

AMA 2634

Introduction to Statistics

2024 Fall

Lecture 5

Discrete Random Variables

- Probability Distributions, mean and variance
- Binomial Distribution
- Multinomial Distribution
- Poisson Distribution
- Hypergeometric Distribution



A **random variable** is a variable whose values are determined by chance.

A **discrete probability distribution** consists of the values a random variable can assume and the corresponding probabilities of the values.

The sum of the **probabilities** of all events in a sample space add up to 1. Each probability is between 0 and 1, inclusively.

Example: Rolling a Die



Construct a probability distribution for rolling a single die.

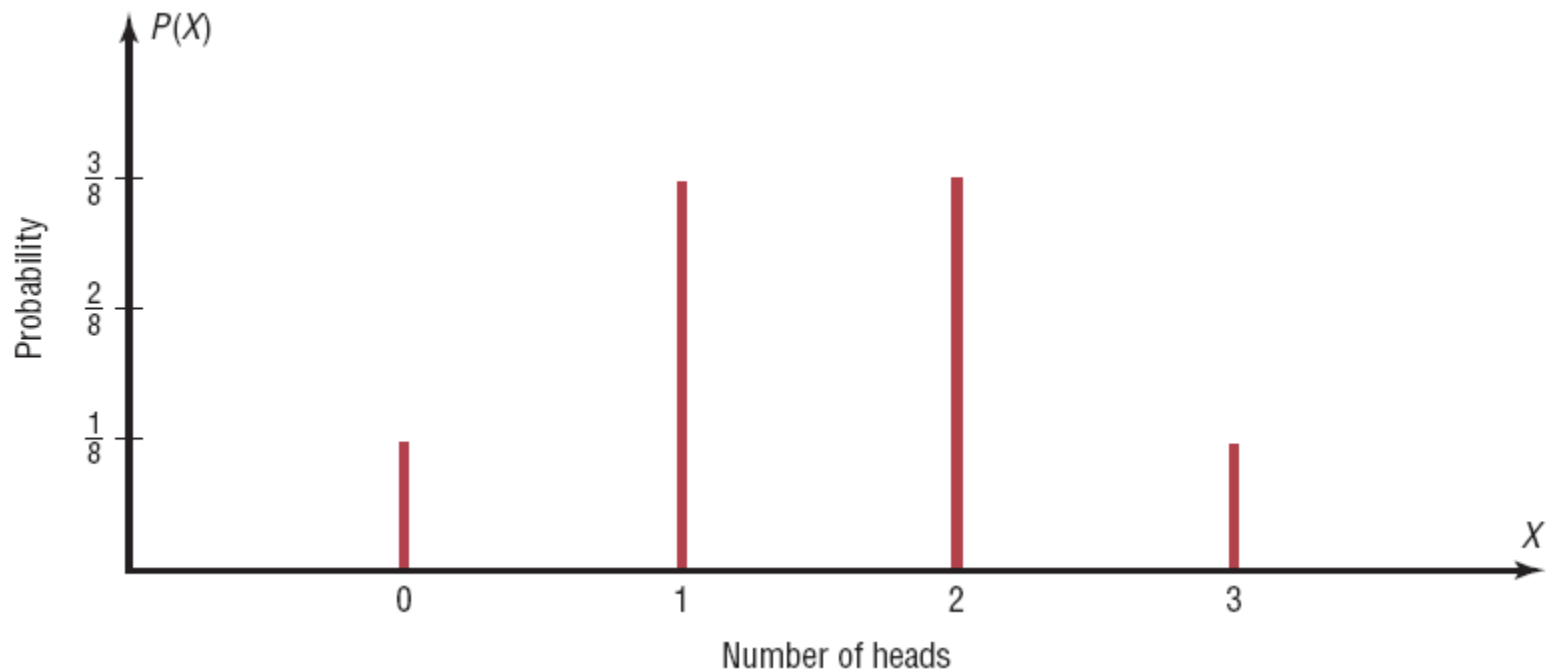
Outcome X	1	2	3	4	5	6
Probability $P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Example: Tossing Coins



Represent graphically the probability distribution for the sample space for tossing three coins.

Number of heads X	0	1	2	3
Probability $P(X)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$





MEAN: $\mu = \sum X \cdot P(X)$

VARIANCE:

$$\sigma^2 = \sum [X^2 \cdot P(X)] - \mu^2$$

Example: Rolling a Die



Find the mean of the number of spots that appear when a die is tossed.

Outcome X	1	2	3	4	5	6
Probability $P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\mu = \sum X \cdot P(X)$$

$$= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6}$$

$$= \frac{21}{6} = \boxed{3.5}$$

Example: Rolling a Die



Compute the variance and standard deviation

Outcome X	1	2	3	4	5	6
Probability $P(X)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

$$\sigma^2 = \sum [X^2 \cdot P(X)] - \mu^2$$

$$\begin{aligned}\sigma^2 &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} \\ &\quad + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} - (3.5)^2\end{aligned}$$

$$\sigma^2 = \boxed{2.9}, \quad \sigma = \boxed{1.7}$$

Example: On Hold for Talk Radio



A talk radio station has four telephone lines. If the host is unable to talk (i.e., during a commercial) or is talking to a person, the other callers are placed on hold. When all lines are in use, others who are trying to call in get a busy signal. The probability that 0, 1, 2, 3, or 4 people will get through is shown in the distribution. Find the variance and standard deviation for the distribution.

X	0	1	2	3	4
$P(X)$	0.18	0.34	0.23	0.21	0.04

X	0	1	2	3	4
$P(X)$	0.18	0.34	0.23	0.21	0.04

$$\begin{aligned}\mu &= 0(0.18) + 1(0.34) + 2(0.23) \\ &\quad + 3(0.21) + 4(0.04) = 1.6\end{aligned}$$

$$\begin{aligned}\sigma^2 &= 0^2(0.18) + 1^2(0.34) + 2^2(0.23) \\ &\quad + 3^2(0.21) + 4^2(0.04) - (1.6)^2\end{aligned}$$

$$\sigma^2 = \boxed{1.2}, \quad \sigma = \boxed{1.1}$$



A talk radio station has four telephone lines. If the host is unable to talk (i.e., during a commercial) or is talking to a person, the other callers are placed on hold. When all lines are in use, others who are trying to call in get a busy signal.

Should the station have considered getting more phone lines installed?

No, the four phone lines should be sufficient.

The mean number of people calling at any one time is 1.6.

Since the standard deviation is 1.1, most callers would be accommodated by having four phone lines because $\mu + 2\sigma$ would be

$$1.6 + 2(1.1) = 1.6 + 2.2 = 3.8.$$

Very few callers would get a busy signal since at least 75% of the callers would either get through or be put on hold. (By Chebyshev's theorem)



The **expected value**, or **expectation**, of a discrete random variable of a probability distribution is the theoretical average of the variable.

The expected value is, by definition, the mean of the probability distribution.

$$E(X) = \mu = \sum X \cdot P(X)$$

Binomial Probability Distribution

Criteria for a Binomial Probability Experiment

An experiment is said to be a **binomial experiment** if

1. The experiment is performed a fixed number of times. Each repetition of the experiment is called a **trial**.
2. The trials are independent. This means the outcome of one trial will not affect the outcome of the other trials.
3. For each trial, there are two mutually exclusive (or disjoint) outcomes, success or failure.
4. The probability of success is fixed for each trial of the experiment.

A chance experiment with exactly two possible outcomes

Who is Bernoulli?



1654-1705

- The Bernoulli trial is called after the famous Swiss mathematician Jacob Bernoulli studied it in his book "*The Art of Conjecture*" published in 1713.
- He was compelled to study philosophy and theology by his parents, which he greatly resented, and he graduated from the University of Basel with a master's degree in philosophy in 1671 and a licentiate in theology in 1676.



Polynomial expression containing two terms, for example, $x + y$. The binomial theorem, or binomial formula, gives the expansion of the n th power of a binomial $x + y$ for $n=1,2,3,\dots$

$$(x + y)^0 = 1$$

$$(x + y)^1 = x + y$$

$$(x + y)^2 = x^2 + 2xy + y^2$$

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$(x + y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

$$(x + y)^5 = x^5 + 5x^4y + 10x^3y^2 + 10x^2y^3 + 5xy^4 + y^5$$

Pascal Triangle (Blaise Pascal)



If we detach the **coefficients** and display them in a triangular array, we get the **Pascal's triangle**

						1					
					1		1				
			1		2		1				
		1		3		3		1			
	1		4		6		4		1		
1		5		10		10		5		1	



1623-1662

Notation Used in the Binomial Probability Distribution

- There are n independent trials of the experiment.
- Let p denote the probability of success so that $1 - p$ is the probability of failure.
- Let X be a binomial random variable that denotes the number of successes in n independent trials of the experiment. So, $0 \leq x \leq n$.



Binomial Probability Distribution Function

The probability of obtaining x successes in n independent trials of a binomial experiment is given by

$$P(x) = {}_nC_x p^x (1 - p)^{n-x} \quad x = 0, 1, 2, \dots, n \quad (1)$$

where p is the probability of success.



Which of the following are binomial experiments?

(a) A player rolls a pair of fair die 10 times. The number X of 7's rolled is recorded. **Binomial experiment**

(b) The 11 largest airlines had an on-time percentage of 84.7% in November, 2001 according to the Air Travel Consumer Report. In order to assess reasons for delays, an official with the FAA randomly selects flights until she finds 10 that were not on time. The number of flights X that need to be selected is recorded.

Not a binomial experiment – not a fixed number of trials.

(c) In a class of 30 students, 55% are female. The instructor randomly selects 4 students. The number X of females selected is recorded. **Not a binomial experiment – the trials are not independent.**

Quality Control (Acceptance Sampling)

One of the major fields of statistical quality control is acceptance sampling.

A company receives a shipment of goods. It samples the shipment and either accepts it as conforming to its standards or rejects it.

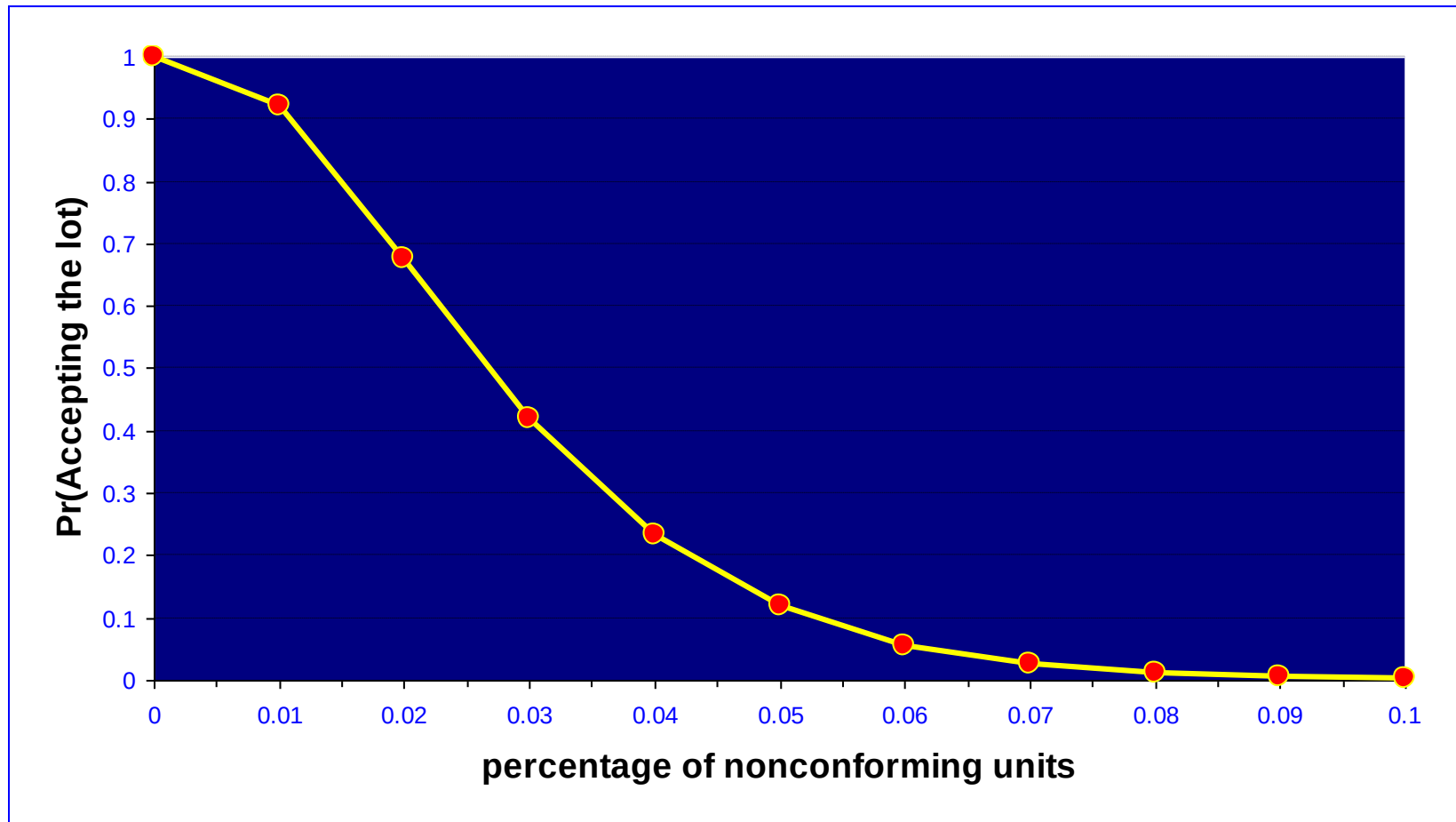
OC curve: reviews how the probability of accepting a lot varies with the quality of the material offered for inspection.

Example of OC curve



$n = 100, c = 2$

(n = sample size, c = number of nonconforming units that cannot be exceeded without the lot's being rejected)



Marketing research



STARBUCKS.CO

Survey 100 people to study
whether they like coffee





According to the Experian Automotive, 35% of all car-owning households have three or more cars.

(a) In a random sample of 20 car-owning households, what is the probability that exactly 5 have three or more cars?

$$\begin{aligned} P(5) &= {}_{20}C_5 (0.35)^5 (1 - 0.35)^{20-5} \\ &= 0.1272 \end{aligned}$$

(b) In a random sample of 20 car-owning households, what is the probability that less than 4 have three or more cars?

$$\begin{aligned} P(X < 4) &= P(X \leq 3) \\ &= P(0) + P(1) + P(2) + P(3) \\ &= 0.0444 \end{aligned}$$

(c) In a random sample of 20 car-owning households, what is the probability that at least 4 have three or more cars?

$$\begin{aligned} P(X \geq 4) &= 1 - P(X \leq 3) \\ &= 1 - 0.0444 \\ &= 0.9556 \end{aligned}$$

Mean (or Expected Value) and Standard Deviation of a Binomial Random Variable

A binomial experiment with n independent trials and probability of success p has a mean and standard deviation given by the formulas

$$\mu_X = np \quad \text{and} \quad \sigma_X = \sqrt{np(1 - p)} \quad (2)$$

EXAMPLE *Finding the Mean and Standard Deviation of a Binomial Random Variable*

According to the Experian Automotive, 35% of all car-owning households have three or more cars. In a simple random sample of 400 car-owning households, determine the mean and standard deviation number of car-owning households that will have three or more cars.

$$\begin{aligned} \mu_X &= np \\ &= (400)(0.35) \\ &= 140 \end{aligned}$$

$$\begin{aligned} \sigma_X &= \sqrt{np(1 - p)} \\ &= \sqrt{(400)(0.35)(1 - 0.35)} \\ &= 9.54 \end{aligned}$$

Which one is better? 2 Engines OR 4 Engines?



Airbus A300



Douglas DC-6A



Suppose that an airplane engine will fail, when in flight, with probability $1 - p$ independently from engine to engine; suppose that the airplane will make a successful flight if at least 50 percent of its engines remain operative. For what values of p is a 4-engine plane preferable to a 2-engine plane?

The solution

- Prob. that a 4-engine plane makes a successful flight = ?
- Prob. that a 2-engine plane makes a successful flight = ?
- Comparison!

The probability that a 4-engine plane makes a successful flight is

$$\begin{aligned} & \binom{4}{2} p^2 (1-p)^2 + \binom{4}{3} p^3 (1-p) + \binom{4}{4} p^4 (1-p)^0 \\ &= 6p^2(1-p)^2 + 4p^3(1-p) + p^4 \end{aligned}$$

whereas the corresponding probability for a 2-engine plane is

$$\binom{2}{1} p(1-p) + \binom{2}{2} p^2 = 2p(1-p) + p^2$$

Hence the 4-engine plane is safer if

$$6p^2(1-p)^2 + 4p^3(1-p) + p^4 \geq 2p(1-p) + p^2$$

or equivalently,

$$p \geq \frac{2}{3}$$

- The **multinomial distribution** is similar to the binomial distribution but has the advantage of allowing one to compute probabilities when there are more than two outcomes.

$$P(X) = \frac{n!}{X_1! X_2! X_3! \dots X_k!} \cdot p_1^{X_1} \cdot p_2^{X_2} \cdot p_3^{X_3} \dots p_k^{X_k}$$

- The binomial distribution is a special case of the multinomial distribution.

In a large city, 50% of the people choose a movie, 30% choose dinner and a play, and 20% choose shopping as a leisure activity. If a sample of 5 people is randomly selected, find the probability that 3 are planning to go to a movie, 1 to a play, and 1 to a shopping mall.

$$P(X) = \frac{n!}{X_1! X_2! X_3! \cdots X_k!} \cdot p_1^{X_1} \cdot p_2^{X_2} \cdot p_3^{X_3} \cdots p_k^{X_k}$$

$$P(X) = \frac{5!}{3!1!1!} \cdot (0.50)^3 (0.30)^1 (0.20)^1 = \boxed{0.15}$$

- The **Poisson distribution** is a distribution useful when n is large and p is small and when the independent variables occur over a period of time.
- The Poisson distribution can also be used when a density of items is distributed over a given area or volume, such as the number of plants growing per acre or the number of defects in a given length of videotape.

Poisson Distribution

The probability of X occurrences in an interval of time, volume, area, etc., for a variable, where λ (Greek letter lambda) is the mean number of occurrences per unit (time, volume, area, etc.), is

$$P(X; \lambda) = \frac{e^{-\lambda} \lambda^X}{X!} \quad \text{where } X = 0, 1, 2, \dots$$

The letter e is a constant approximately equal to 2.7183.

If there are 200 typographical errors randomly distributed in a 500-page manuscript, find the probability that a given page contains exactly 3 errors.

- First, find the mean number of errors. With 200 errors distributed over 500 pages, each page has an average of errors.

$$\lambda = \frac{200}{500} = 0.4$$

- Then probability that a given page contains exactly 3 errors.

$$P(X; \lambda) = \frac{e^{-\lambda} \lambda^X}{X!} = \frac{e^{-0.4} (0.4)^3}{3!} = \boxed{0.0072}$$

- Thus, there is less than 1% chance that any given page will contain exactly 3 errors.



- The **hypergeometric distribution** is a distribution of a variable that has two outcomes when sampling is done without replacement.



Given a population with only two types of objects (females and males, defective and nondefective, successes and failures, etc.), such that there are a items of one kind and b items of another kind and $a + b$ equals the total population, the probability $P(X)$ of selecting without replacement a sample of size n with X items of type a and $n - X$ items of type b is

$$P(X) = \frac{{}_a C_X \cdot {}_b C_{n-X}}{{}_{a+b} C_n}$$

Example: House Insurance



A recent study found that 2 out of every 10 houses in a neighborhood have no insurance. If 5 houses are selected from 10 houses, find the probability that exactly 1 will be uninsured.

$$a = 2, a + b = 10 \rightarrow b = 8, \quad X = 1, n = 5 \rightarrow n - X = 4$$

$$P(X) = \frac{{}_a C_X \cdot {}_b C_{n-X}}{{}_{a+b} C_n}$$

$$P(X) = \frac{{}_2 C_1 \cdot {}_8 C_4}{{}_{10} C_5} = \frac{2 \cdot 70}{252} = \frac{140}{252} = \boxed{\frac{5}{9}}$$

Example: Stock



The Dow Jones Industrial Average (DJIA) is a collection of thirty publicly traded companies that are meant to be representative of the United States economy. In one certain month 18 of the 30 stocks in the DJIA increased in value. What is the probability that an investor randomly invests in four stocks at the beginning of this month and three of the stocks increased in value?

$$N = 30; n = 4; k = 18; x = 3$$

$$\begin{aligned} P(x) &= \frac{{}_k C_x {}_{N-k} C_{n-x}}{{}_N C_n} \\ &= \frac{{}_{18} C_3 {}_{30-18} C_{4-3}}{{}_{30} C_4} \\ &= \frac{(816)(12)}{27,405} \\ &= 0.3573 \end{aligned}$$



Population Size of a certain specie in a particular territory?

Capture-recapture Method

<http://www.youtube.com/watch?v=lKyj7gEAAS8>



An unknown number, say N , of animals inhabit a certain region. To obtain some information about the population size, ecologists often perform the following experiment: They first catch a number, say r , of these animals, mark them in some manner, and release them. After allowing the marked animals time to disperse throughout the region, a new catch of size, say n , is made. Let X denote the number of marked animals in this second capture. If we assume that the population of animals in the region remained fixed between the time of the two catches and that each time an animal was caught it was equally likely to be any of the remaining uncaught animals, it follows that X is a hypergeometric random variable.

The question is: **How to estimate N ?**



Country A inadvertently launches 10 guided missiles - 6 armed with nuclear warheads - at Country B.

In response, Country B fires 7 anti-ballistic missiles, each of which will destroy exactly one of the incoming rockets.

The anti-ballistic missiles have no way of detecting, though, which of the 10 rockets are carrying nuclear warheads.

What are the chances that Country B will be hit by at least one nuclear missile?

A purchaser of electrical components buys them in lots of size **10**. It is his policy to inspect **3** components randomly from a lot and to accept the lot only if all 3 are non-defective.

If 30 percent of the lots have **4** defective components and 70 percent have only **1**, what proportion of lots does the purchaser reject?