P8130: Biostatistical Methods I

Lecture 4: Discrete Probability Distributions

Cody Chiuzan, PhD
Department of Biostatistics
Mailman School of Public Health (MSPH)

Lecture 3: Recap

- Definitions and basic concepts
- Sets/Events/Rules
- Independent and conditional probability
- “Law of Total Probability” and “Bayes’ Theorem”

Lecture 3: Outline

- Randomness and random variables
- Binomial distribution: definition and statistical properties
- Poisson distribution: definition and statistical properties

Randomness and Random Variables

Variable: a characteristic of each element of a population or sample; a characteristics, number, or quantity that can be measured or counted.

Idea of randomness: adding a probability to the values that the random variable can assume.

Random variable (r. v.): A numerical quantity that takes different values with specified probabilities; the numerical outcome of an experiment or random phenomenon.

Discrete vs Continuous Random Variables

Discrete Random Variable: A numerical r. v. for which there exists a discrete set of values with specified probabilities; there are gaps in the range of possible values

E.g. : number of complains received daily by a cable company

Continuous Random Variable: A numerical r. v. whose values form a continuum (there are no gaps between the values)

E.g. : running times recorded during NYC marathon

Probability Distribution

The probability distribution of a random variable is represented by a table, graph, or formula which denotes what possible values a r. v. can take and the associated probabilities.

Notation: $P(X = x) = P(x)$

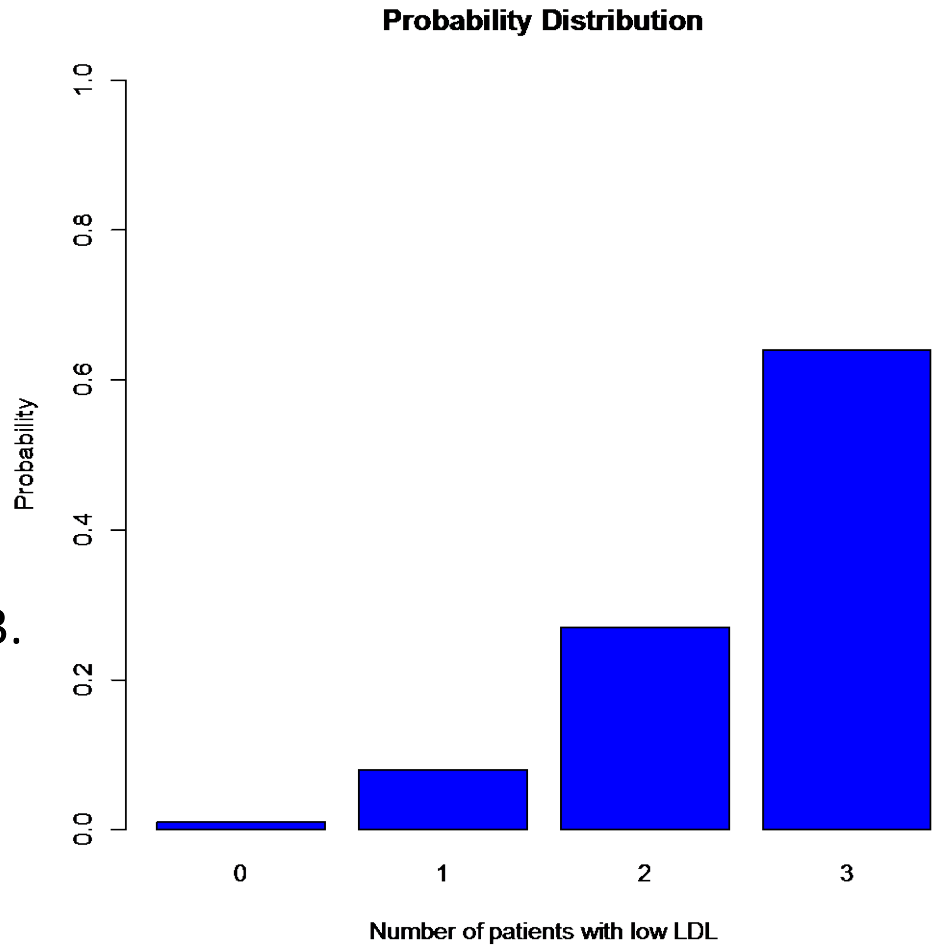
For any probability distribution the following hold true:

1. $P(x) \geq 0$, for any value of x
2. $\sum P(x) = 1$; the sum of probabilities for all x values is 1.

Probability Distribution

<u>Example:</u> An investigator is testing a new medication for lowering the (LDL) cholesterol levels. He expects the following probabilities for the next three patients:

# Patients that respond to medication (X)	Probability of response P(X=r)
0	0.01
1	0.08
2	0.27
3	0.64



X is a discrete random variable that takes values 0, 1, 2, or 3. The probability distribution in this case is called <u>probability mass function</u> (*pmf*).

Cumulative Distribution Function

The cumulative distribution of a discrete r. v. is denoted by:

$$F(x) = P(X \leq x)$$

In the previous example, calculate $F(2) = P(X = 0) + P(X = 1) + P(X = 2) = 0.36$.

Probability that two or less (at most two) patients will respond to medication is 36%.

The complete (step) cumulative function is given by:

$$F(x) = \begin{cases} 0, x < 0 \\ 0.01, 0 \leq x < 1 \\ 0.09, 1 \leq x < 2 \\ 0.36, 2 \leq x < 3 \\ 1, x \geq 3 \end{cases}$$

Expected Value of a Discrete Random Variable

In general, the expected value of a r. v. is its mean or the 'long-run' average value of multiple trials (repetitions).

<u>The expected value of a discrete r. v.</u> is the probability-weighted average of all possible values defined as:

$$\mu = \sum xP(x)$$

Calculate the expected value of X for the cholesterol example:

$$\mu = E(X) = (0 \cdot 0.01) + (1 \cdot 0.08) + (2 \cdot 0.27) + (3 \cdot 0.64) = 2.54$$

Thus, the average number of patients to respond to cholesterol medication is 2.5.

Variance of a Discrete Random Variable

The variance of a discrete random variable is the expected value of the squared deviations from the mean and it's defined as:

Recall that in general: $\sigma^2 = E[(X-\mu)^2]$

For discrete r. v.:

$$\sum (x-\mu)^2 P(x) = \left[\sum x^2 P(x)\right] - \mu^2$$

Calculate the variance of X for the cholesterol example:

$$\sigma^2 = var(X) = [(0^2 \cdot 0.01) + (1^2 \cdot 0.08) + (2^2 \cdot 0.27) + (3^2 \cdot 0.64)] - 2.54^2 = 0.47$$

Thus, the variance of the number of patients to respond to cholesterol medication is 0.47.

Binomial Random Variable

Many experiments have only two options as possible outcomes: e.g., pass/fail, yes/no, etc.

These experiments are called binomial experiments, they generate a discrete r. v. called binomial random variable that follows a binomial distribution.

Characteristics of binomial distribution:

1. Fixed number of *n* trials
2. Trials are independent
3. Only two possible exclusive outcomes
4. The probability of success (*p*) is fixed and the same for each trial

   The probability of failure is denoted by *1-p*.

The random variable of interest is the number of successes in *n* trials.

Binomial Distribution

The probability distribution function of a binomial random variable *X* with *n* trials and probability of success *p* on any trial $X \sim Bin(n, p)$, is denoted by:

$$P(X = x) = f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, x = 0, 1, \; 2, \ldots, n$$

$n! = 1 \cdot 2 \cdot 3 \cdot \; \ldots \cdot n$ is called *'n factorial'*

$\frac{n!}{x!(n-x)!}$ is called the binomial coefficient; it denotes the total number of possible combinations, i.e., choosing *x* objects of *n,* without order being important.

Note: why can we multiple the probabilities of success and failure in *n* successive trials?

Binomial Distribution: Example

Recently, a pediatrician observed that children tend to develop asthma in the first 2 yrs of life in 5 out of 20 households situated within city limits. The national probability (rate) of developing asthma for infants this age is 0.06.

Help the doctor answer the following questions:

1. What is the probability that exactly 5 infants in this sample develop asthma?
2. How often do we expect infants in at least 10 households out of 20 to have asthma?

Binomial Distribution: Example

Recently, a pediatrician observed that children tend to develop asthma in the first 2 yrs of life in 5 out of 20 households situated in the city limits.

Binomial distribution with:

X – random variable denoting the number of asthma cases in 20 households (trials)

$$X \sim Bin(20, 0.06)$$

$n = 20$, number of households

$p = 0.06$, probability of success (developing asthma), $1 - p = 0.94$

Binomial Distribution: Example

Q1: $P(X = 5) = \frac{20!}{5!15!}(0.06)^5(1-0.06)^{20-5}$

$= \frac{16 \cdot 17 \cdot 18 \cdot 19 \cdot 20}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5}(0.06)^5(1-0.06)^{20-5} = \mathit{0.005}$

Q2: $\mathrm{P}(X \geq 10) = \sum_{x=10}^{20} \frac{20!}{x!(20-x)!}(0.06)^x(1-0.06)^{20-x} = 1 - P(X < 10)$

$= 1 - \sum_{x=0}^{9} \frac{20!}{x!(20-x)!}(0.06)^x(1-0.06)^{20-x} = 6.38 \cdot 10^{-8} \approx 0$

These calculations were done in R software (see corresponding code), but you can also use binomial tables (see textbook: Rosner, page 811)

Binomial Distribution

The <u>expected value</u> of a binomial distribution is given by:

$$\mu = E(X) = np$$

The <u>variance</u> of a binomial distribution is given by:

$$\sigma^2 = var(X) = np(1 - p)$$

Back to our asthma example:

$E(X) = 20 \cdot 0.06 = 1.20$; on average you would expect about 1 infant to develop asthma in this sample of 20 households

$\text{var}(X) = 20 \cdot 0.06 \cdot 0.94 = 1.12$; 1.12 is the variance for number of infants to develop asthma in the sample. How do we interpret this?

Poisson Distribution

Poisson process: discrete random variable of the number of occurrences of an event in a continuous interval of time or space: e.g., number of accidents per day at an intersection

Characteristics of Poisson distribution:

1. Events occur one at a time; two or more events cannot occur exactly at the same time and location
2. The occurrence of an event in a given period is independent of the occurrence of an event in a non-overlapping period
3. The expected number of events during any period is constant

Poisson Distribution

The probability distribution of a Poisson random variable $X \sim Poi(\lambda)$ is denoted by:

$$P(X = x) = f(x) = \frac{\lambda^x e^{-\lambda}}{x!}, x = 0, 1,\ 2, \dots, n$$

Where λ represents the expected number of events for a specific period of time (rate)

e is called Euler's number and it's approximately 2.71828

Notes:

1. This distribution depends only on one parameter λ>0
2. There is an infinite number of possible events, but the probabilities will get small as *x* increases.

Poisson Distribution: Example

A cable company averages about 10 calls/complains per hour. Assume that the number of calls follow a Poisson distribution, what is the probability of receiving exactly 4 calls in the next hour?

X – random variable denoting the number of calls per hour, $X \sim Poi(10)$

$\lambda = 10$, rate of hourly calls

Calculate: $P(X = 4) = \frac{10^4 e^{-10}}{4!} = 0.019$

Calculate: $P(X \leq 8) = \sum_{x=0}^{8} \frac{10^x e^{-10}}{x!} = \frac{10^0 e^{-10}}{0!} + \frac{10^1 e^{-10}}{1!} + \ldots + \frac{10^8 e^{-10}}{8!}$ = 0.332

Approximately 33% probability that the company would receive 8 or less calls in the next hour.

Poisson Distribution: Example

Consider that the number of deaths due to typhoid fever over a period of time follows a Poisson distribution. Assume that the expected number of events in one year is 4.6.

Calculate the probability of having zero deaths in the next 6 months.

X – random variable denoting the number of deaths in one year: $X \sim Poi(4.6)$

Calculate: $P(X = 0)$

Poisson Distribution

For a Poisson distribution, the <u>expected value</u> equals the <u>variance</u>:

$$\mu = \sigma^2 = \lambda$$

In other words, the mean and variance are exactly the average rate over a time interval.

Importance?

1. We always compare the mean with the variance (actually standard deviation)
2. A known factor that affects the mean also impacts the variance of the data

Poisson Approximation to Binomial

Under certain conditions, the Poisson distribution is a very good approximation to the binomial distribution where $\lambda = np$.

What are these conditions?

1. n must be large (> 100)
2. Probability of success, p, should be small (p < 0.01)

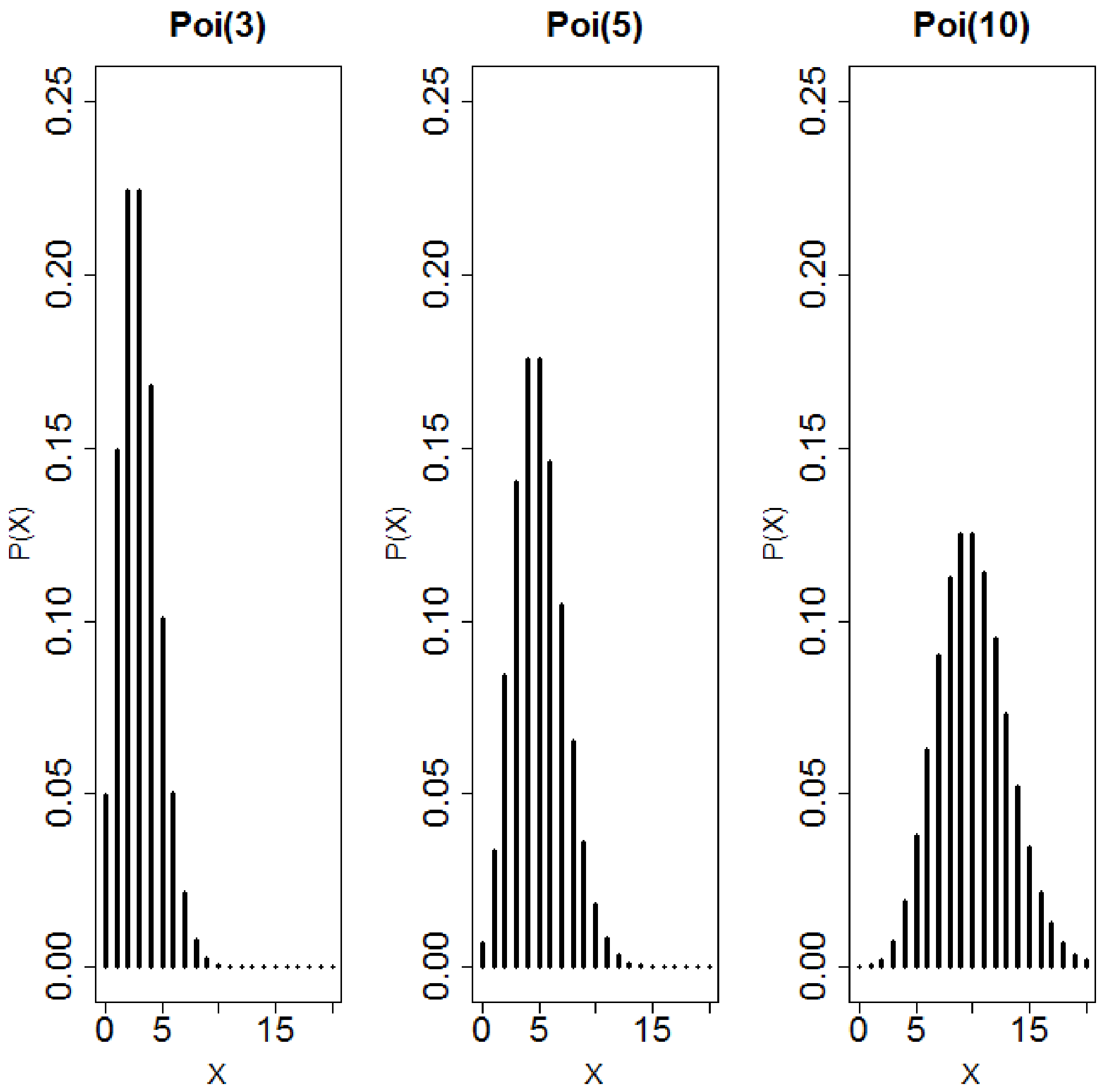
<u>Example:</u> A rare birth defect occurs with probability 0.0001. Assuming that 4,000 babies are born at a large hospital within a year, calculate the probability of having at least 10 babies with a birth defect.

Compute this probability using both Poisson and Binomial formulae and comment on the results – Class Exercise for next time.

Other Shapes of Poisson Distribution

Notice that as the value of parameter λ increases, the distribution becomes more of a bell-shaped (normal distribution).

This indicates that for sufficiently large values of λ, the Normal distribution is a good approximation to the Poisson distribution.



Readings

Rosner, *Fundamentals of Biostatistics*, Chapter 4

- Sections: 4.1 – 4.13