Discrete Distributions - Binomial, Poisson

Sumanta Basu

Reading

Reading: Textbook Sections 3.4.1, 3.5.2

Recommended Exercise: 3.25, 3.27, 3.33, 3.37, 3.43, 3.44

Example: Overbooking flights

Airlines often sell more seats than the flight capacity in anticipation that some passengers will not show up.

Assume an United Airlines (UA) flight from Ithaca to Washington, DC has 100 passenger seats. UA sold 102 tickets, since their historical data suggests 2% of passengers who bought ticket do not show up or change to a different flight.

- What are the chances that the flight will be overbooked?
- If a flight is overbooked, UA offers a $600 voucher to volunteer for a different flight. How much money UA is expected to lose due to their decision to overbook the flight?

> Binomial, Poisson or Normal (next lecture) distributions can help us answer such questions!

Important Discrete Distributions

Overview

Bernoulli distribution - single binary trial ("success" or "failure")

Binomial distribution - "success" count based on n binary trials

Poisson distribution - "event" counter (time/space)

And others…

Bernoulli distribution definition

$X \sim \text{Bernoulli}(p)$ means

$$P(X=1)=p \text{ and } P(X=0)=1-p$$

- single trial with binary outcome
- success/failure

PMF:

$$P(X=x)=p^x(1-p)^{1-x} \quad (\text{ for } x=0 \text{ or } 1 \text{ only}).$$

Bernoulli mean?

$$E(X) = \sum_i x_i P(X = x_i) = 1 \cdot P(X = 1) + 0 \cdot P(X = 0) = p$$

Bernoulli variance?

$$\mathrm{Var}(X) = \sum_i (x_i - \mu)^2 P(X = x_i)$$

$$= (1 - p)^2 \cdot P(X = 1) + (0 - p)^2 \cdot P(X = 0)$$

(after some algebra…)

$$= p(1 - p)$$

Binomial distribution definition

$X \sim \text{Binomial}(n, p)$ means

Consider n binary trials in which

- each has **same probability** p of success
- result of each trial is **independent** of all others
- X counts number of successes in these n trials

Note: $X = X_1 + \cdots + X_n$ where each X_i is an independent Bernoulli(p)

Binomial mean and variance

Mean:

$$E(X) = np$$

Variance:

$$\mathrm{Var}(X) = np(1 - p)$$

Standard deviation:

$$\mathrm{SD}(X) = \sqrt{np(1 - p)}$$

Binomial pmf

> $X \sim \text{Binomial}(n, p)$ has probability mass function (pmf)

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n$$

where $\binom{n}{x}$, read "n choose x," is number of ways to choose x items from n items.

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

and $n! = 1 \times 2 \times \cdots \times n$ and $0! = 1$.

Understanding binomial pmf

If $n = 10$, and suppose $X = 3$.

Examples:

$$\begin{gathered} HHHTTTTTTT \\ TTTTHHTHTT \\ THTTTHTTTH \\ \vdots \end{gathered}$$

Each is equally likely:

$$p^3(1-p)^7$$

So $P(X = 3)$ equals

$$(\text{\# of ways to choose 3 items from 10 items}) \times p^3(1-p)^7$$

Counting

How many distinct orderings are there of these letters?

> ABC, ACB, BAC, BCA, CBA, CAB … 3!

But what if we have two identical letters…

> AAC, ACA, AAC, ACA, CAA, CAA

We've overcounted! Don't distinguish between orderings of the letter A. Should divide by $2!$

Counting

How many orderings of $AAABB$?

If question were about $A_1A_2A_3B_1B_2$ (i.e., "same letters" can be distinguished)… $5!$

But what if all A's are indistinguishable and likewise for B's?

$5!$ overcounts…

- e.g., for $ABBAA$ we would have counted $A_1B_1B_2A_2A_3$ and $A_2B_1B_2A_1A_3$ …
- $3!$ orderings of A's and $2!$ orderings of B's

> Number of orderings: $\frac{5!}{3!2!}$

Counting

\# of ways to have 3 H out of 10 tosses

10 letter strings of H and T with 3 H's.

$$\frac{10!}{3!7!}$$

Written:

$$\binom{10}{3}$$

Back to 3 coins

When $n = 3$,

$X = 0$ in $\binom{3}{0} = 1$ way: TTT

$X = 1$ in $\binom{3}{1} = 3$ ways: HTT, THT, TTH

$X = 2$ in $\binom{3}{2} = 3$ ways: THH, HTH, HHT

$X = 3$ in $\binom{3}{3} = 1$ way: HHH

We calculated earlier

$$P(X = x) = \begin{cases} p^3 & \text{if } x = 3 \\ 3p^2(1-p) & \text{if } x = 2 \\ 3p(1-p)^2 & \text{if } x = 1 \\ (1-p)^3 & \text{if } x = 0 \end{cases}$$

Verify that this agrees with

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, \ldots, n$$

Poisson distribution definition

Let X be the **number of occurrences** of a specified "event" over a fixed time period (or defined region of space).

Assume

- events to be counted occur at completely random points (in time or space)
- **no upper bound** to the number of events we might observe

Write $X \sim \text{Poisson}(\lambda)$ where $\lambda > 0$ is a parameter.

Example

Consider a hospital with birth **rate** of $\lambda = 2$ births/hour.

$X =$ # of births in a specific hour

$X \sim \text{Poisson}(2)$

Poisson distribution definition

It has pmf

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!} \text{ for } k = 0, 1, 2, 3, \ldots$$

Mean:

$$E(X) = \lambda$$

Variance:

$$\mathrm{Var}(X) = \lambda$$

Standard deviation:

$$\mathrm{SD}(X) = \sqrt{\lambda}$$

Example

Hospital needs to call for extra staff if there are more than 3 births in an hour.

What's the chance of this?

$$P(X > 3) = 1 - P(X \leq 3)$$
$$= 1 - [P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)]$$
$$= 1 - e^{-\lambda}\left[\frac{\lambda^0}{0!} + \frac{\lambda^1}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!}\right]$$
$$= 1 - e^{-2}\left[\frac{1}{1} + \frac{2}{1} + \frac{4}{2} + \frac{8}{6}\right] = 1 - \frac{19}{3e^2} \approx 14\%$$

Or more simply in R…

```r
1-ppois(3, lambda=2)
```

```
## [1] 0.1428765
```

Example

Maternity unit is charged if they call for extra staff on two or more separate hours during a given day.

$Y =$ # of hours (out of 24 hour) in which they call for extra staff

$$Y \sim \text{Binomial}(24, 0.14)$$

$$P(Y > 1) = 1 - [P(Y = 0) + P(Y = 1)]$$

$$= 1 - \left[\binom{24}{0}(0.14)^0(1 - 0.14)^{24} + \binom{24}{1}(0.14)^1(1 - 0.14)^{23}\right]$$

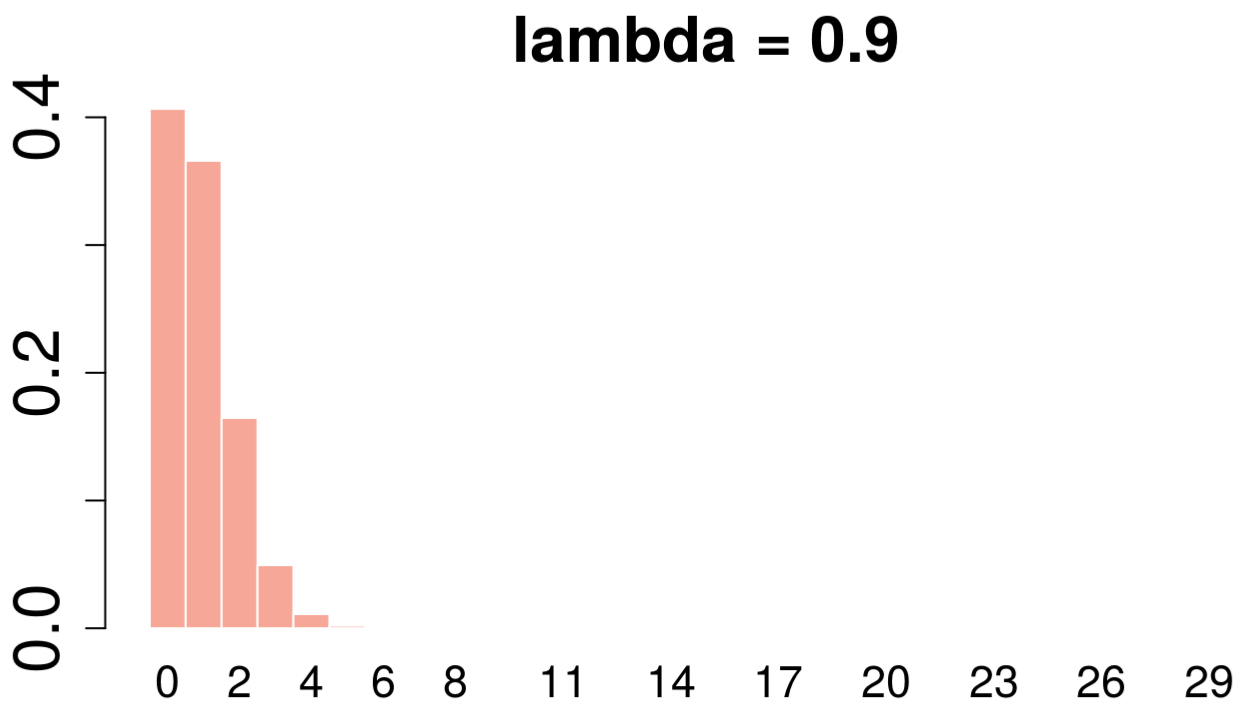
$$= 1 - \left[(0.86)^{24} + 24(0.14)(0.86)^{23}\right] \approx 88\%$$

Or in R…

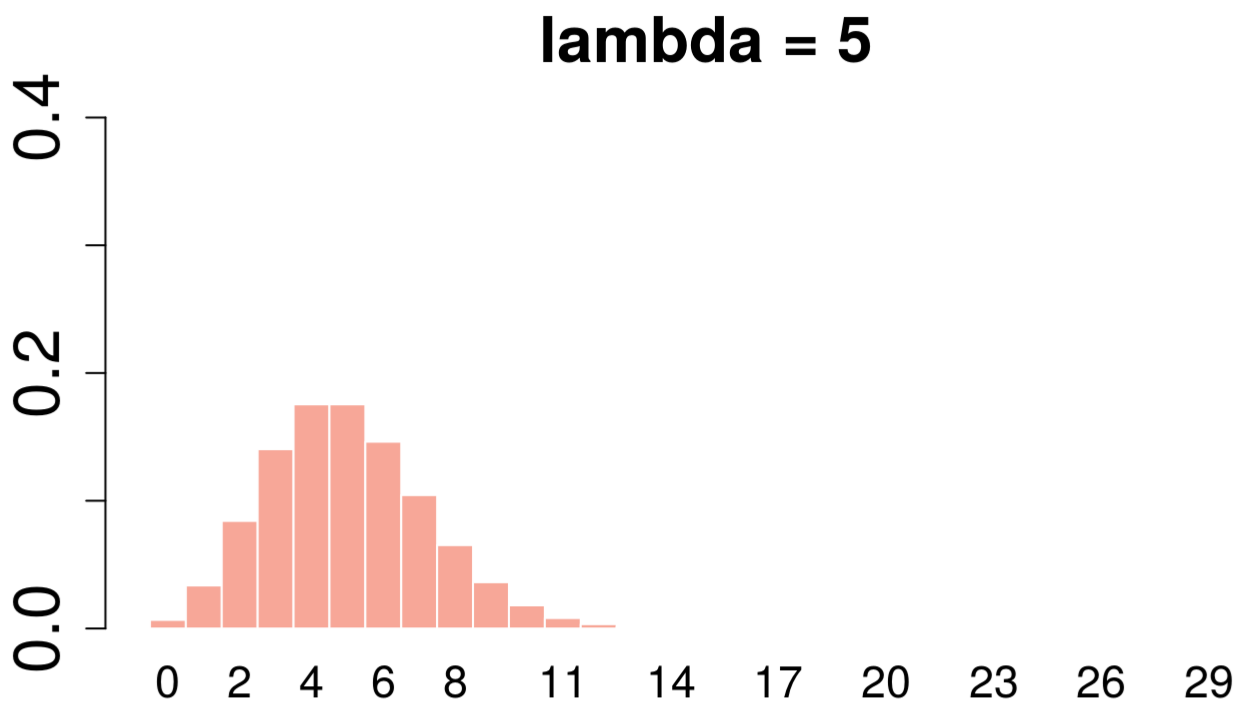
```r
p=1-ppois(3, lambda=2)
1-pbinom(1,size=24,prob=p)
```

```
## [1] 0.8763864
```

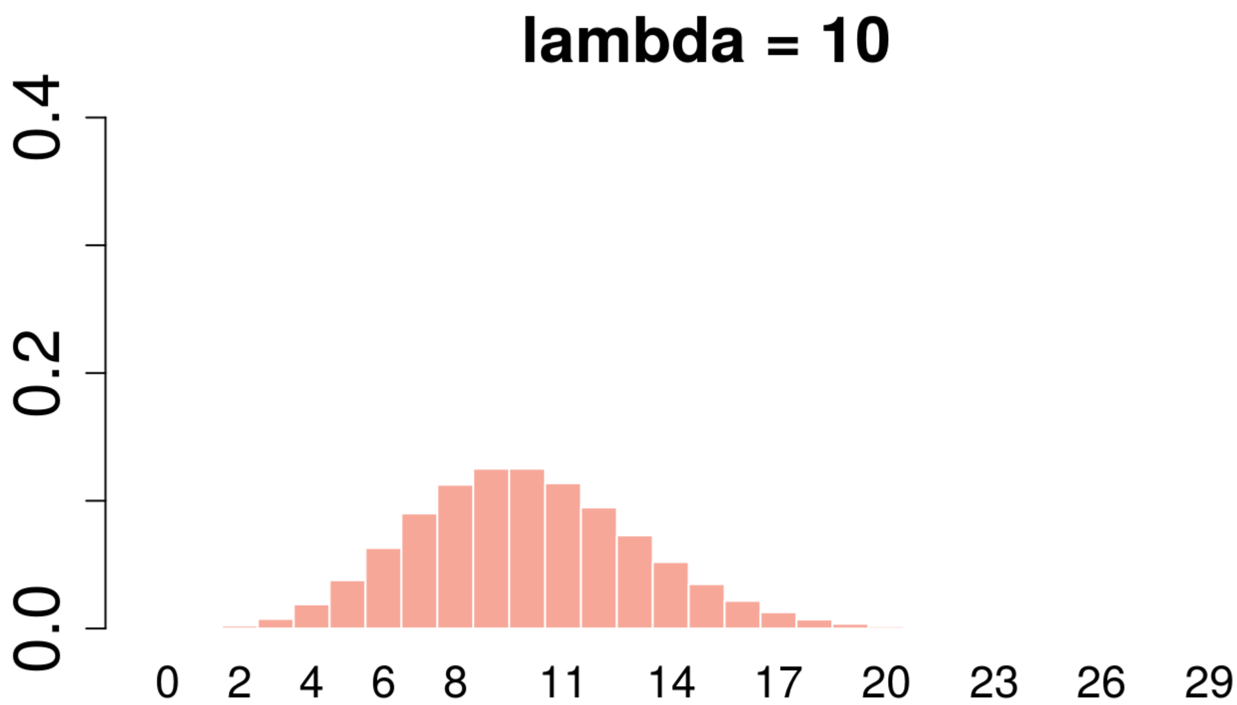
A look at the pmf



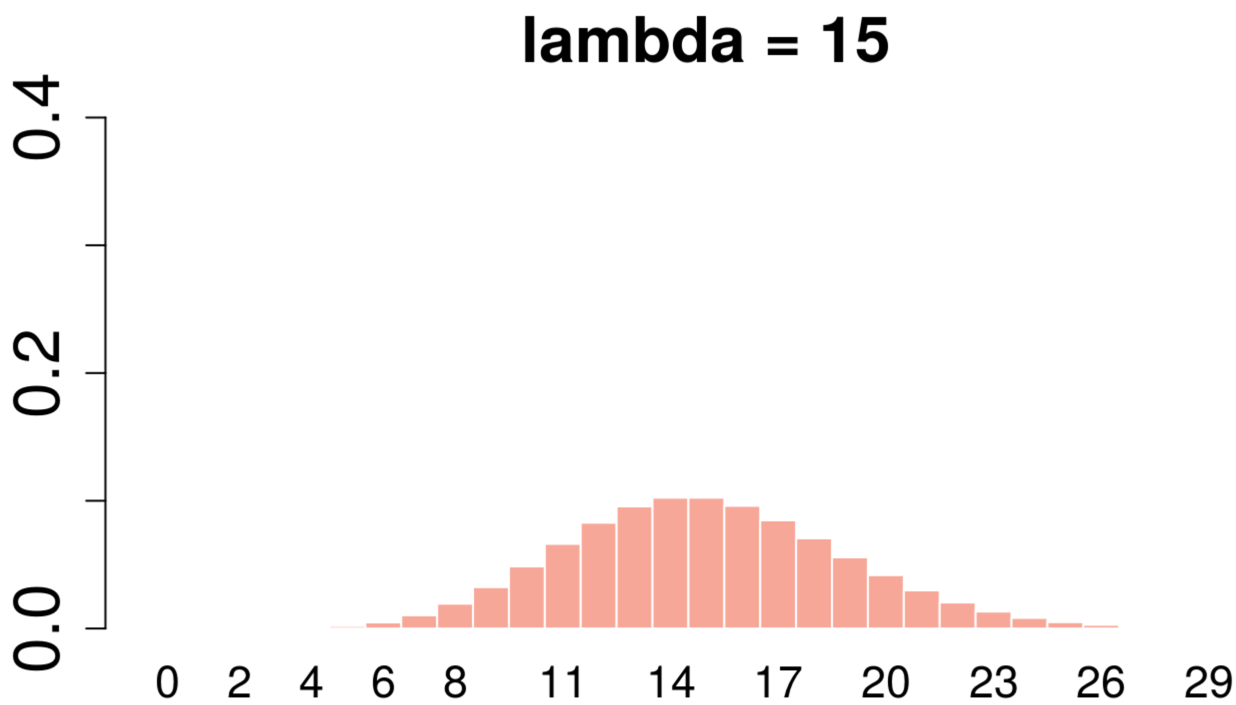
A look at the pmf



A look at the pmf



A look at the pmf



Poisson's relation to Binomial

A binomial with p **small** and n **large** can be *approximated* by a $\text{Poisson}(np)$.

E.g., Blue: $\text{Binomial}(500, 0.01)$ and Green: $\text{Poisson}(5)$