270: Special Distributions

D. Alex Hughes

Math Camp

September 18, 2015

Introduction

What do we require of a probability distribution/density model?

- It must be non-negative for all outcomes in its sample space
- It must sum or integrate to one across the sample space.

By this definition, $f(y) = \frac{y}{4} + 7\frac{y^3}{2}, 0 \leq y \leq 1$ is a valid pdf.

- $f(y) \geq 0, \forall y 0 \leq 1$, and;
- $\int_0^1 \frac{y}{4} + 7\frac{y^3}{2} dy = 1$

But, how useful is this model?

A Model's Merit

- A pdf has utility as a probability model if it actually models the behavior of real- world behavior.
- A surprisingly small number of distributions describe these real world outcomes.
- Many measurements/outcomes are the result of the same set of assumptions about the data generating process
  - Number of fraud incidents
  - Number of latrines built
  - Feeding patterns of zebra muscles

The Three Most Important Discrete Distributions

1. Uniform
2. Binomial
3. Poisson

Arriving at the Poisson

- Suppose we have some occurrence that happens at random though time. Call it 911 calls in Oakland.
- We might, reasonably, be interested in the probability that more than 10 calls are made to the police in a 5 minute window. You know, we've got to have someone guarding the inmates, not just yaking on the phone.
- It seems reasonable to assume that there is a more or less constant rate of calls. Name it λ. If this is defined at the average rate/minute, then in any five minute window, there will be 5λ calls.
- It also seems reasonable to assume that the calls are independent across time. Unless there is a fire. Or a sharknado.
- **How would you model the number of calls?**

Model this with a binomial?

- If we make the time windows short (very short?), then the probability of two calls in a single time is zero, and we have a Bernoulli trial in each interval. Sounds like a Binomial to me!
- If the rate is λ, how many calls would we expect in 3 units time?
- What if we cut this into n subintervals? What is the probability of a call in any of those subintervals?

$$\frac{\lambda t}{n} = p$$

What is the probability of *no* calls in a window?

Model this with a binomial?

Looks like a binomial to me!

$$\binom{n}{0}(p)^0(1-p)^n$$

We can substitute in here :

$$\begin{aligned} P(X=0) &= \left(1-\frac{\lambda t}{n}\right)^n \\ &= \left(1-\frac{\lambda}{n}\right)^n \end{aligned}$$

For a large n, this simplifies to:

$$e^{-\lambda}$$

Model this with a binomial?

We can figure out, with just a little bit of combinatoric magic, that the for any fixed number of successes, the incremental ratio of a success can be written as,

$$\frac{b(n,p,k)}{b(n,p,k-1)} = \frac{\lambda - (k-1)p}{kq}$$

Finally, when n is large (and so p is small) simplifies to $\frac{\lambda}{k}$. So if we are looking for one success, we can substitute into the earlier eq:

$$P(X = 1) \approx \lambda e^{-\lambda}$$

, and for any k,

$$P(X = k) \approx \frac{\lambda^k}{k!} e^{-\lambda}$$

Introduction to Poisson

Think of the binomial distribution:

- $p(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{1-k}$
- We were evaluating this for relatively small numbers of trials
- Suppose we were to evaluate this for $n = 1000, k = 500$. Does this work on your TI-89? What about $n = 10,000, k = 5,000$

Introduction to Poisson

Think of the binomial distribution:

- $p(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{1-k}$
- We were evaluating this for relatively small numbers of trials
- Suppose we were to evaluate this for $n = 1000, k = 500$. Does this work on your TI-89? What about $n = 10,000, k = 5,000$
- And this is 2012 – the future!

The Poisson Limit

Theorem

Suppose X is a binomial random variable, where

$$P(X = k) = p(k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, ...n$$

Then, if $n \to \infty$ and $p \to 0$ in such a way that $\lambda = np$ remains constant then

$$\lim_{\{n \to \infty; p \to; np = const.\}} p(k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-np}(np)^k}{k!}$$

$$= \frac{e^{-\lambda}\lambda^k}{k!}$$

The Poisson Limit

Proof.

Let $\lambda = np$. Then, rewrite the binomial probability in terms of λ

$$\begin{aligned}
\lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} &= \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k!)} \lambda^k \left(\frac{1}{n^k}\right) \left(1-\frac{\lambda}{n}\right)^{-k} \left(1-\frac{\lambda}{n}\right) \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{(n-\lambda)^k} \left(1-\frac{\lambda}{n}\right)^n
\end{aligned}$$

As $n \to \infty, [1-(\lambda/n)]^n \to e^{-\lambda}$ Se we need to demonstrate only that the first term approaches 1.

$$\frac{n!}{(n-k)!(n-\lambda)^k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{(n-\lambda)(n-lambda)\cdots(n-\lambda)} \to_{n\to\infty} 1$$

Great, so what?

Before the eye-roll, note that the proof is an asymptotic result – it holds when n is large? But at what point does the proof breakdown? Or, how small an n is too small? Typically $n = 100$ is large enough to well-approximate data.

Example

A hospital serves 12,000 residents. There is a 1/8,000 chance that a patient will need an AED on *any given day*. The hospital has three AED. What is the probability that the hospital will have insufficient equipment for tomorrow?

Answer

Let X be a binomial random variable (n=12,000, p=1/8,000). Then we are interested in P(X¿3)

$$\begin{aligned} P(X > 3) &= 1 - P(X \leq 3) \\ &= 1 - \sum_{k=0}^{3} \binom{12,000}{k} \left(\frac{1}{8000}\right)^{k} \left(\frac{7999}{8000}\right)^{12,000-k} \\ &= 1 - \sum_{k=0}^{n} \frac{e^{-1.5}(1.5)^{k}}{k!} \\ &= 0.0656 \end{aligned}$$

Typing Slides

Example

Alex averages one typo per 3250 works. What is the chance that a 6000 word slide-deck is free of typos? First, do it using the exact binomial, then use the Poisson approximation.

Typing Slides

Example

Alex averages one typo per 3250 works. What is the chance that a 6000 word slide-deck is free of typos? First, do it using the exact binomial, then use the Poisson approximation.

Answer

p=1/3250; n=6000

$$P(X=0) = \binom{6000}{0}\left(\frac{1}{3250}\right)^{0}\left(\frac{3249}{3250}\right)^{6000} = 0.158$$

For the Poisson approximation, $\lambda = np = 6000(1/3250) = 1.846$.

$$P(X=0) = \frac{e^{-1.846}(1.846)^{0}}{0!} = 0.158$$

The Poisson Distribution

Theorem

The random variable X is said to have a Poisson distribution if

$$P(X = k) = p(k) = \frac{e^{-\lambda}\lambda^k}{k!}, \text{for} k = 0, 1, 2, ...$$

where λ is a positive constant.
For any Poisson random variable $E[X] = \lambda$, and $VAR[X] = \lambda$.

- Then, if we think that a random variable is likely to be Poisson distributed, we can fit a model to our data by calculating little more than the "average" value of the data.
- *How do we know if we can use the Poisson distribution?*
- Driven by the numbers of 0s, 1s, 2s, ... in the sample.

Example

Football example, p. 283

"Law of Small Numbers"

Why does the Poisson model *work?*

- Examples presented so far have all used data generated in a similar way
- Data are 1,0
- Data are independent events
- Probability is fixed

The Exponential Distribution

Sometimes, the time between events is the object of our affections – time to failure of a computer, amp (Jack White), time between elections, and so on.

Theorem

Suppose a series of events satisfying the Poisson model (recall, what are those?) are occuring at the rate of λ per unit time. Let the random variable Y denote the interval between consecutive events. Then Y is distributed exponential:

$$f(y) = \lambda e^{-\lambda y}, y > 0$$

The Exponential Distribution

Proof.

Suppose an event occurs at time a. Consider the interval that extends from a to $a + y$. Then, since they are poisson events occurring at rate λ per unit time,

$$p(0) = \frac{e^{-}\lambda y(\lambda y)^0}{0!} = e^{-\lambda y}.$$

Now, define Y to be the interval between occurrences. There will be no occurrences in $(a, a + y)$ *iff* $Y > y$: $P(Y > y) = e^{-\lambda y}$ or, equivalently, $P(Y \leq y) = 1 - P(Y > y) = 1 - e^{-\lambda y}$ Let f(y) be some (unknown) pdf. It must be true that

$$P(Y \leq y) = \int_0^y f(t)dt$$

$$\frac{d}{dy}\int_0^y f(t)dt = \frac{d}{dy}(1 - e^{-\lambda y})$$

Meteor Shower

Example

The Perseids happened last month. The number of visible meteors can be as high as 40 per hour. Assume such sightings are Poisson events, then calculate the probability that an observer who has just seen a meteor will have to wait at least five minutes before seeing another.

Meteor Shower

Example

The Perseids happened last month. The number of visible meteors can be as high as 40 per hour. Assume such sightings are Poisson events, then calculate the probability that an observer who has just seen a meteor will have to wait at least five minutes before seeing another.

Answer

Let's express the rate in terms of minutes $40/60 = 0.67/\text{minute}$.

$$\begin{aligned} P(Y > 5) &= \int_5^\infty 0.67 e^{-0.67y} dy \\ &= \int_{5/60}^\infty e^{-u} du \\ &= -e^{-u}|_{3.33}^\infty = e^{3.33} \\ &= \mathbf{0.036} \end{aligned}$$

Example

Fifty bodyguards have been placed outside the Lybian embassy. Given the turbulent times, 1.1 of these body guards will be killed every 100 hours. Given no replacement of bodyguards, how many will not survive 750 hours?

Example

Fifty bodyguards have been placed outside the Lybian embassy. Given the turbulent times, 1.1 of these body guards will be killed every 100 hours. Given no replacement of bodyguards, how many will not survive 750 hours?

Answer

Let X be the number of contractor killed in 1 hour. Then $E(X) = \lambda = 0.011$ Let Y be the number of hours a contractor is alive. Then

$$P(Y < 75) = \int_0^{750} 0.011 e^{-0.011y} dy = -e^{-u}|_0^{.825} = 0.56$$

(Substitute $u = 0.11y$).

Since there were 50 contractors initially, 28 will not survive the firefight.

Normal Distribution

The Poisson is useful, but the Normal distribution is by a wide margin the most utilized distribution in statistics. (Why?)

Theorem

Let X be a binomial random variable defined on n independent trials for which $p = P(\text{success})$. For any numbers a and b,

$$\lim_{n\to\infty} P\left(a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} dz$$

(DeMoivre-Laplace Limit)

Alex' Morning

Example

Airlines overbook their flights because they know that on average only 90% of ticket-holders show up. Imagine that all ticket-holders' showing up is an independent event. An airline sells 178 tickets for a 168 seat airplane. What is the probability that the flight is overbooked?

Alex' Morning

Example

Airlines overbook their flights because they know that on average only 90% of ticket-holders show up. Imagine that all ticket-holders' showing up is an independent event. An airline sells 178 tickets for a 168 seat airplane. What is the probability that the flight is overbooked?

Answer

$$
\begin{aligned}
&= P(\textit{Overbooked}) \\
&= P(169 \leq X \leq 178) \\
&= \frac{169 - (178)(0.9)}{\sqrt{(178)(0.9)(0.1)}} \leq \frac{X - (178)(0.9)}{\sqrt{(178)(0.9)(0.1)}} \leq \frac{178 - (178)(0.9)}{\sqrt{(178)(0.9)(0.1)}} \\
&= P(2.07 \leq Z \leq 4.57) \\
&= 0.0192
\end{aligned}
$$

Central Limit Theorem

Theorem

Let W_1, W_2, ... be an infinite sequence of independent random variables, each with the same distribution. Suppose that the mean μ and the variance σ^2 of f(w) are both finite. The any numbers a and b,

$$\lim_{n\to\infty} P\left(a \leq \frac{W_1 + ... + W_n - n\mu}{\sqrt{n}\sigma} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} dz$$

The proof of this is beyond the goals of the class.

Individual Measurements

Imagine that you're an old-school astronomer trying to figure out how far away is some star. Then, when you take a particular measurement, the measurement you write down, call it Y, is the sum of two components:

1. The star's *true* location μ^* (which is unknown); and,
2. measurement error.

We could write a formula for our measurement that looks like:

$$Y = \mu^* + W_1 + W_2 + \ldots + W_t$$

Then, if the equation above is a good representation of the measurement process, the CLT will apply, and

$$E(Y) = E(\mu^* + W_1 + \ldots + W_t) = \mu$$

Normal Distribution

Definition

A random variable Y is said to be normally distributed with mean μ and variance σ^2 if

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, -\infty \leq x \leq \infty$$

Comment

Areas under an "arbitrary" normal distribution, $f_Y(y)$ are calculated by finding the equivalent area under the standard normal distribution $f_Z(z)$

$$P(a \leq Y \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{Y-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right)$$

This is a Z-transformation