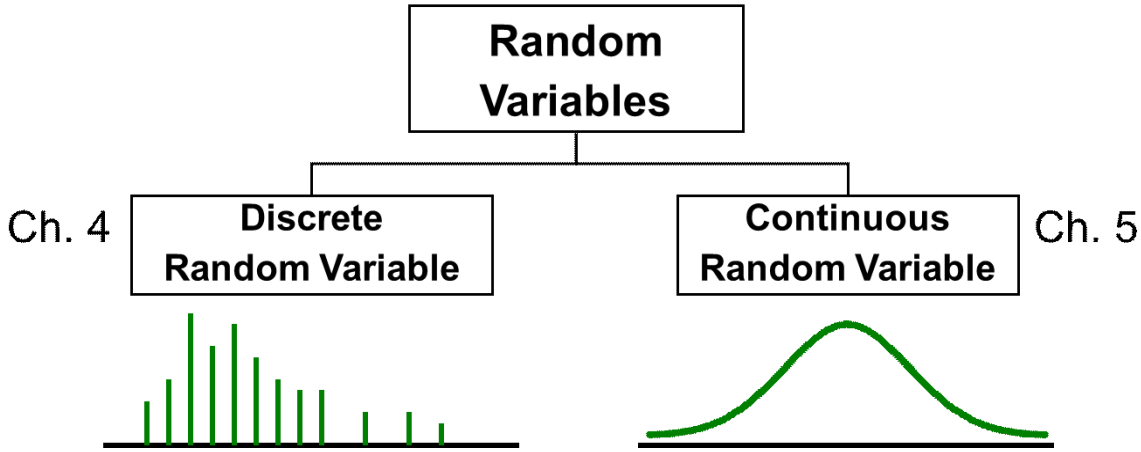
Lecture 3. Discrete Random Variables (Chapter 4)

Ping Yu

HKU Business School
The University of Hong Kong

Plan of This Lecture



- Random Variables
- Probability Distributions for Discrete Random Variables
- Properties of Discrete Random Variables
- Binomial Distribution
- Poisson Distribution
- Hypergeometric Distribution
- Jointly Distributed Discrete Random Variables

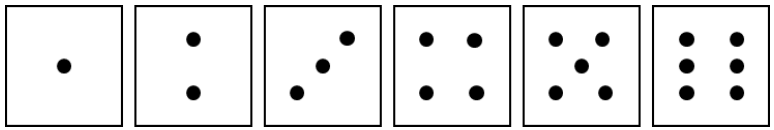
Random Variables

Discrete Random Variables

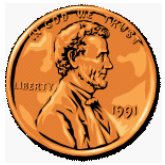
- A random variable (r.v.) is a variable that takes on numerical values realized by the outcomes in the sample space generated by a random experiment.
  - Mathematically, a random variable is a function from S to $\mathbb{R}$.
  - In this and next lectures, we use capital letters, such as X, to denote the random variable, and the corresponding lowercase letter, x, to denote a possible value.
- A discrete random variable is a random variable that can take on no more than a countable number of values.
  - "Countable" includes "finite" and "countably infinite" (see the examples below).

Examples of Discrete Random Variables

- Roll a die twice. Let X be the number of times 4 comes up (then X could be 0, 1, or 2 times).



- Toss a coin 5 times. Let X be the number of heads (then $X = 0, 1, 2, 3, 4,$ or 5).



- The number of customers visiting a store in one day.
- The number of claims on a medical insurance policy in a particular year.
- X in the above two examples can be $0, 1, 2, \cdots$

Continuous Random Variables

- A continuous random variable is a random variable that can take any value in an interval (i.e., for any two values, there is some third value that lies between them).
  - Possible values are measured on a continuum, e.g., the yearly income for a family, the highest temperature in one day, weight of packages filled by a mechanical filling process, etc.
  - The probability can only be assigned to a range of values since the probability of a single value is always zero.
- Recall the distinction between discrete numerical variables and continuous numerical variables in Lecture 1.
- Modeling a r.v. as continuous is usually for convenience as the differences between adjacent discrete values (e.g., $35,276.21 and $35,276.22) are of no importance.
- On the other hand, we model a r.v. as discrete when probability statements about the individual possible outcomes have worthwhile meaning.

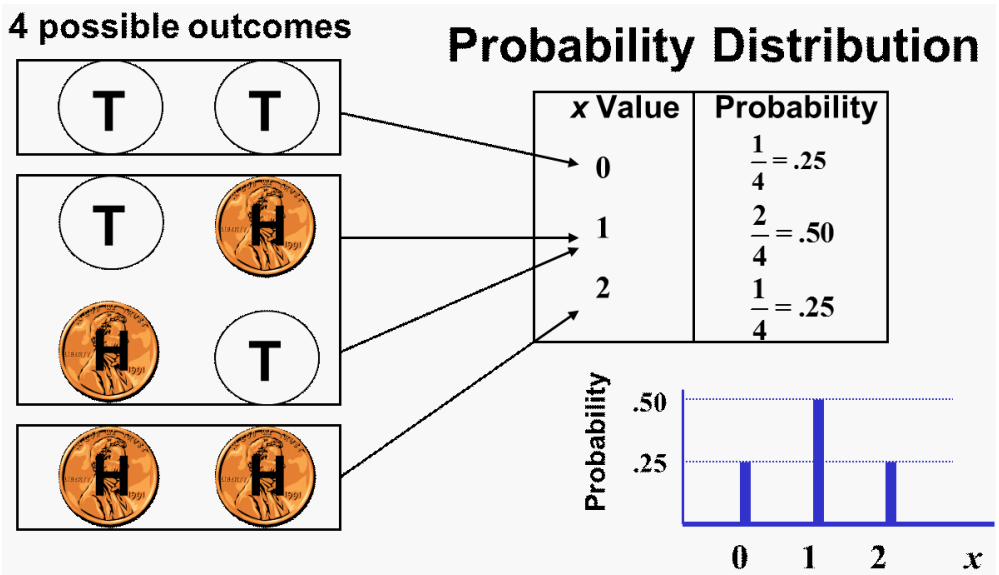
Probability Distributions for Discrete Random Variables

Probability Distribution Function

- The probability distribution (function), $p(x)$, of a discrete r.v. X represents the probability that X takes the value x, as a function of x, i.e.,

$$p(x) = P(X = x) \text{ for } \underline{\text{all}} \text{ values of } x.$$

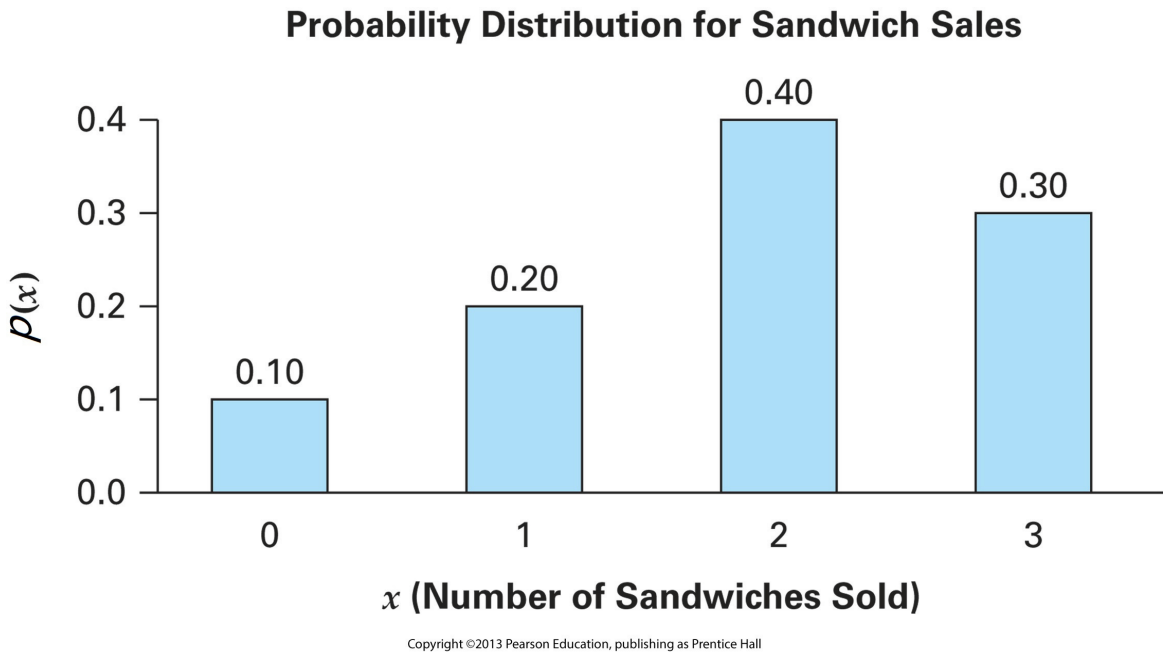
  - Sometimes, the probability distribution of a discrete r.v. is called the probability mass function (pmf).
  - Note that $X = x$ must be an event; otherwise, $P(X = x)$ is not well defined.
- PMF can be shown algebraically, graphically, or with a table.
- Experiment: Toss 2 Coins. Let $X = \#$ heads.



Example 4.1: Number of Product Sales

Table 4.1 Probability Distribution for Example 4.1

x	$p(x)$
0	0.10
1	0.20
2	0.40
3	0.30



Properties of PMF

- $p(x)$ must satisfy the following properties (implied by the probability postulates in Lecture 2):

1. $0 \leq p(x) \leq 1$ for any value x,
2. $\sum_{x \in \mathscr{S}} p(x) = 1$, where $\mathscr{S}$ is called the support of X, i.e., the set of all x values such that $p(x) > 0$.

- In the tossing-two-coins example, $\mathscr{S} = \{0, 1, 2\}$.

- Notation: I will use $p(x)$ and p (rather than $P(x)$ and P as in the textbook) for pmf and an interested probability to avoid confusion with the probability symbol P.

Cumulative Distribution Function

- The cumulative distribution function (cdf), $F(x_0)$, of a r.v. X, represents the probability that X does not exceed the value x_0, as a function of x_0, i.e.,

$$F(x_0) = P(X \leq x_0).$$

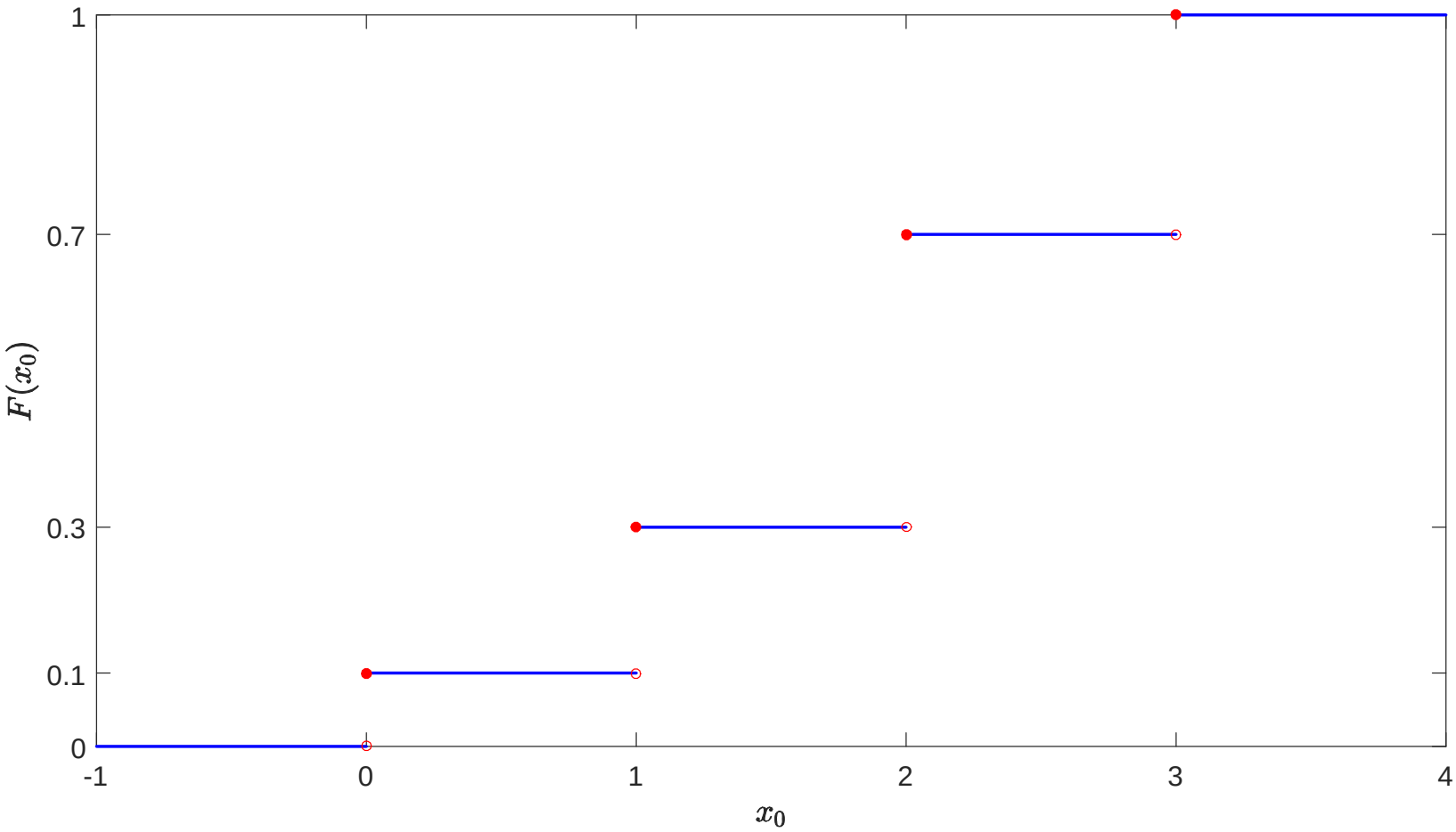
  - The definition of cdf applies to both discrete and continuous r.v.'s, and $x_0 \in \mathbb{R}$.
  - $F(x_0)$ for a discrete r.v. is a step function with jumps only at support points in $\mathscr{S}$. [figure here]
  - $p(\cdot)$ and $F(\cdot)$ are probabilistic counterparts of histogram and ogive in Lecture 1.

- Relationship between pmf and cdf for discrete r.v.'s:

$$F(x_0) = \sum_{x \leq x_0} p(x).$$

- Derived Properties of cdf: (i) $0 \leq F(x_0) \leq 1$ for any x_0; (ii) if $x_0 < x_1$, $F(x_0) \leq F(x_1)$, i.e., $F(\cdot)$ is an (weakly) increasing function.
- From the figure in the next slide, we can also see (iii) $F(x_0)$ is right continuous, i.e., $\lim_{x \downarrow x_0} F(x) = F(x_0)$; (iv) $\lim_{x_0 \to -\infty} F(x_0) = 0$ and $\lim_{x_0 \to \infty} F(x_0) = 1$.

Example 4.1 Continued



Properties of Discrete Random Variables

Mean

- The pmf contains all information about the probability properties of a discrete r.v., but it is desirable to have some summary measures of the pmf's characteristics.
- The mean (or expected value, or expectation), $E[X]$, of a discrete r.v. X is defined as

$$E[X] = \mu := \sum_{x \in \mathscr{S}} x p(x).$$

- The mean of X is the same as the population mean in Lecture 1, $\mu = \frac{\sum_{i=1}^{N} x_i}{N}$, but we use the probability language here: think of $E[X]$ in terms of relative frequencies,

$$\frac{\sum_{i=1}^{N} x_i}{N} = \sum_{x \in \mathscr{S}} x \frac{N_x}{N},$$

weighting each possible value x by its probability.
- In other words, the mean of X is a weighted average of all possible values of X.
- In the tossing-two-coins example, the expected number of heads is

$$E[X] = 0 \times 0.25 + 1 \times 0.5 + 2 \times 0.25 = 1.$$

Variance

- The variance, $Var(X)$, of a discrete r.v. X is defined as

$$Var(X) = \sigma^2 := E\left[(X-\mu)^2\right] = \sum_{x\in\mathcal{S}} (x-\mu)^2 p(x).$$

- This definition of $Var(X)$ is the same as the population variance in Lecture 1.
- It is not hard to see that

$$\begin{aligned}\sigma^2 &= \sum_{x\in\mathcal{S}} (x-\mu)^2 p(x) = \sum_{x\in\mathcal{S}} x^2 p(x) - 2\mu \sum_{x\in\mathcal{S}} x p(x) + \mu^2 \sum_{x\in\mathcal{S}} p(x)\\ &= E\left[X^2\right] - 2\mu E[X] + \mu^2 = E\left[X^2\right] - 2\mu^2 + \mu^2\\ &= E\left[X^2\right] - \mu^2,\end{aligned}$$

i.e., the second moment $-$ first moment[2],[1] where in the third equality, $p(x)$ is the probability of $X = x$,[2] and $\sum_{x\in\mathcal{S}} p(x) = 1$.

[1] σ^2 is also called the second central moment. The term "moment" is borrowed from physics.

[2] What will happen if X can take both 1 and -1? $\sum_{x^2=1} x^2 \times P\left(X^2 = x^2\right) = \sum_{x^2=1} x^2 \times (p(1)+p(-1))$ $= p(1)+p(-1) = 1^2 \times p(1) + (-1)^2 \times p(-1)$.

Standard Deviation

- The standard deviation, $\sigma = \sqrt{\sigma^2}$, is the same as the population standard deviation in Lecture 1.
- In the tossing-two-coins example, recall that $\mu = 1$, so

$$\begin{aligned}
\sigma &= \sqrt{\sum_{x\in\{0,1,2\}} (x-\mu)^2 p(x)} \\
&= \sqrt{(0-1)^2 \times 0.25 + (1-1)^2 \times 0.5 + (2-1)^2 \times 0.25} \\
&= \sqrt{0.5} \\
&= 0.707.
\end{aligned}$$

Mean of Functions of a R.V.

- For a function of X, $g(X)$, its mean, $E[g(X)]$, is defined as

$$E[g(X)] := \sum_{x \in \mathscr{S}} g(x) p(x).$$

  - e.g., X is the time to complete a contract, and $g(X)$ is the cost when the completion time is X; we want to know the expected cost.

- $E[g(X)] \neq g(E[X])$ in general, e.g., if $g(X) = X^2$, then

$$E[g(X)] - g(E[X]) = E\left[X^2\right] - \mu^2 = \sigma^2 > 0.$$

  - However, when $g(X)$ is linear in X, $E[g(X)] = g(E[X])$.

Mean and Variance of Linear Functions

- For $Y = a + bX$ with a and b being constant fixed numbers,

$$\mu_Y := E[Y] = E[a + bX] = a + bE[X] =: a + b\mu_X,$$
$$\sigma_Y^2 := Var(Y) = Var(a + bX) = b^2 Var(X) =: b^2\sigma_X^2,$$

and

$$\sigma_Y = \sqrt{Var(Y)} = |b|\sigma_X.$$

- The proof can follow similar steps as in the last[3] slide. [exercise]
- The constant a will not contribute to the variance of Y.

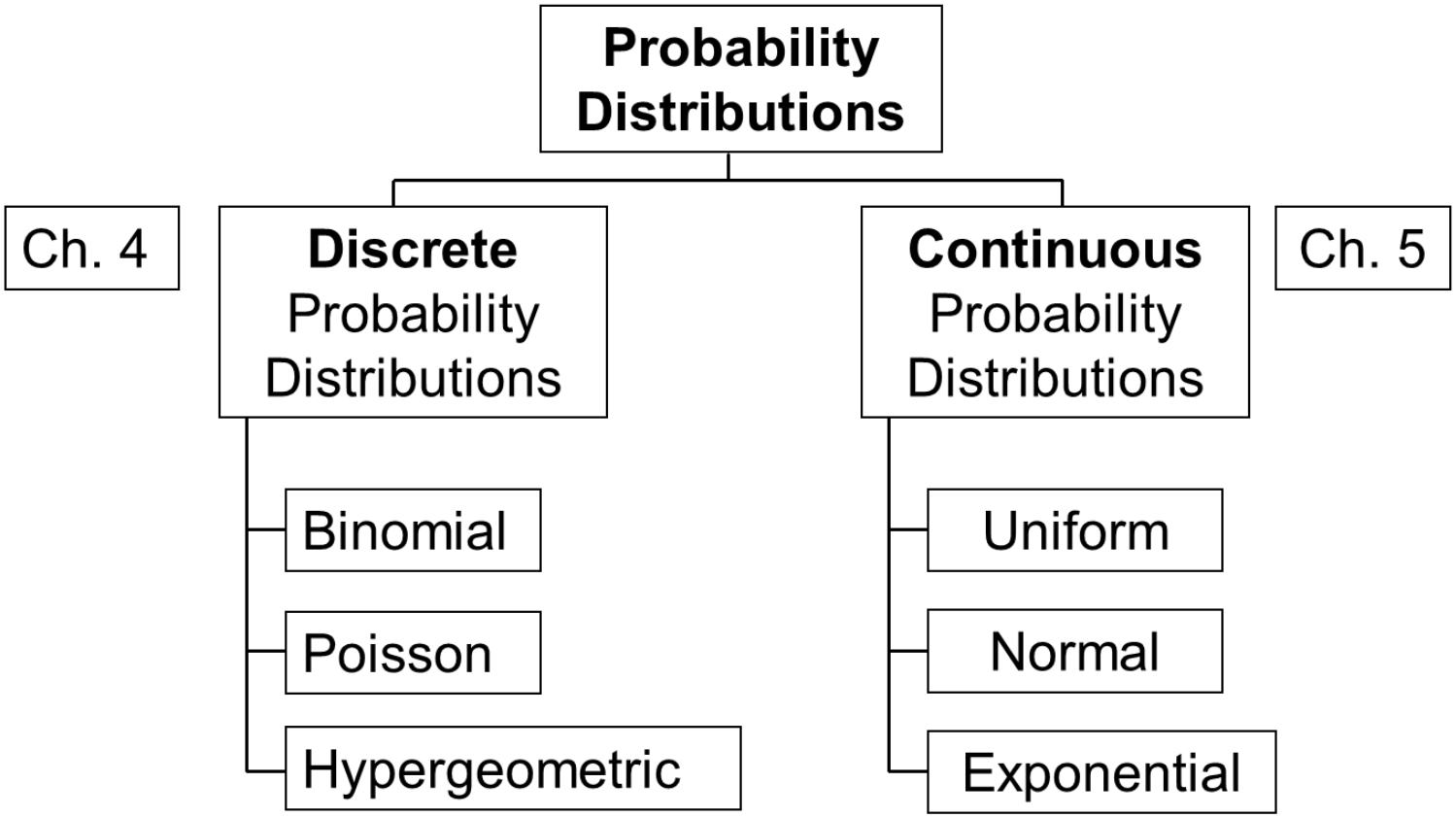
- Some Special Linear Functions:
- If $b = 0$, i.e., $Y = a$, then $E[a] = a$ and $Var(a) = 0$.
- If $a = 0$, i.e., $Y = bX$, then $E[bX] = bE[X]$ and $Var(bX) = b^2 Var(X)$.
- If $a = -\mu_X/\sigma_X$ and $b = 1/\sigma_X$, i.e., $Y = \frac{X-\mu_X}{\sigma_X}$ is the z-score of X, then

$$E\left[\frac{X-\mu_X}{\sigma_X}\right] = \frac{\mu_X}{\sigma_X} - \frac{\mu_X}{\sigma_X} = 0$$

and

$$Var\left(\frac{X-\mu_X}{\sigma_X}\right) = \frac{Var(X)}{\sigma_X^2} = 1.$$

Probability Distributions We Will Study



Binomial Distribution

Bernoulli Distribution

- The Bernoulli r.v. is a r.v. taking only two values, 0 and 1, labeled as "failure" and "success". [figure here]
- If the probability of success, $p(1) = p$, then the probability of failure, $p(0) = 1 - p(1) = 1 - p$. This distribution is known as the Bernoulli distribution, and we denote a r.v. X with this distribution as $X \sim \text{Bernoulli}(p)$.
- The mean of a Bernoulli(p) r.v. X is

$$\mu_X = E[X] = 1 \times p + 0 \times (1-p) = p,$$

and the variance is

$$\begin{aligned} \sigma_X^2 &= Var(X) = (1-p)^2 \times p + (0-p)^2 \times (1-p) \\ &= p(1-p). \end{aligned}$$

- When $p = 0.5$, σ_X^2 achieves its maximum; when $p = 0$ and 1, $\sigma_X^2 = 0$. [why?]

History of the Bernoulli Distribution

Jacob Bernoulli (1655-1705), Swiss

- Jacob Bernoulli (1655-1705) was one of the many prominent mathematicians in the Bernoulli family.

Binomial Distribution

- The binomial r.v. X is the number of successes in n independent trials of a Bernoulli(p) r.v., denoted as $X \sim \text{Binomial}(n,p)$.
- Denote X_i as the outcome in the ith trial, then the binomial r.v. $X = \sum_{i=1}^{n} X_i$.
- After some thinking, we can figure out that the number of sequences with x successes in n trials is C_x^n, and the probability of any sequence with x successes is $p^x (1-p)^{n-x}$ by the multiplication rule.
- By the addition rule, the binomial distribution is

$$p(x|n,p) = C_x^n p^x (1-p)^{n-x}, x = 0,1,\cdots,n.$$

Conditions to Define a Binomial Distribution

1. A fixed number of observations n.
   - e.g., 15 tosses of a coin; ten light bulbs taken from a warehouse.
2. Two mutually exclusive and collectively exhaustive categories.
   - e.g., head or tail in each toss of a coin; defective or not defective light bulb.
   - Generally called “success” and “failure”.
   - Probability of success is p, probability of failure is $1-p$.
3. Constant probability for each observation.
   - e.g., probability of getting a tail is the same each time we toss the coin.
4. Observations are independent.
   - The outcome of one observation does not affect the outcome of the other.

Possible Binomial Distribution Settings:

- A manufacturing plant labels items as either defective or acceptable.
- A firm bidding for contracts will either get a contract or not.
- A marketing research firm receives survey responses of “yes I will buy” or “no I will not”.
- New job applicants either accept the offer or reject it.

Calculating a Binomial Probability

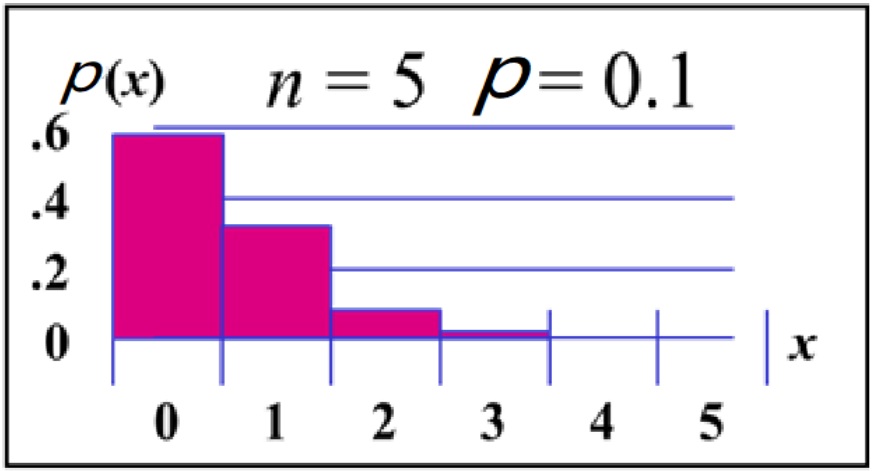
- What is the probability of one success in five observations if the probability of success is 0.1?
- $x = 1, n = 5$, and $p = 0.1$.
- So

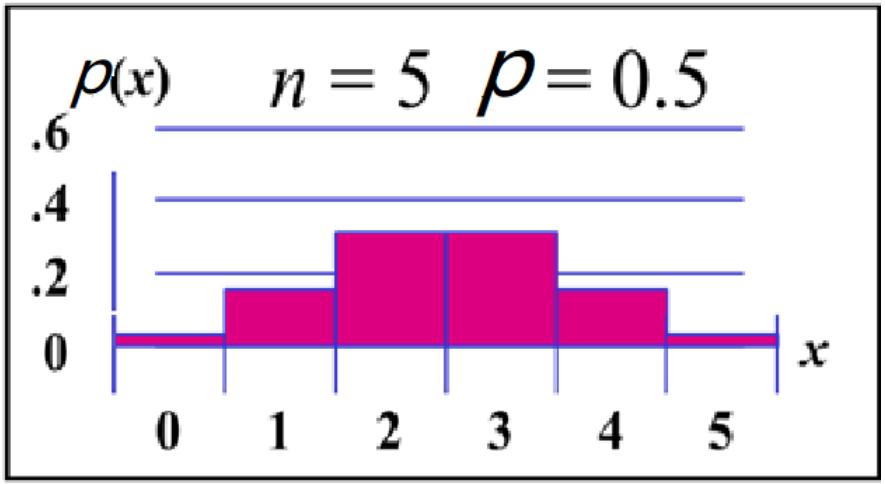
$$\begin{aligned} p(x|n,p) &= \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \\ &= \frac{5!}{1!(5-1)!}0.1^1(1-0.1)^{5-1} \\ &= 5 \times 0.1 \times 0.9^4 \\ &= 0.32805. \end{aligned}$$

- This probability can be easily obtained from Binomial Tables (Appendix Table 2).

Shape of Binomial Distribution

- The shape of the binomial distribution depends on the values of p and n.





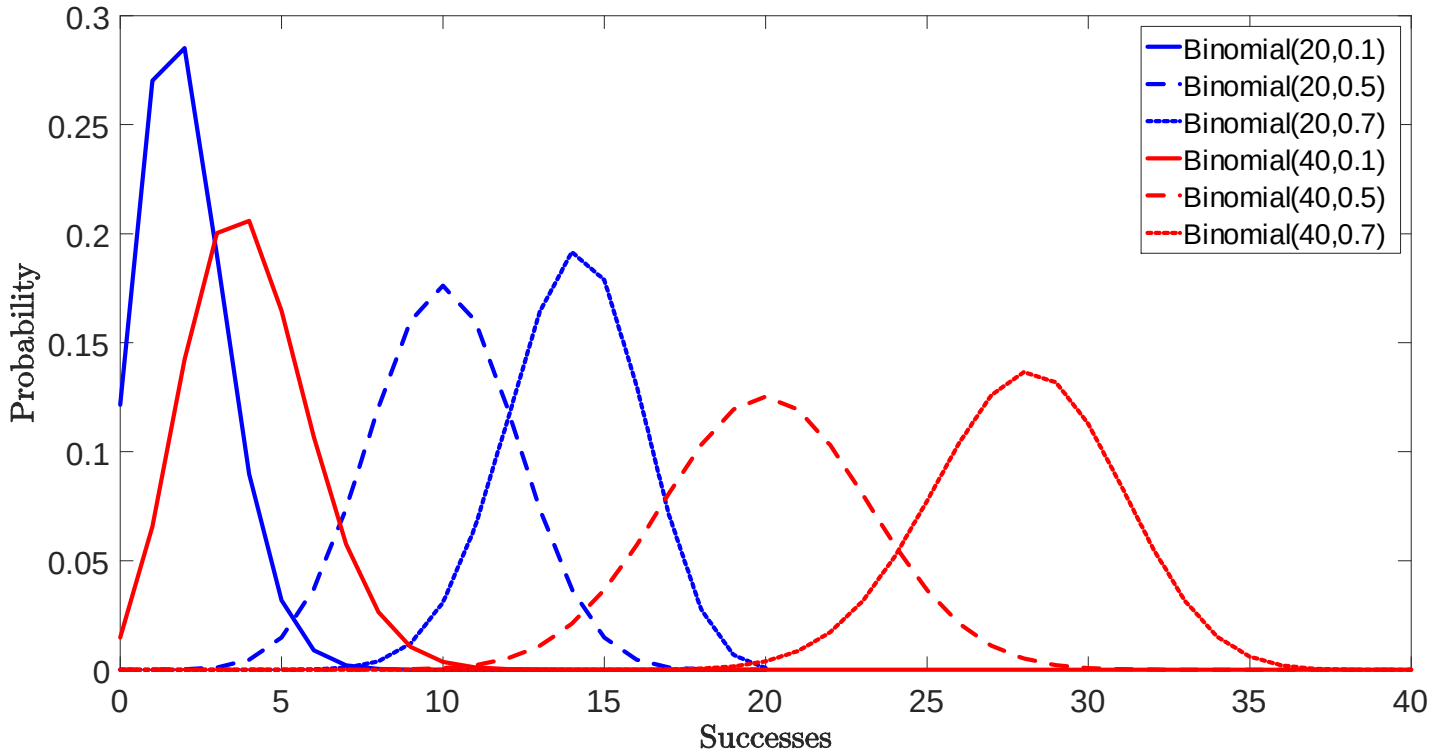


Figure: Binomial Distributions with Different n and p

Mean and Variance

- From the discussion on multivariate r.v.'s below, we can show

$$\mu_X = E[X] \overset{(*)}{=} \sum_{i=1}^{n} E[X_i] = np,$$

and

$$\sigma_X^2 = Var(X) \overset{(**)}{=} \sum_{i=1}^{n} Var(X_i) = np(1-p).$$

- Check the figure in the last slide for intuition.
- (*) holds even if X_i's are dependent, while (**) depends on the independence of X_i's; see the slides on jointly distributed r.v.'s.

- For Binomial$(5, 0.1)$,

$$\begin{aligned} \mu_X &= 5 \times 0.1 = 0.5, \\ \sigma_X^2 &= 5 \times 0.1 \times (1-0.1) = 0.45, \\ \sigma_X &= \sqrt{\sigma_X^2} = 0.6708. \end{aligned}$$

Poisson Distribution

Poisson Distribution

- The Poisson distribution was proposed by Siméon Poisson in 1837. [figure here]
- It is used to determine the probability of a random variable which characterizes the number of occurrences or successes of a certain event in a given continuous interval (such as time, surface area, or length).
- Assume that an interval is divided into a very large number of equal subintervals so that the probability of the occurrence of an event in any subinterval is very small (e.g., ≤ 0.05). The Poisson distribution models the number of events occuring on that inverval, assuming

1. The probability of the occurrence of an event is constant for all subintervals.
2. There can be no more than one occurrence in each subinterval.
3. Occurrences are independent, i.e., an occurrence in one interval does not influence the probability of an occurrence in another interval.

History of the Poisson Distribution

Siméon D. Poisson (1781-1840), French

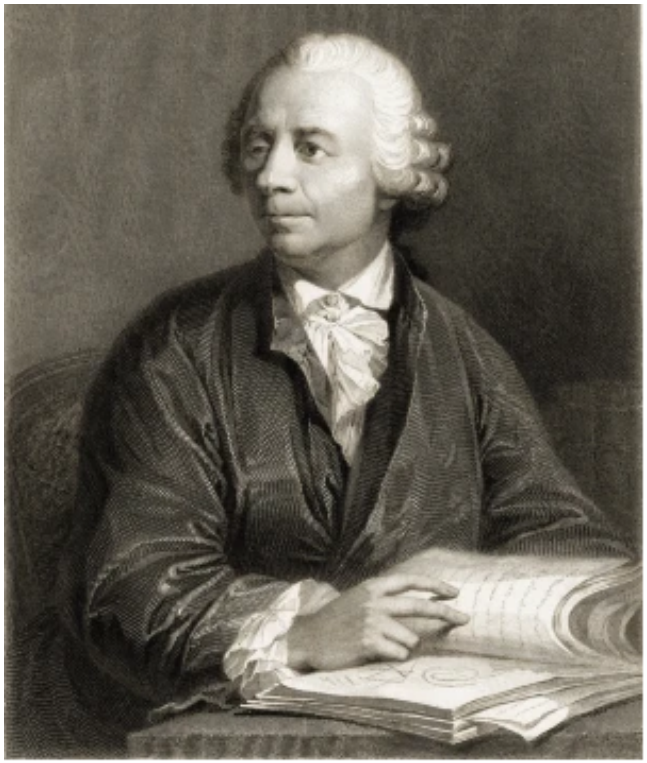
Applications

- From these assumptions, we can see the Poisson distribution can be used to model, e.g., the number of failures in a large computer system during a given day, the number of ships arriving at a dock during a 6-hour loading period, the number of defective products in large production runs, etc.
- The Poisson distribution is particularly useful in waiting line, or queuing, problems, e.g., the probability of various numbers of customers waiting for a phone line or waiting to check out of large retail store.
  - For a store manager, how to balance long lines (too few checkout lines, losing customers) and idle customer service associates (too many lines, resulting waste)?

PMF and Properties

- Intuitively, the Poisson r.v. is the binomial r.v. taking limit as $p \to 0$ and $n \to \infty$. If $np \to \lambda$ which specifies the average number of occurrences (successes) for a particular time (and/or space), then the binomial distribution converges to the Poisson distribution:
$$p(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}, x = 0, 1, 2, \cdots,$$
where $e = 2.71828\cdots$ is the base for natural logarithms, called Euler's number [figure here]. [proof not required]
- We denote a r.v. X with the above Poisson distribution as $X \sim \text{Poisson}(\lambda)$.
- For example, when $\lambda = 0.5$, $p(2|\lambda) = \frac{e^{-0.5}0.5^2}{2!} = 0.0758$, which can be easily obtained from Poisson Tables (Appendix Table 5).
- $\mu_X = E[X] = \lambda$, and $\sigma_X^2 = Var(X) = \lambda$.
- $np \to \lambda$, and $np(1-p) = np - np \cdot p \to \lambda - \lambda \cdot 0 = \lambda$.
- The sum of independent Poisson r.v.'s is also a Poisson r.v., e.g., the sum of K Poisson(λ) r.v. is a Poisson$(K\lambda)$ r.v..

History of Euler's Number

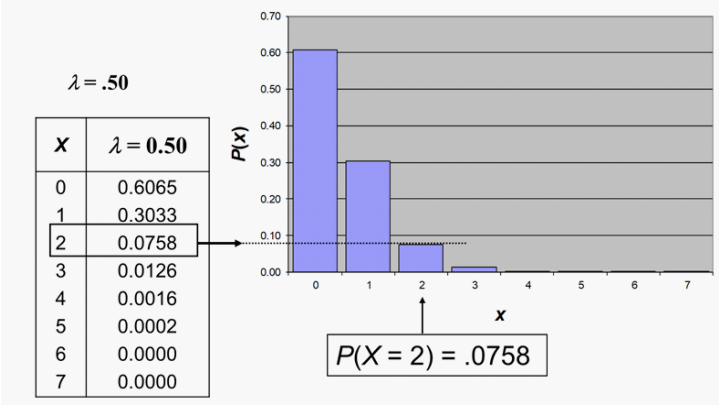


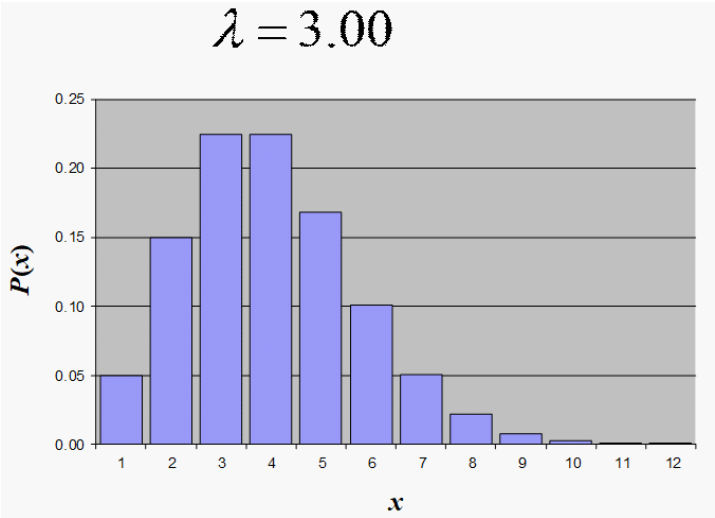
Leonhard Euler (1707-1783), Swiss

- Euler is widely considered to be the most prolific mathematician – 866 publications. He lost his vision in his latter life, which, however, did not affect his productivity: in 1775 he produced, on average, one mathematical paper every week.

Shape of Possion Distribution

- The shape of the Poisson Distribution depends on the parameter λ :





Poisson Approximation to the Binomial Distribution

- When n is large and np is of only moderate size (preferably $np \leq 7$), the binomial distribution can be approximated by Poisson(np).

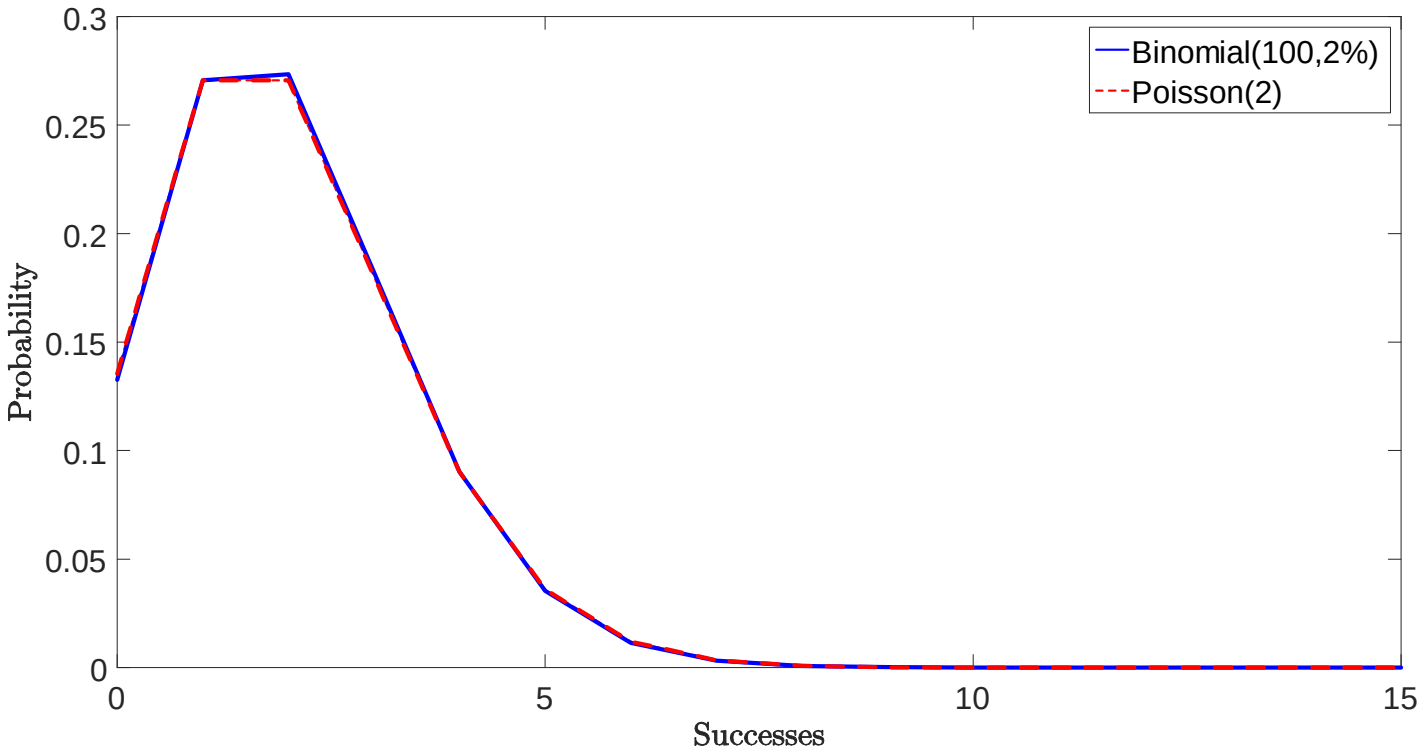


Figure: Poisson Approximation

- For an example where the approximation is not this good, see Assignment II.8(ii).

Hypergeometric Distribution

Hypergeometric Distribution

- If the binomial distribution can be treated as from random sampling with replacement from a population of size N, S of which are successes and $S/N = p$, then the hypergeometric distribution models the number of successes from random sampling without replacement.
  - These two random sampling schemes will be discussed more in Lecture 5.
- The hypergeometric distribution is

$$p(x|n,N,S) = \frac{C_x^S C_{n-x}^{N-S}}{C_n^N}, x = \max\{0, n-(N-S)\}, \cdots, \min(n,S),$$

  where n is the size of the random sample, and x is number of successes.
  - A r.v. with this distribution is denoted as $X \sim \text{Hypergeometric}(n,N,S)$.
- The binomial distribution assumes the items are drawn independently, with the probability of selecting an item being constant.
- This assumption can be met in practice if a small sample is drawn (without replacement) from a large population (e.g., $N > 10,000$ and $n/N < 1\%$). [figure here]
- When we draw from a small population, the probability of selecting an item and the probability of success are changing with each selection because the numbers of remaining items and remaining success items are changing, i.e., outcomes of trials are dependent.

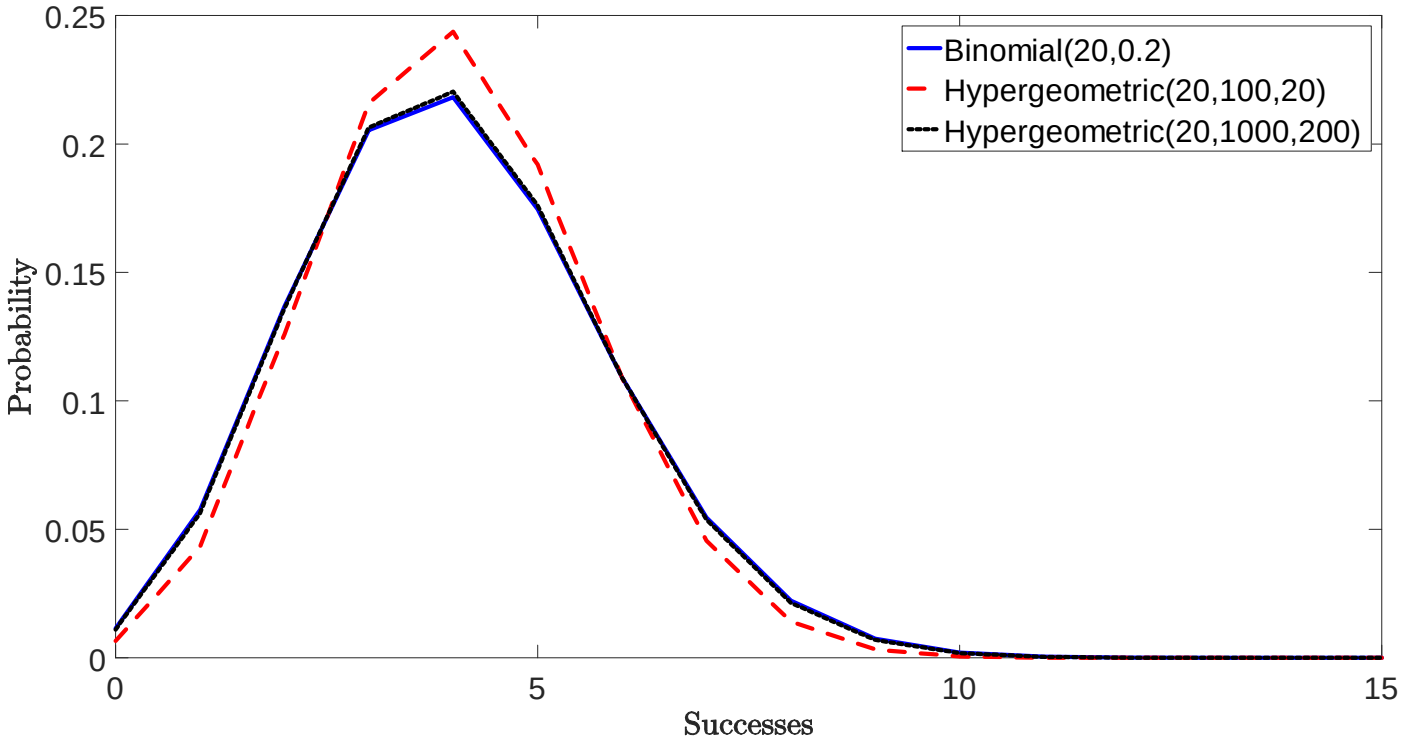


Figure: Comparison of Binomial and Hypergeometric Distributions

- $\mu_X = E[X] = np$, and $\sigma_X^2 = Var(X) = np(1-p)\frac{N-n}{N-1} \leq np(1-p)$,[3] where $p = \frac{S}{N}$. [proof not required]

[3] When $\frac{n}{N}$ is small, $\frac{N-n}{N-1}$ is close to 1, matching the variance of the binomial r.v.

[Example] Using the Hypergeometric Distribution

- 3 different computers are checked from 10 in the department. 4 of the 10 computers have illegal software loaded. What is the probability that 2 of the 3 selected computers have illegal software loaded?
- Solution: $N=10, S=4, n=3, x=2$.
- So

$$p(2)=\frac{C_x^S C_{n-x}^{N-S}}{C_n^N}=\frac{C_2^4 C_{3-2}^{10-4}}{C_3^{10}}=\frac{6\times 6}{120}=0.3,$$

  i.e., the probability that 2 of the 3 selected computers have illegal software loaded is 0.30, or 30%.

- Furthermore,

$$\mu_X=np=3\times\frac{4}{10}=1.2,$$

  and

$$\sigma_X^2=np(1-p)\frac{N-n}{N-1}=3\times\frac{4}{10}\times\left(1-\frac{4}{10}\right)\times\frac{10-3}{10-1}=0.56.$$

Jointly Distributed Discrete Random Variables

Bivariate Discrete R.V.'s: Joint and Marginal Probability Distributions

- We can use bivariate probability distribution to model the relationship between two univariate r.v.'s.
- For two discrete r.v.'s X and Y, their joint probability distribution expresses the probability that simultaneously X takes the specific value x and Y takes the value y, as a function of x and y:

$$p(x,y) = P(X = x \cap Y = y), x \in \mathscr{S}_X \text{ and } y \in \mathscr{S}_Y.$$

  - $p(x,y)$ is a straightforward extension of joint probabilities in Lecture 2, where $X = x$ and $Y = y$ are two events with x and y indexing them.
  - From probability postulates in Lecture 2, $0 \leq p(x,y) \leq 1$, and $\sum_{x \in \mathscr{S}_X} \sum_{y \in \mathscr{S}_Y} p(x,y) = 1$.
- The marginal probability distribution of X is

$$p(x) = \sum_{y \in \mathscr{S}_Y} p(x,y),$$

  and the marginal probability distribution of Y is

$$p(y) = \sum_{x \in \mathscr{S}_X} p(x,y),$$

Conditional Probability Distribution and Independence of Bivariate R.V.'s

- These two concepts are parallel to conditional probabilities and independent events in Lecture 2.
- The conditional probability distribution of Y, given that X takes the value x, expresses the probability that Y takes the value y, as a function of y, when the value x is fixed for X:

$$p(y|x) = \frac{p(x,y)}{p(x)};$$

similarly, the conditional probability distribution of X, given $Y = y$, is

$$p(x|y) = \frac{p(x,y)}{p(y)}.$$

- Two r.v.'s X and Y are independent iff

$$p(x,y) = p(x)p(y)$$

for all $x \in \mathscr{S}_X$ and $y \in \mathscr{S}_Y$, i.e., independence of r.v.'s can be understood as a set of independencies of events. E.g., "height" and "musical talent" are independent.
- Generally, k r.v.'s are independent iff $p(x_1,\cdots,x_k) = p(x_1)p(x_2)\cdots p(x_k)$ for all possible $(x_1,\cdots,x_k)$ values. [Question: how are $p(x_1,\cdots,x_k)$ and $p(x_j)$, $j = 1,\cdots,k$ defined?]
- X and Y are independent iff $p(y|x) = p(y)$ or $p(x|y) = p(x)$ for all possible (x,y) values ($\Longrightarrow$symmetric).

Conditional Mean and Variance

- The conditional mean of Y, given that X takes the value x, is given by

$$\mu_{Y|X=x} = E[Y|X=x] = \sum_{y \in \mathscr{S}_Y} y p(y|x),$$

where note that $p(y|x) \geq 0$ and $\sum_{y \in \mathscr{S}_Y} p(y|x) = 1$.
- E.g., if Y is wage and X is education, then $\mu_{Y|X=12}$ is the average wage for the "slice" of population who are high school graduates.
- For any constants a and b, $E[a+bY|X=x] = a+bE[Y|X=x]$.

- The conditional variance of Y, given that X takes the value x, is given by

$$\sigma^2_{Y|X=x} = Var(Y|X=x) = \sum_{y \in \mathscr{S}_Y} \left(y - \mu_{Y|X=x}\right)^2 p(y|x).$$

- For any constants a and b, $Var(a+bY|X=x) = b^2 Var(Y|X=x)$.

- Notation: The notations used in the textbook, $\mu_{Y|X}$ and $\sigma^2_{Y|X}$, are not clear.

Mean and Variance of (Linear) Functions

- For a function of (X,Y), $g(X,Y)$, its mean, $E[g(X,Y)]$, is defined as

$$E[g(X,Y)] = \sum_{x\in\mathscr{S}_X}\sum_{y\in\mathscr{S}_Y} g(x,y)p(x,y).$$

- For a linear function of (X,Y), $W = aX + bY$,

$$\mu_W := E[W] = a\mu_X + b\mu_Y, \text{[verified in the next slide]}$$
$$\sigma_W^2 := Var(W) = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY} \text{ [see the next}^4\text{ slide for } \sigma_{XY}\text{]}.$$

- e.g., W is the total revenue of two products with (X,Y) being the sales and (a,b) the prices.
- If $a = b = 1$, then $E[X+Y] = E[X] + E[Y]$, i.e., the mean of sum is the sum of means.
- If $a = 1$ and $b = -1$, then $E[X-Y] = E[X] - E[Y]$, i.e., the mean of difference is the difference of means.
- If $a = b = 1$ and $\sigma_{XY} = 0$, then $Var(X+Y) = Var(X) + Var(Y)$, i.e., the variance of sum is the sum of variances.
- If $a = 1, b = -1$ and $\sigma_{XY} = 0$, then $Var(X-Y) = Var(X) + Var(Y)$, i.e., the variance of difference is the sum of variances.

(*) Verification and Extensions

- μ_W:

$$\begin{aligned}\mu_W &= \sum_{x\in\mathcal{S}_X}\sum_{y\in\mathcal{S}_Y}(ax+by)p(x,y)\\ &= a\sum_{x\in\mathcal{S}_X}\left[x\sum_{y\in\mathcal{S}_Y}p(x,y)\right]+b\sum_{y\in\mathcal{S}_Y}\left[y\sum_{x\in\mathcal{S}_X}p(x,y)\right]\\ &= a\sum_{x\in\mathcal{S}_X}xp(x)+b\sum_{y\in\mathcal{S}_Y}yp(y)\\ &= a\mu_X+b\mu_Y,\end{aligned}$$

- σ_W^2 can be derived based on this result. [exercise]
- Extension I: If $W=\sum_{i=1}^K a_iX_i$, then

$$\mu_W=E[W]=\sum_{i=1}^K a_iE[X_i]=:\sum_{i=1}^K a_i\mu_i,$$

continue

- and

$$\sigma_W^2 = Var(W) = \sum_{i=1}^K a_i^2 Var(X_i) + 2\sum_{i=1}^{K-1}\sum_{j>i} a_i a_j Cov(X_i, X_j)$$
$$=: \sum_{i=1}^K a_i^2\sigma_i^2 + 2\sum_{i=1}^{K-1}\sum_{j>i} a_i a_j \sigma_{ij},$$

where note that $2\sum_{i=1}^{K-1}\sum_{j>i} = 2\sum_{i=1}^{K-1}\sum_{j=i+1}^K = \sum_{i=1}^K\sum_{j\neq i}$. [figure here]
- If $a_i = 1$ for all i, then we have

$$E\left[\sum_{i=1}^K X_i\right] = \sum_{i=1}^K \mu_i \text{ and } Var\left(\sum_{i=1}^K X_i\right) = \sum_{i=1}^K \sigma_i^2 + 2\sum_{i=1}^{K-1}\sum_{j>i}\sigma_{ij},$$

where $Var\left(\sum_{i=1}^K X_i\right)$ reduces to $\sum_{i=1}^K \sigma_i^2$ if $\sigma_{ij} = 0$ for all $i \neq j$.

- Extension II: For $W = aX + bY$ and a r.v. Z different from (X, Y),

$$\begin{aligned} E[W|Z=z] &= a\mu_{X|Z=z} + b\mu_{Y|Z=z}, \\ Var(W|Z=z) &= a^2\sigma^2_{X|Z=z} + b^2\sigma^2_{Y|Z=z} + 2ab\sigma_{XY|Z=z}, \end{aligned}$$

where $Var(W|Z=z)$ reduces to $a^2\sigma^2_{X|Z=z} + b^2\sigma^2_{Y|Z=z}$ if $\sigma_{XY|Z=z} = 0$.

(*) Details of σ_W^2

- Note that

$$\begin{aligned} Var(W) &= Cov(W,W) = Cov\left(\sum_{i=1}^K a_iX_i, \sum_{j=1}^K a_jX_j\right) = \sum_{i=1}^K \sum_{j=1}^K Cov\left(a_iX_i, a_jX_j\right) \\ &= \sum_{i=1}^K \sum_{j=i} Cov\left(a_iX_i, a_jX_j\right) + \sum_{i=1}^K \sum_{j\neq i} Cov\left(a_iX_i, a_jX_j\right) \\ &= \sum_{i=1}^K Var\left(a_iX_i\right) + 2\sum_{i=1}^{K-1} \sum_{j>i} Cov\left(a_iX_i, a_jX_j\right), \end{aligned}$$

where $Cov\left(a_iX_i, a_jX_j\right) = Cov\left(a_jX_j, a_iX_i\right)$.

	$j=1$	$j=2$	$j=3$	$\cdots$	$j=K$
$i=1$	$Cov(a_1X_1, a_1X_1)$	$Cov(a_1X_1, a_2X_2)$	$Cov(a_1X_1, a_3X_3)$	$\cdots$	$Cov(a_1X_1, a_KX_K)$
$i=2$	$Cov(a_2X_2, a_1X_1)$	$Cov(a_2X_2, a_2X_2)$	$Cov(a_2X_2, a_3X_3)$	$\cdots$	$Cov(a_2X_2, a_KX_K)$
$i=3$	$Cov(a_3X_3, a_1X_1)$	$Cov(a_3X_3, a_2X_2)$	$Cov(a_3X_3, a_3X_3)$	$\cdots$	$Cov(a_3X_3, a_KX_K)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$i=K$	$Cov(a_KX_K, a_1X_1)$	$Cov(a_KX_K, a_2X_2)$	$Cov(a_KX_K, a_3X_3)$	$\cdots$	$Cov(a_KX_K, a_KX_K)$

Covariance and Correlation

- These two concepts are the same as those in Lecture 1 but in the probability language.
- The covariance between X and Y

$$Cov(X,Y)=\sigma_{XY}=E[(X-\mu_X)(Y-\mu_Y)]=\sum_{x\in\mathscr{S}_X}\sum_{y\in\mathscr{S}_Y}(x-\mu_X)(y-\mu_Y)p(x,y).$$

  - It is not hard to show that $Cov(X,Y)=E[XY]-\mu_X\mu_Y$, which reduces to $Var(X)=E\left[X^2\right]-\mu_X^2$ when $X=Y$.
- The correlation between X and Y

$$Corr(X,Y)=\rho_{XY}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y}.$$

- Recall that σ_{XY} is not unit-free so is unbounded, while $\rho_{XY}\in[-1,1]$ is more useful.
- Recall that σ_{XY} and ρ_{XY} have the same sign: if they are positive, X and Y are called positively dependent; when they are negative, X and Y are called negatively dependent; when they are zero, there is no linear relationship between X and Y.

Covariance and Independence

- If X and Y are independent, then $Cov(X,Y) = Corr(X,Y) = 0$. [exercise]
  - The converse is not true; recall the figure in Lecture 1.
- Here is a concrete example: if the distribution of X is

$$p(-1) = 1/4,\ p(0) = 1/2 \text{ and } p(1) = 1/4,$$

  then $Cov(X,Y) = 0$ with $Y = X^2$. Why?
- Because X can determine Y, X and Y are not independent (rigorously, check $p(x,y) \neq p(x)p(y)$ for some (x,y) below).
- The distribution of X implies $E[X] = 0$.
- The distribution of Y is $p(0) = p(1) = 1/2$, i.e., Y is a Bernoulli r.v., which implies $E[Y] = 1/2$.
- The joint distribution of (X,Y) is

$$p(-1,1) = 1/4,\ p(0,0) = 1/2,\ p(1,1) = 1/4,$$

  which implies $E[XY] = 0$, so $Cov(X,Y) = E[XY] - E[X]E[Y] = 0$.
- Portfolio analysis in the textbook (Pages 190-192, 236-240) will be discussed in the next tutorial class.