Probability and Mathematical Statistics in Data Science

Lecture 07: Section 3.3: Binomial Distribution
Section 3.4: The Hypergeometric Distribution

The Binomial Model

- A **Binomial model** tells us the probability for a random variable that counts the number of successes in a fixed number of Bernoulli trials.

- Two parameters define the Binomial model: *n*, the number of trials; and, *p*, the probability of success. We denote this Binom(*n*, *p*).

The Binomial Distribution

There are many experiments that conform either exactly or approximately to the following list of requirements:

1. The experiment consists of a sequence of n smaller experiments called *trials,* where n is fixed in advance of the experiment.

2. Each trial can result in one of the same two possible outcomes (dichotomous trials), which we generically denote by success (S) and failure (F).

3. The trials are **independent**, so that the outcome on any particular trial does not influence the outcome on any other trial.

4. The probability of success $P(S)$ is constant from trial to trial; we denote this probability by p.

Independence and Sampling Without Replacement

Population: 50 people, 40 Insured - P(Insured) = 0.80

Select 10 people from pool at random w/o replacement

P(10th person insured | 1st 9 insured) = 31/41 = 0.75

Population: 500,000 people, 400,000 Insured - P(Insured) = 0.80

Select 10 people from pool at random w/o replacement

P(10th person insured | 1st 9 insured) = 399,991/499,991

= 0.80 (approximately)

Independence

- One of the important requirements for Bernoulli trials is that the trials be independent.

- When we don't have an infinite population, the trials are not independent. But, there is a rule that allows us to pretend we have independent trials:

  - **The 10% condition**: Bernoulli trials must be independent. If that assumption is violated, it is still okay to proceed as long as the sample is smaller than 10% of the population.

Binomial Distribution: Example

- Consider a box with one red ball and eleven blue ones.

- One draw is made. What is the probability that the ball is red?
  - n=1, p=1/12

- Now 4 draws are made, ***with replacement***. What is the probability that *exactly* 1 draw is red (out of the 4)?
  - Notice that this is like a tossing a coin 4 times, with P(head) = 1/12.

- P(RBBB) =
- How many such sequences are there?
- What is the probability of all such sequences (1 R, 3B)?

Binomial Example: Tossing a Coin Four Times

HTHT **TTHH** **HHHT** **HHHH**

They are 16 possible outcomes

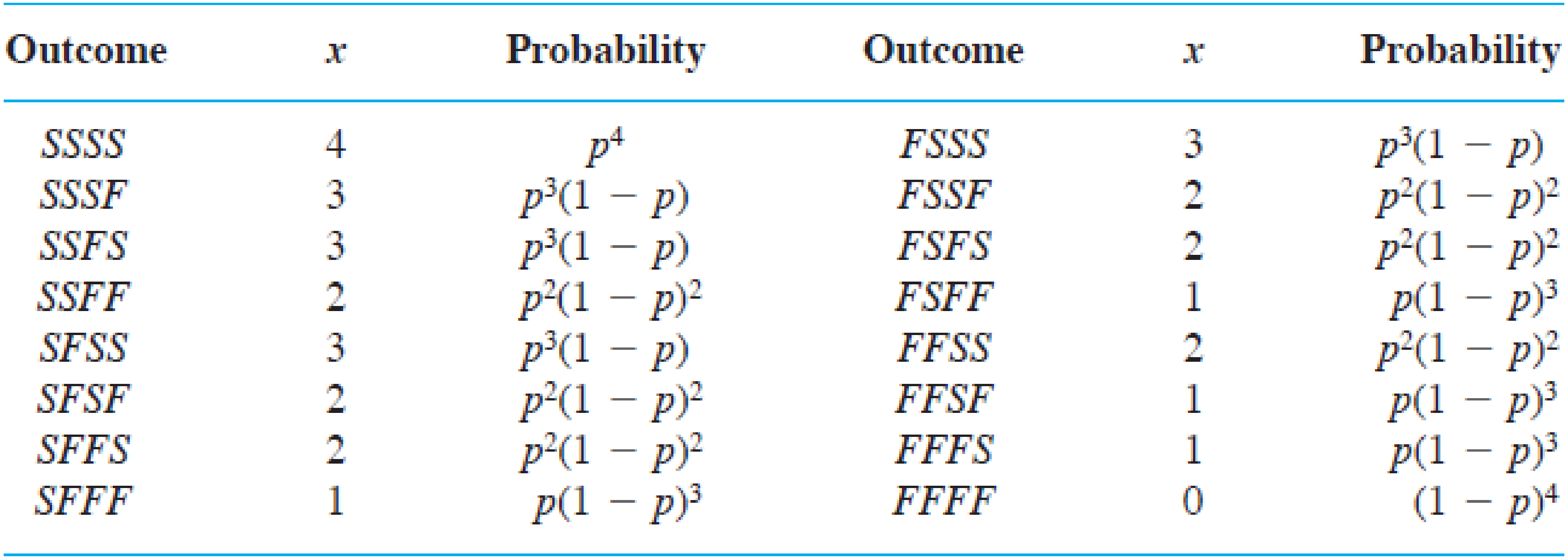
HHHH
HHTH
HTHH
HTTH
THHH
THTH
TTHH
TTTH
HHHT
HHTT
HTHT
HTTT
THHT
THTT
TTHT
TTTT

The probability of getting all heads is 1/16 or (0.5) (0.5) (0.5) (0.5) equal to 0.0625. The probability of getting 50% heads and 50% tails is 6/16 (0.375).

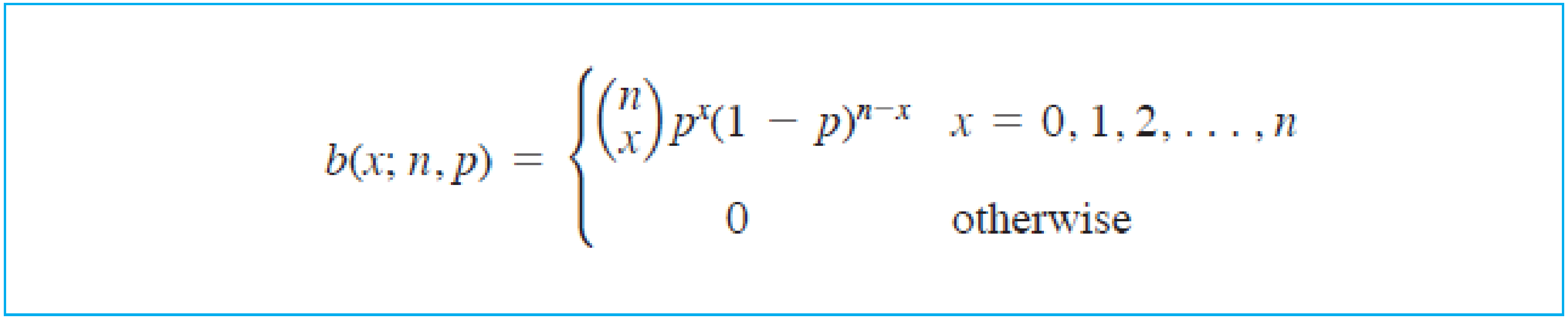
Probability Distribution for the number of heads

No. of Heads	0	1	2	3	4
Proportion:	0.0625	0.25	0.375	.25	0.0625

Outcomes and Probabilities for Binomial Experiment with 4 Trials



Outcome	x	Probability	Outcome	x	Probability
SSSS	4	p^4	*FSSS*	3	$p^3(1-p)$
SSSF	3	$p^3(1-p)$	*FSSF*	2	$p^2(1-p)^2$
SSFS	3	$p^3(1-p)$	*FSFS*	2	$p^2(1-p)^2$
SSFF	2	$p^2(1-p)^2$	*FSFF*	1	$p(1-p)^3$
SFSS	3	$p^3(1-p)$	*FFSS*	2	$p^2(1-p)^2$
SFSF	2	$p^2(1-p)^2$	*FFSF*	1	$p(1-p)^3$
SFFS	2	$p^2(1-p)^2$	*FFFS*	1	$p(1-p)^3$
SFFF	1	$p(1-p)^3$	*FFFF*	0	$(1-p)^4$



$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Devore 7th edition

Binomial Probability Mass Function

- The pmf of a binomial random variable is

$$b(x;n,p)=\begin{cases}\boxed{\binom{n}{x}}\boxed{p^x(1-p)^{n-x}} & x=0,1,2,3,\ldots,n\\ 0 & \text{otherwise}\end{cases}$$

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

n = number of trials

p = probability of success

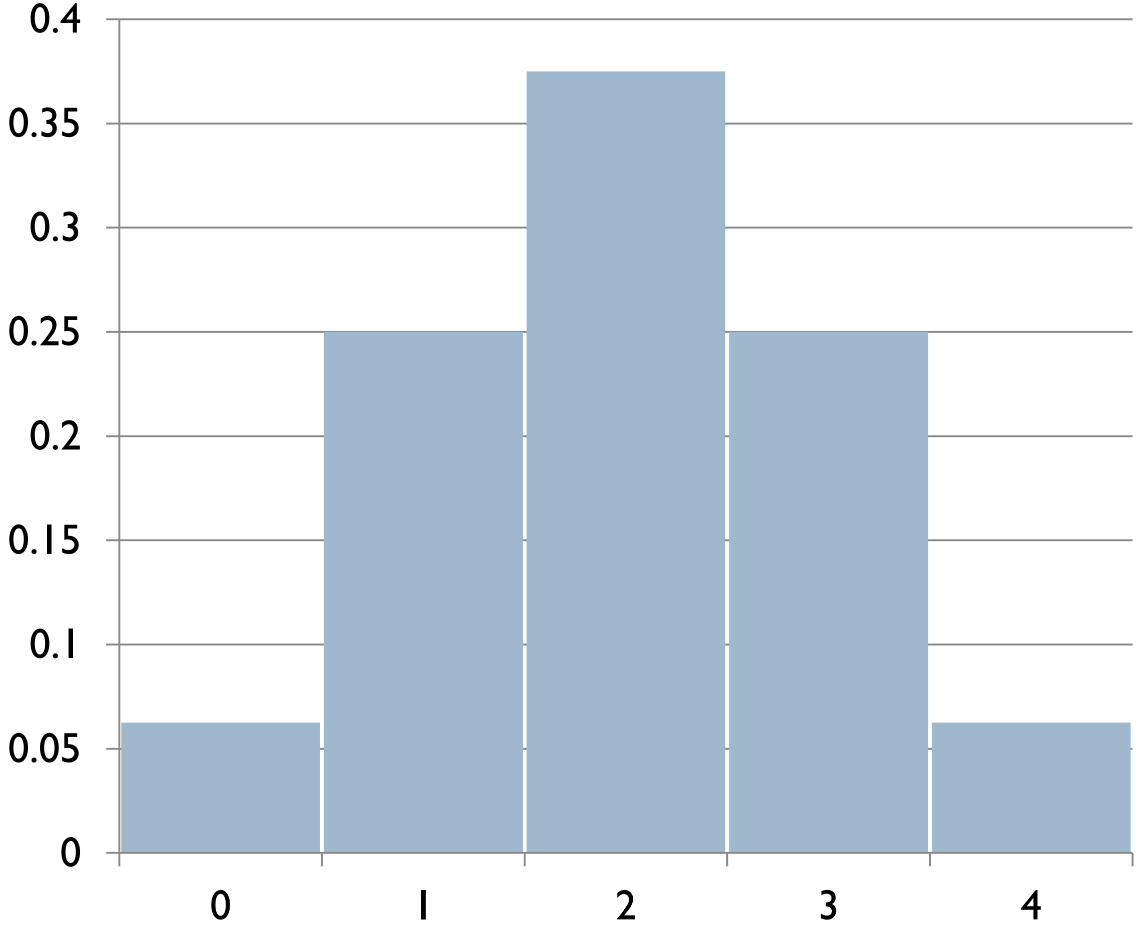
q = 1 – *p* = probability of failure

x = # of successes in *n* trials

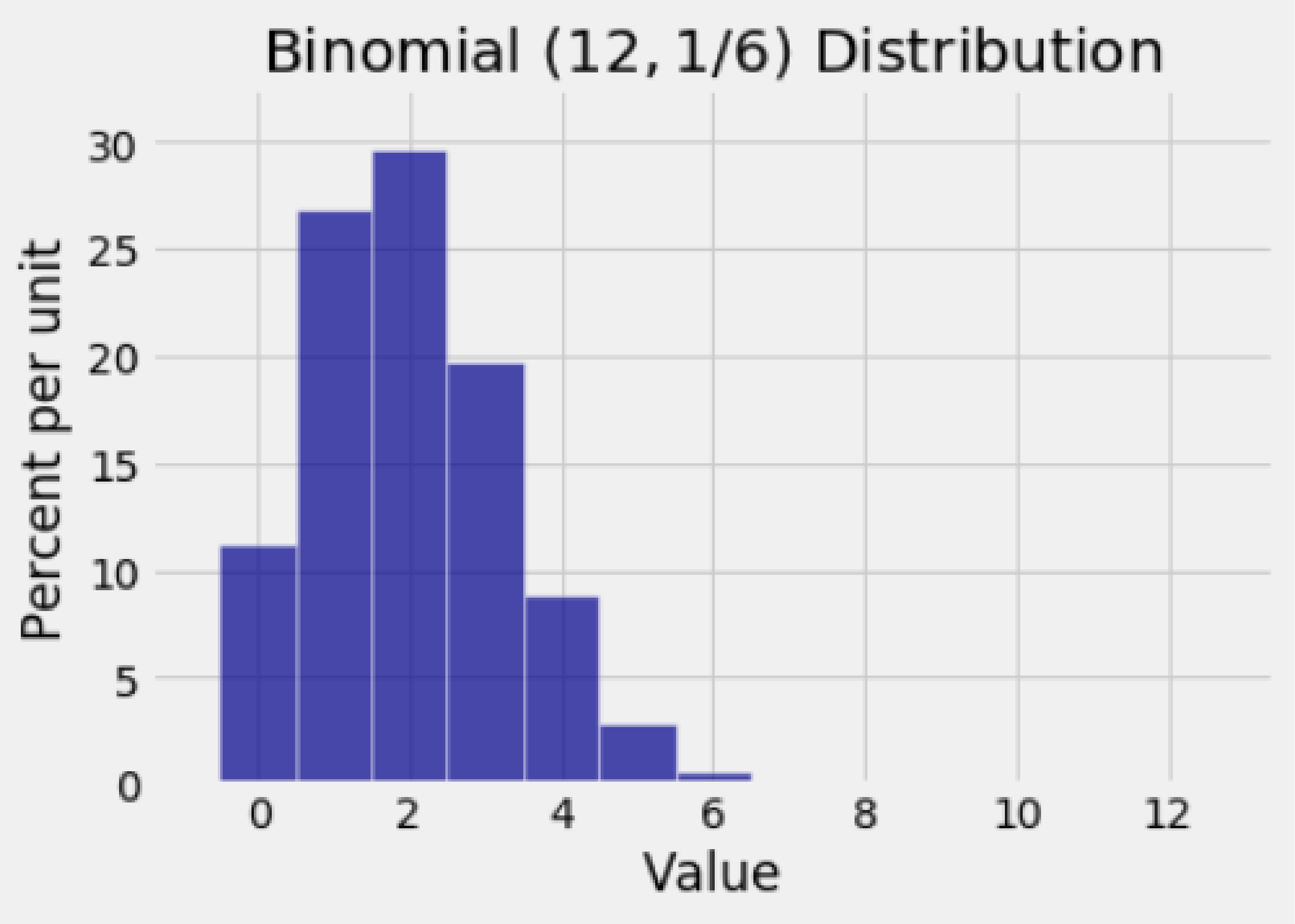
Note: *n!* = *n* × (*n* – 1) × … × 2 × 1, and *n*! is read as "*n*factorial."

Binomial Example

Distribution of Number of Heads from Repeatedly Tossing a Coin Four Times



Number of Sixes in 12 rolls of a die



Identifying binomial random variables

Which of the following are binomial random variables?

- Number of heads in 12 tosses of a fair coin.
- Number of tosses until we see two heads.
- Number of queens in a five card hand
- Number of Democrats in a simple random sample of 500 adult voters drawn from the SF Bay Area.

Exercise 3.6.3

- Yi likes to bet on "red" at roulette. Each time she bets, her chance of winning is 18/38, independently of all other times. Suppose she bets repeatedly on red. Find the chance that:

a) she wins four of the first 10 bets

b) she wins at most four of the first 10 bets

c) the third time she wins is on the 10th bet

Hypergeometric Distribution

- Suppose there are 50 colored socks in the drawer, of which 16 are red and the other 34 are blue. We are going to randomly draw 10 sock out of the drawer without replacement. What is the probability that we will have exactly 2 blue socks?

- In this example, when we have a *finite* or *small* population, and we sample without replacement. Therefore the binomial will not be appropriate.

- Notice that any subset of 10 socks in this example is equally likely to be chosen.

- Let X = the number of successes (blue socks) in the sample we draw, then X is said to have the hypergeometric distribution.

Parameters

- We see that the probability distribution of X depends on three parameters:
  n = sample size (10 in socks example).
  G = total number of successes in the population (34 in socks example).
  N = total number of individuals in the population (50 in socks example). We wish to obtain P(X=*g*) = *h*(*x*; *n*, *G*, *N*).

- P(X=2) = *h*(2; 10, 34, 50) = {# of outcomes with X=2} / {# of possible outcomes}.

- Thus we have

# of ways of selecting 2 blue socks

# of ways of selecting 8 red socks

$$h(2; 10, 34, 50) = \frac{\boxed{\binom{34}{2}}\boxed{\binom{16}{8}}}{\boxed{\binom{50}{10}}}$$

The Hypergeometric Distribution

- The assumptions leading to the hypergeometric distribution are as follows:

- **1.** The population or set to be sampled consists of *N* individuals, objects, or elements (a *finite* population).

- **2.** Each individual can be characterized as a success (*S*) or a failure (*F*), and there are *G* successes in the population.

- **3.** A sample of *n* individuals is selected without replacement in such a way that each subset of size *n* is equally likely to be chosen.

Hypergeometric Probability Mass Function

- If X is the number of successes in a completely random sample of size *n* drawn from a population consisting of *G* successes and (*N* – *G*) failures, then the distribution of X is given by

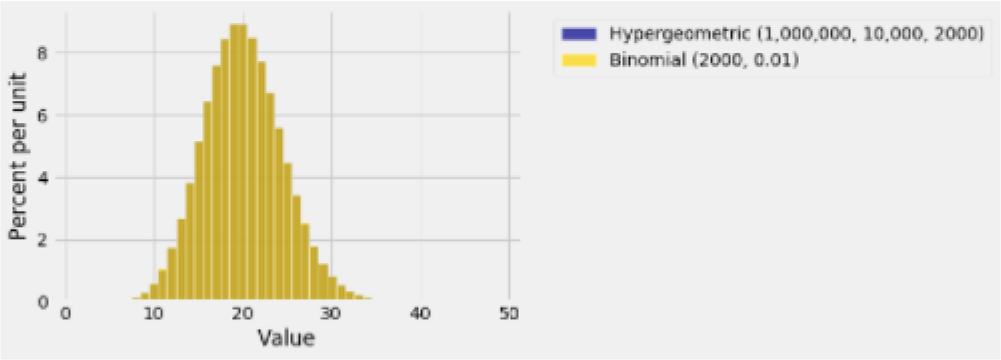
$$P(X = g) = \frac{\binom{G}{g}\binom{N-G}{n-g}}{\binom{N}{n}}$$

Binomial Approximation

- Let X be the number of Independent voters in a simple random sample of 2000 voters drawn from a population of one million voters of whom 1% are Independent.

- We know that X has the hypergeometric distribution with parameters N = 1,000,000 G = 10,000 and n= 2000

- If we sample with replacement then X – Binomial (2000, 0.01)


Percent per unit
Value
Hypergeometric (1,000,000, 10,000, 2000)
Binomial (2000, 0.01)