Lecture 11 Discrete Probability Distribution

BIO210 Biostatistics

Xi Chen

Spring, 2024

School of Life Sciences
Southern University of Science and Technology

南方科技大学生命科学学院
SUSTech · SCHOOL OF
LIFE SCIENCES

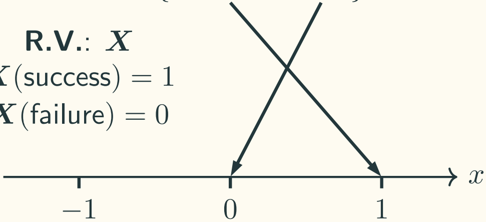
Bernoulli Random Variables

$\Omega = \{\text{success, failure}\}$

R.V.: $\boldsymbol{X}$

$\boldsymbol{X}(\text{success}) = 1$

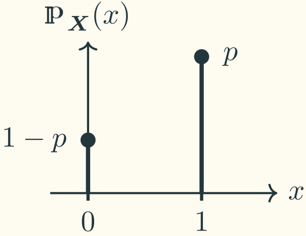
$\boldsymbol{X}(\text{failure}) = 0$



PMF: $\mathbb{P}_{\boldsymbol{X}}(x)$

$\mathbb{P}_{\boldsymbol{X}}(0) = 1 - p$

$\mathbb{P}_{\boldsymbol{X}}(1) = p$



$$\mathbb{P}_{\boldsymbol{X}}(x) = \begin{cases} 1 - p, & \text{if } x = 0 \\ p, & \text{if } x = 1 \\ 0, & \text{otherwise} \end{cases}$$

- $\mathbb{E}[X] = ?$
- $\mathbb{V}\mathrm{ar}(X) = ?$

Binomial Random Variables

Experiment: Perform n independent Bernoulli trials. Let the random variable $\boldsymbol{X}$ represent the number of successes in the n trials and $\mathbb{P}\,(\text{success}) = p$.

Task: Construct a PMF of the random variable $\boldsymbol{X}$.

$n = 2$

ω	$\boldsymbol{X}$	$\mathbb{P}_{\boldsymbol{X}}(x)$
FF	0	$(1-p)^2$
FS	1	$(1-p)p$
SF		$p(1-p)$
SS	2	p^2

$n = 3$

ω	$\boldsymbol{X}$	$\mathbb{P}_{\boldsymbol{X}}(x)$
FFF	0	$(1-p)^3$
FFS		$(1-p)(1-p)p$
FSF	1	$(1-p)p(1-p)$
SFF		$p(1-p)(1-p)$
FSS		$(1-p)pp$
SFS	2	$p(1-p)p$
SSF		$pp(1-p)$
SSS	3	p^3

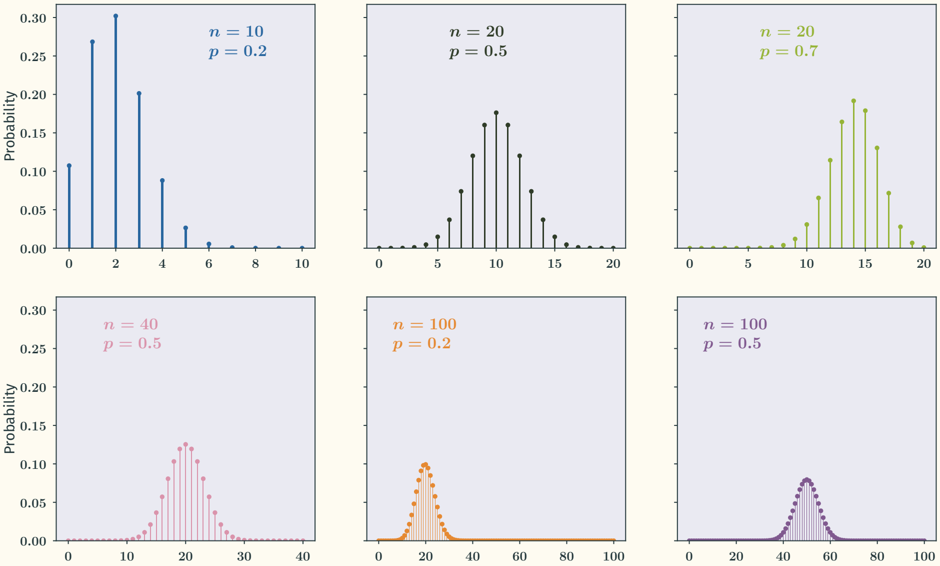
ω	$\boldsymbol{X}$	$\mathbb{P}_{\boldsymbol{X}}(x)$
FFFF	0	$(1-p)^4$
FFFS		$(1-p)(1-p)(1-p)p$
FFSF	1	$(1-p)(1-p)p(1-p)$
FSFF		$(1-p)p(1-p)(1-p)$
SFFF		$p(1-p)(1-p)(1-p)$
FFSS		$(1-p)(1-p)pp$
FSFS		$(1-p)p(1-p)p$
SFFS	2	$p(1-p)(1-p)p$
SSFF		$pp(1-p)(1-p)$
FSSF		$(1-p)pp(1-p)$
SFSF		$p(1-p)p(1-p)$
FSSS		$(1-p)ppp$
SFSS	3	$p(1-p)pp$
SSFS		$pp(1-p)p$
SSSF		$ppp(1-p)$
SSSS	4	p^4

The Binomial PMF

$$\mathbb{P}(k) = \mathbb{P}(\boldsymbol{X} = k) = \binom{n}{k} p^k (1-p)^{n-k},\ k = 0, 1, 2, 3, ..., n$$

$$\mathbb{P}_{\boldsymbol{X}}(x) = \mathbb{P}(\boldsymbol{X} = x) = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, 2, 3, ..., n$$

Different Binomial PMFs



Expectation & Variance of a Binomial Random Variable

Expectation

$$\mathbb{E}\left[\boldsymbol{X}\right] = np$$

Variance

$$\mathbb{V}\mathrm{ar}\left(\boldsymbol{X}\right) = np(1-p) = npq$$

Binomial Distribution Assumptions

Basic assumptions when we use the binomial distribution to solve problems:

1. There are a fixed number (n) of Bernoulli trials;
2. The outcome of the n trials are independent;
3. The probability of p is constant for each trial.

An Example in Lecture 1

Probability vs. Statistics

Probability: Previous studies showed that the drug was 80% effective. Then we can anticipate that for a study on 100 patients, on average 80 will be cured and at least 65 will be cured with 99.99% chance.

Statistics: We observe that 78/100 patients were cured by the drug. We will be able to conclude that we are 95% confident that for other studies the drug will be effective on between 69.88% and 86.11% of patients.

A Special Case of The Binomial Distribution

Experiment: monitoring number of emails received per day.

Question: Let the random variable $\boldsymbol{X}$ represent the number of email received per day. What is the probability distribution of $\boldsymbol{X}$?

Counting emails

Mar, 2023: 1414 days, 23,651 emails

$$\lambda = 16.73$$

$$\mathbb{E}\left[\boldsymbol{X}\right] = \lambda = np$$

Monitoring Emails

$$\lambda = \boxed{24} \cdot \boxed{\frac{\lambda}{24}} \longrightarrow \mathbb{P}_{\boldsymbol{X}}(k) = \binom{24}{k} \cdot \left(\frac{\lambda}{24}\right)^k \cdot \left(1 - \frac{\lambda}{24}\right)^{24-k}$$

$n = 24$
each hour is a
Bernoulli trial

This is p, the probability of
receiving an email in an hour

$$\lambda = \boxed{1440} \cdot \boxed{\frac{\lambda}{1440}} \longrightarrow \mathbb{P}_{\boldsymbol{X}}(k) = \binom{1440}{k} \cdot \left(\frac{\lambda}{1440}\right)^k \cdot \left(1 - \frac{\lambda}{1440}\right)^{1440-k}$$

$n = 1440$
each minute is a
Bernoulli trial

This is p, the probability of
receiving an email in a minute

$n \to \infty$
Binomial $\to$ Poisson

The Poisson Random Variables

Let $n \to \infty$ in a Binomial PMF:

$$\lim_{n\to\infty} \binom{n}{k} \cdot \left(\frac{\lambda}{n}\right)^k \cdot \left(1 - \frac{\lambda}{n}\right)^{n-k}$$

We get:

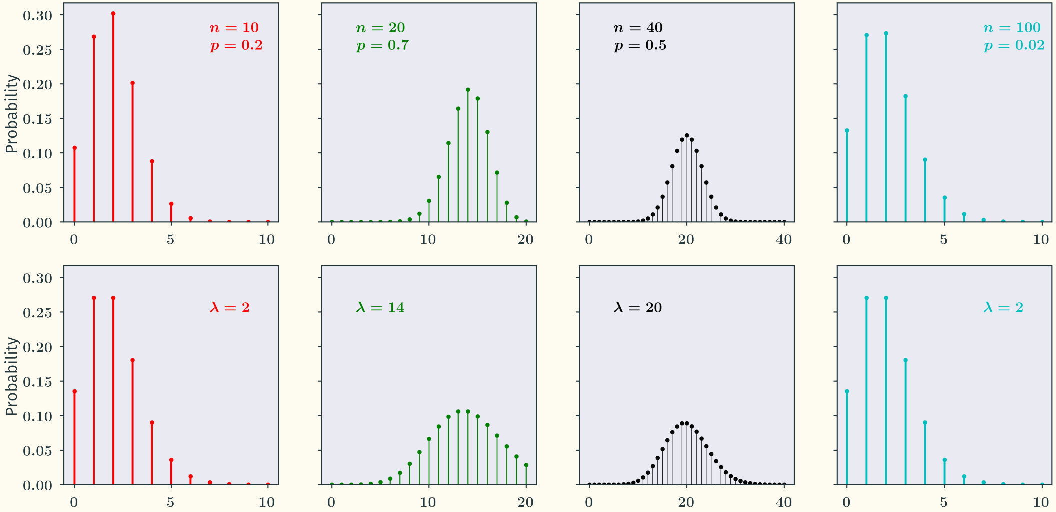
Poisson PMF

$$\mathbb{P}_{\boldsymbol{X}}(k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda},\ k = 0, 1, 2, 3, \ldots \qquad \mathbb{E}\left[\boldsymbol{X}\right] = \lambda,\ \mathbb{V}\mathrm{ar}\left(\boldsymbol{X}\right) = \lambda$$

Interpretation of n when $n \to \infty$:

1. n becomes "moments in time" where you can only receive one or zero emails.
2. You check your email continuously in time.

Binomial vs Poisson



Common usage:

- Monitor discrete rare event that happen in a fixed interval of time or space.
- In a binomial distribution where n is large and p is small, such that $0 < np < 10$, the binomial distribution is well approximated by the Poisson distribution with $\lambda = np$.

Examples of Poisson Distributions

A classical example: the number of Prussian soldiers accidentally killed by horse-kick.

# of deaths	Predicted probability	Expected # of occurrences	Actual # of occurrences
0	54.34	108.67	109
1	33.15	66.29	65
2	10.11	20.22	22
3	2.05	4.11	3
4	0.32	0.63	1
5	0.04	0.08	0
6	0.01	0.01	0

Examples of Poisson Distributions

Other examples:

- The number of mutations on a given strand of DNA per time/length unit.
- The number of stars found in a unit of space.
- The number of network failures per day.

Poisson Distribution Assumptions

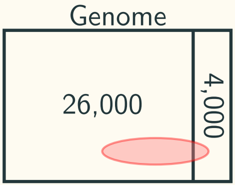
Basic assumptions when using the Poisson distribution:

1. The probability that a certain number of events occur within an interval is proportional to the length of the interval and is only dependent on the length of the interval;
2. Within a single interval, an infinite number of occurrences of the event are theoretically possible, *i.e.* not restricted to a fixed number of trials;
3. For a particular interval, the events occur independently both within and outside that interval.

$$\mathbb{P}_{\boldsymbol{X}}(k, \tau) = \mathbb{P}\left(\text{exactly } k \text{ events during an interval of length } \tau\right) = \frac{(\lambda\tau)^k}{k!} e^{-\lambda\tau}$$

The simplified gene ontology analysis

Experiment: There are 30,000 genes in the genome, and 4,000 of them are cell cycle related genes. If an experiment returns 500 genes of your interest, what is the probability that within this 500 genes, 30 of them are from those cell cycle related genes?



- Event of interest $A = \{$ choose 30 genes are from the 4,000 cell cycle related genes and 470 genes from the rest of the genome $\}$
- Sample space $\Omega = \{$ choose 500 genes from the genome $\}$

Hypergeometric Distributions

$$|A| = \binom{4000}{30} \cdot \binom{26000}{470}$$

$$|\Omega| = \binom{30000}{500}$$

Definition

An urn contains N balls, out of which K are red. We select n of the balls at random without replacement. The probability of drawing k red balls is:

$$\mathbb{P}_{\boldsymbol{X}}(k) = \frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}$$

Probability Distributions And Parameter(s)

- Probability distribution: describes the behaviour of the random variable.
- Parameter(s): numerical quantities that summarise the characteristics of a probability distribution.

Probability distribution and parameter(s)

	PMF $\mathbb{P}_{\boldsymbol{X}}(k)$	Parameter(s)
Geometric	$(1-p)^{k-1}p$	p
Bernoulli	p, if $k=1$ $1-p$, if $k=0$	p
Binomial	$\binom{n}{k}p^k(1-p)^{n-k}$	$n,\ p$
Poisson	$\frac{\lambda^k}{k!} \cdot e^{-\lambda}$	λ
Hypergeometric	$\frac{\binom{K}{k} \cdot \binom{N-K}{n-k}}{\binom{N}{n}}$	$N,\ K,\ n$