This time:	continuous rvs; joint, marginal & conditional distributions
next time:	functions of rvs; expected value

Read: DS ch. 3, 4

AMS 131 3 Aug 16

Case Study (doc. ① com. notes)

The Poisson process | DS p. 294

You work for a bank that has lately been receiving complaints about long waiting times in the bank teller lines. To quantify the extent of the problem you gather a data set like the one at left, on Mon, Tue, ..., Sat in a randomly chosen week. This is a new type of data set for us: it unfolds in time.

time	event
9 AM	bank opens
9.01 am	arrival
9.03 am	arrival
9.04 am	arrival
9.05 am	departure
⋮	⋮

There are two equivalent ways to keep track of the arrivals: you could let $N(t) = \#$ arrivals in $[0, t]$, or you could keep track of the inter-arrival times (times between arrivals) $T_1, T_2, \ldots$

Here we'll look at $N(t)$. **Definition:**

A stochastic process is a collection of random variables indexed by elements of an indexing set, usually $t \in T$, where t represents time. $N(t)$, $t \in [0, t_{max}]$

9AM

e.g. 5pm

$N(t)$ = # arrivals in $[0, t]$ is an example of a stochastic process, in which the time index is continuous; $\{T_1, T_2, \ldots\}$ (interarrival times) is an example of a stochastic process with discrete time index. (T_t, $t = 1, 2, \ldots$)

What may reasonably be assumed about the random behavior of $N(t)$?

Assumption 1 The numbers of arrivals in any collection of disjoint time intervals are (mutually) independent (this is reasonable if unrelated customers arrive at the bank haphazardly in time).

Assumption 2 P(1 or more arrivals in short time interval $[s, s+t]$, $t > 0$ small) is proportional to t, for example $\lambda \cdot t$ for a rate parameter $\lambda > 0$ (this is reasonable if the arrivals process is smooth rather than lumpy).

Definition To say that a function $f(t)$ of t (small) is $o(t)$ (read little-oh of t) is to say that

$$\lim_{t \to 0} \frac{f(t)}{t} = 0$$

In other words, $f(t)$ approaches 0 as $t \to 0$ at a rate faster than t itself.

The mathematically formal way to state Assumption 2 is then

$$P(\text{at least one arrival in } [s, s+t]) = \lambda t + o(t) \quad (t \text{ small})$$

Assumption 3 (nearly simultaneous arrivals are rare)

$$P\left(\begin{array}{c}\text{2 or more arrivals}\\ \text{in } [s, s+t)\end{array}\right) = o(t) \quad (t \text{ small})$$

Remarkably, these 3 simple & often plausible assumptions specify the probability behavior of $N(t)$ uniquely (see DS exercise 16, p. 296) for a proof of the following result).

λ constant $\leftrightarrow$ $N(t)$ is a stationary stochastic process

Note that Assumption 2 implies that the rate parameter λ is constant in time; this would be unrealistic in the bank problem over an entire day but would be reasonable during stable subsets of a day (e.g. 10am – 11:30am)

⑤

Definition For any $\lambda > 0$, a random variable Y has the <u>Poisson distribution</u> (Poisson(λ)) with parameter λ if its pmf is

S. Poisson (1800s)

$$f_Y(y \mid \lambda) = P(Y = y \mid \lambda) = \begin{cases} \dfrac{\lambda^y e^{-\lambda}}{y!} & \text{for } y = 0, 1, \ldots \\ 0 & \text{else} \end{cases}$$

(<u>Abraham de Moivre</u> did this earlier (1711))

Result Under Assumptions 1-3 above, if $Y = \begin{pmatrix} \text{\# of arrivals in any} \\ \text{time interval of length } t \end{pmatrix}$ then $Y \sim$ Poisson(λt).

Definition

A <u>Poisson process</u> with rate λ per unit time is a stochastic process satisfying

(1) for any $s > 0$, # arrivals in $[s, s+t) \sim$ Poisson(λt)

and (2) # of arrivals in disjoint time intervals are independent

Restatement of Result Under Assumptions 1–3 above, $N(t)$ is a Poisson process with rate parameter λ.

Exploring the Poisson distribution

$$f(y \mid \lambda) = P(Y = y \mid \lambda) = \begin{cases} \dfrac{\lambda^y e^{-\lambda}}{y!} & y = 0, 1, \ldots \\ 0 & \text{else} \end{cases}$$

(for $\lambda > 0$). This is our first example of a discrete rv that takes on a countably infinite number of possible values.

Comment: In reality we don't expect a Poisson rv to be infinite or "nearly infinite"; all the statement ($y = 0, 1, \ldots$) means is that ahead of time we can't place a fixed upper bound on Y.

The first thing to check is that the Poisson probabilities add up to 1:

$$\sum_{y=0}^{\infty} f_Y(y|\lambda) = \sum_{y=0}^{\infty} P(Y=y|\lambda) = \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!}$$

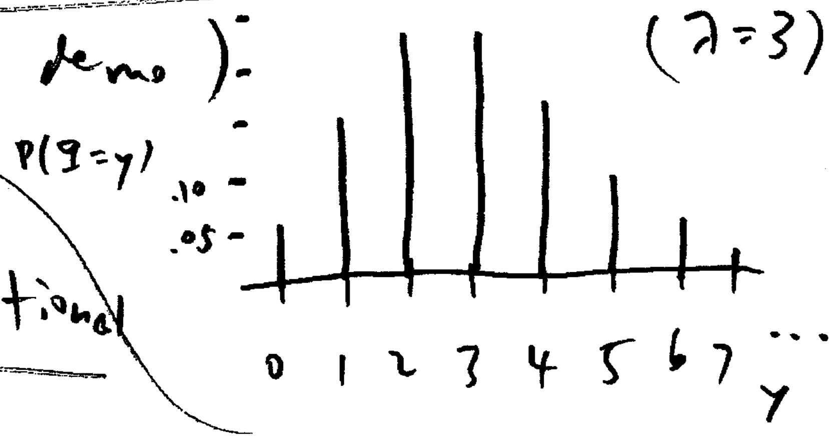
$$= e^{-\lambda} \cdot \underbrace{\sum_{y=0}^{\infty} \frac{\lambda^y}{y!}}_{e^{\lambda}}$$

But you may remember from your calculus class that the Taylor series expansion for e^x about 0 is $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, so

$\sum_{y=0}^{\infty} \frac{\lambda^y}{y!} = e^{\lambda}$ and $\sum_{y=0}^{\infty} f(y|\lambda)$ is indeed 1.

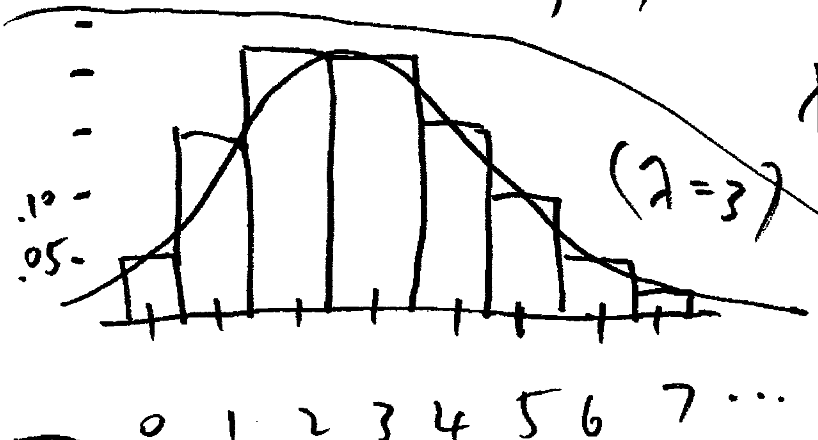
(9.55)

Exploring the shape of the Poisson distribution (R demo):



We need some words to describe distributional shape.

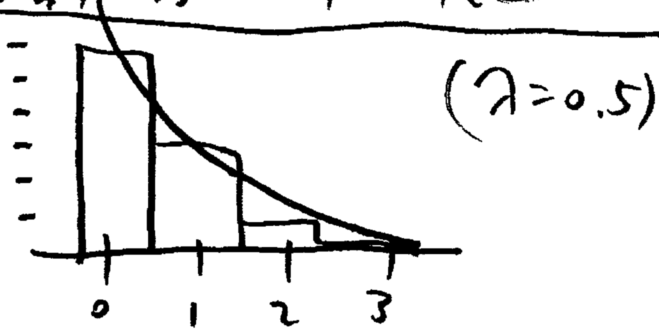
It's a minor but useful distortion of the discreteness of the Poisson distribution to replace the point masses at 0, 1, ... with bars (rectangles) centered at 0, 1, ... and 1 unit wide:



the resulting plot is a special kind of bar graph (relative-frequency) called a histogram.

note: total area of the bars = 1

You can capture the basic shape of the distribution by drawing a smooth curve through or near the centers of the tops of the bars.



(c) (n fixed, $T \uparrow$) $\alpha \uparrow 1$ $\leftrightarrow$ with a small sample from a large population, SRS $\doteq$ IID

Poisson ($\lambda > 0$) $X \sim \text{Poisson}(\lambda)$

$\leftrightarrow$ X has PMF $f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \underbrace{I_{\{0,1,\dots\}}(x)}_{\text{support of } X}$

$E(X) = \lambda$

$V(X) = \lambda$

Thus for the Poisson dist.

$\frac{V(X)}{E(X)} = 1$

Def. If $E(X)$ and $V(X)$ both exist and $E(X) > 0$, $\frac{V(X)}{E(X)}$ is called the variance-to-mean ratio (VTMR)

$\psi_X(t) = e^{\lambda(e^t - 1)}$

$-\infty < t < \infty$

The Poisson can be unrealistic as a consequence of its VTMR of 1, because many rvs that represent counts of occurrences of events in time intervals of fixed length have VTMR > 1.

The Poisson & Binomial distributions both count the numbers of "successes" in a process unfolding in time, so it should not be surprising to find out that these 2 dist. are related:

---

when $\begin{pmatrix} n \text{ is large} \\ p \text{ is close to } 0 \end{pmatrix}$, PMF of Binomial$(n, p) \doteq$ PMF of Poisson$(n \cdot p)$

Theorem | n positive integer, $0 < p < 1$, $X \sim$ Binomial(n, p)

$\lambda > 0$, $Y \sim$ Poisson(λ) / Choose any sequence

$\{p_n\}_{n=1}^{\infty}$ of values between 0 and 1 with

$\lim_{n \to \infty} n \cdot p_n = \lambda$ Then $f_{\underline{X}}(x \mid n, p_n) \xrightarrow[n \to \infty]{}$ $f_{\underline{Y}}(y \mid \lambda)$

PMF Binomial

PMF Poisson

Poisson process, revisited

Def

A poisson process with rate λ per unit time (or space, or volume, or...) is a stochastic process with two properties:

(a) # arrivals in every interval of time of length $t \sim \text{Poisson}(\lambda t)$

(b) #s of arrivals in all disjoint (non-overlapping) time intervals are independent

~~Case Study~~

~~Parasitic~~ Protozoa in drinking water

There's a kind of parasitic

organism called cryptosporidium that's capable of getting into the public drinking water supplies; at one stage in their life cycle they're called oocysts. They can make people sick at a concentration of only 1 oocyst per 5 liters $\doteq$ 1.3 gallons of water.

One problem is that it can be hard to detect these oocysts with water filtration. Suppose that, in the water supply of your city, oocysts occur according to a Poisson process with rate λ oocysts per liter, & that the filtering system your water utility company uses can capture all the oocysts in a water sample but only has probability p of detecting each oocyst that's actually there. (& counting events are independent)

Set Y = # actual oocysts in t liters of water, and $X_i = \begin{cases} 1 & \text{if oocyst } i \text{ gets counted} \\ 0 & \text{else} \end{cases}$

X = # counted oocysts. Then $(X \mid Y = y) = \sum_{i=1}^{y} X_i$

under these assumptions, $(X \mid Y = y) \sim \text{Binomial}(y, p)$

Q: What's the dist. of X? A: By the law of total probability

$$f_X(x) = P(X = x) = \sum_{y=0}^{\infty} P(Y = y) P(X = x \mid Y = y)$$

for all $x = 0, 1, \ldots$

in which $P(Y = y) = \dfrac{(\lambda t)^y e^{-\lambda t}}{y!}$ for $y = 0, 1, \ldots$

and $P(X=x \mid Y=y) = \binom{y}{x} p^x (1-p)^{y-x}$

Notice that if $X=x$, $Y \geq x$ because the actual number of oocysts (Y) has to be at least as large as the number of oocysts detected (X). After a careful calculation you get:

$$f_X(x) = \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \frac{(\lambda t)^y e^{-\lambda t}}{y!}$$

$$= \frac{e^{-p\lambda t} (p\lambda t)^x}{x!}$$

i.e., $X \sim \text{Poisson}(p\lambda t)$: losing a proportion $(1-p)$ of the oocysts to faulty counting just lowers the rate of the Poisson process from λ/liter to $\lambda \cdot p$/liter (makes excellent sense).

In practice oocysts are hard to detect: p is small (not far from 0). **Q:** How much water (t liters) do you need to filter to achieve P(at least 1 oocyte detected) $\geq 1-\alpha$ for small α? **A:** Not hard to work out

$$P(\text{at least 1 detected}) = 1 - P(\text{none detected})$$

$$= 1 - P(\underline{X} = 0) = 1 - e^{-p\lambda t} \geq 1-\alpha$$

$$\leftrightarrow \alpha \geq e^{-p\lambda t} \leftrightarrow \ln\alpha \geq -p\lambda t \leftrightarrow$$

$$\boxed{t \geq -\frac{\ln\alpha}{p\lambda}}$$

Example $\alpha = .01$, $p = 0.1$, $\lambda = 0.2$ / liter (1 per 5 liters) $\updownarrow$ (minimum sickness level)

to achieve prob 99%, t has to be at least 230.3 liters.