

$$\boxed{f(y) = \tfrac{1}{2},\ y = 0, 1}$$

Suppose that $Y \sim Bern(1/2)$.



Then *Median*(Y) = any value in the interval [0,1].

Eg, 1 is a median because $P(Y \le 1) = 1 \ge 1/2$ and $P(Y \ge 1) = 1/2 \ge 1/2$.

Eg, 0.7 is a median because $P(Y \le 0.7) = 1/2 \ge 1/2$ and $P(Y \ge 0.7) = 1/2 \ge 1/2$.

Thus Y has an infinite number of medians. We may also write *Median*(Y) = [0,1].

Note that Y here also has multiple modes: 0 and 1.
We may say that Y is *multimodal* or *bimodal*.
We may also write *Mode*(Y) = {0,1}.

Finally, note that Y's mean is 0(1/2) + 1(1/2) = 1/2.

Summary of discrete distributions

As an exercise, fill in the empty cells below and check against the back inside cover of the text book. You may also wish to add two more columns, one for the mode and one for the median, although these may not always have simple formulas.

distribution $Y \sim$	**pdf** $p(y)$	**mgf** $m(t) = Ee^{Yt}$	**mean** $\mu = EY$	**variance** $\sigma^2 = VarY$
Binomial Bin(n,p)	$\binom{n}{y} p^y (1-p)^{n-y}$ $y = 0,1,\ldots,n$		np	$np(1-p)$
Bernoulli Bern(p) $= ?$				
Geometric				
Hypergeometric				
Poisson				

Completed summary of discrete distributions

distribution $Y \sim$	pdf $p(y)$	mgf $m(t) = Ee^{Yt}$	mean $\mu = EY$	variance $\sigma^2 = VarY$
Binomial Bin(n,p)	$\binom{n}{y} p^y (1-p)^{n-y}$ $y = 0,1,\ldots,n$	$(1-p+pe^t)^n$	np	$np(1-p)$
Bernoulli Bern(p) = Bin(1,p)	$\begin{cases} 1-p, & y=0 \\ p, & y=1 \end{cases}$	$1-p+pe^t$	p	$p(1-p)$
Geometric Geo(p)	$(1-p)^{y-1}p$ $y = 1,2,3,\ldots$	$\dfrac{pe^t}{1-(1-p)e^t}$	$\dfrac{1}{p}$	$\dfrac{1-p}{p^2}$
Hypergeometric Hyp(N,r,n)	$\dfrac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}}$ $y = 0,\ldots,r$ $0 \le n-y \le N-r$	no simple expression	$\dfrac{nr}{N}$	$\dfrac{nr(N-r)(N-n)}{N^2(N-1)}$
Poisson Poi(λ) $= \lim_{\substack{n\to\infty \\ np=\lambda}} \text{Bin}(n,p)$	$p(y) = \dfrac{e^{-\lambda}\lambda^y}{y!}$ $y = 0,1,2,\ldots$	$e^{\lambda(e^t-1)}$	λ	λ

Negative binomial
Neg(r, p)
Neg(1,p) = Geo(p)
(Section 3.6)
	$\boxed{\binom{y-1}{r-1} p^r q^{y-r}}$
	$y = r, r+1, \ldots$
		$\left(\dfrac{pe^t}{1-(1-p)e^t}\right)^r$	$\dfrac{r}{p}$	$\dfrac{r(1-p)}{p^2}$

assessable!

We ~~toss a~~ roll a die until we get the ~~3~~5th 6

5 $\underbrace{3\,3\,1\,2\,5\,6}\,\underbrace{6}\,\underbrace{2\,1\,1\,1\,6}\,\underbrace{3\,6}\,\underbrace{1\,1\,1\,2\,6}$ stop

Y = # rolls $\sim$ NegBin($r=5$, $p=1/6$)

Nb: $Y = Y_1 + Y_2 + \cdots + Y_5$ where

$Y_1, \ldots, Y_5 \overset{iid}{\sim} \text{Geo}(p=1/6)$

Y_1 = # rolls until 1st 6

Y_2 = - - ——— - 2nd 6

$\vdots$

Y_5 = - - - - - 5th 6

eg $y_1 = 6$
$y_2 = 1$
$y_3 = 5$
$y_4 = 2$
$y_5 = 5$

$y = \boxed{19}$