# Elementary Statistics Lecture 4

## Probability and Probability Distribution

Chong Ma

Department of Statistics
University of South Carolina

# Outline

# Probability Quantifies Randomness

Probability is the way we quantify uncertainty, i.e., measure the chances of the possible outcomes for random phenomena.

- Randomness is an essential component in statistics.
- Randomization is common in our life, such as rolling dice, spinning a wheel and flipping a coin etc.
- With a large number of observations for a random phenomenon, summary statistics settle down and get closer to particular numbers.

## Probability

With any random phenomenon, the **probability** of a particular outcome is the proportion of times that the outcome occurs in a long run of observations.

# Simulation Of A Fair Die

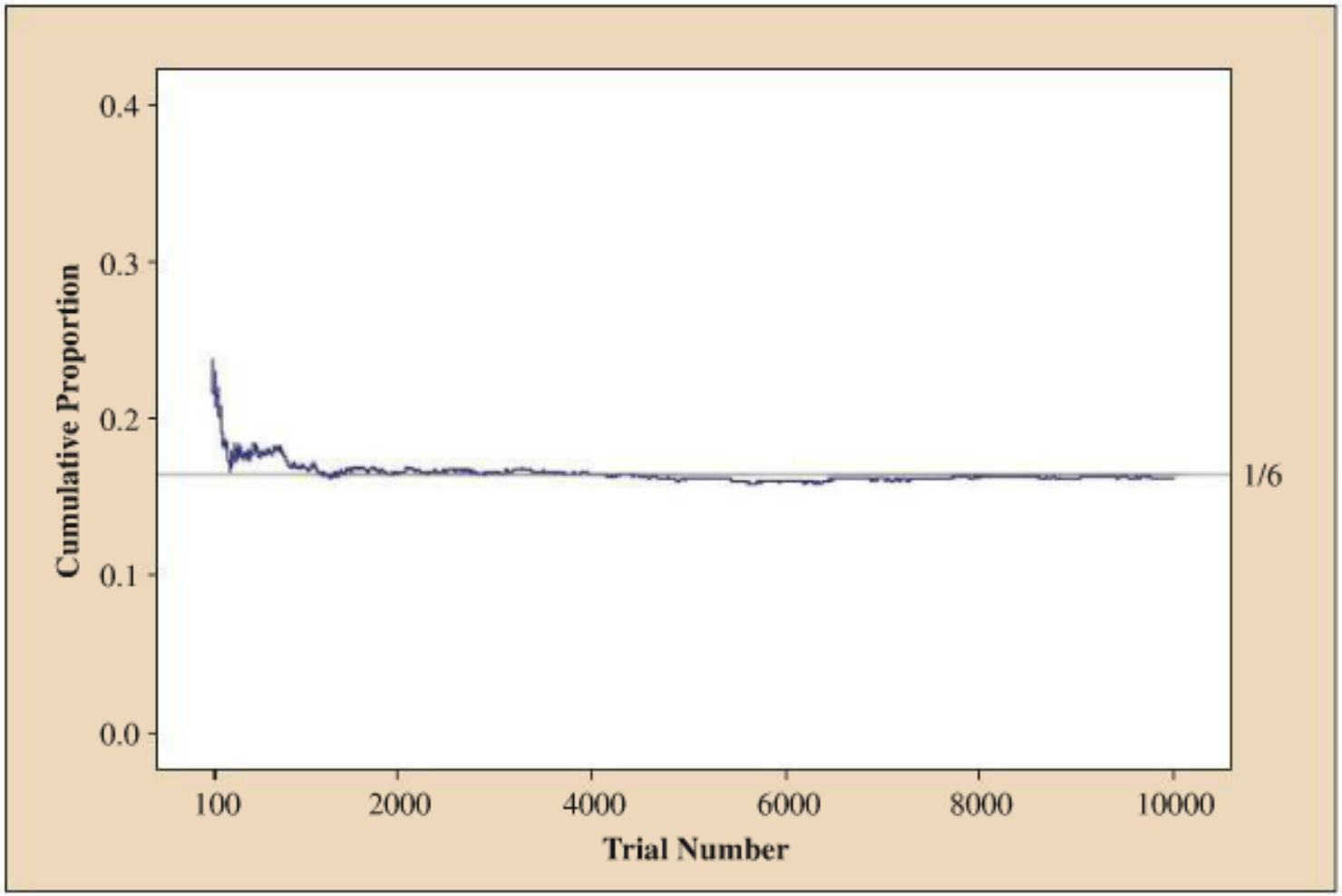



Figure 1: The cumulative proportion of times that a 6 occurs for a simulation of 10,000 rolls of a fair die. As the trial numbers go to large, the proportion of times that a 6 occurs converges to $\frac{1}{6}$.

# Outline

# Probability

**Sample Space** The set of all possible outcomes.

**Event** A subset of the sample space. An event corresponds to a particular outcome or a group of possible outcomes.

Examples

- roll a die once
  - **sample space**: $\Omega=\{1,2,3,4,5,6\}$
  - **even numbers**: $A=\{2,4,6\}$
- flip a coin twice
  - **sample space**: $\Omega=\{HH, HT, TH, TT\}$
  - **at least one head**: $A=\{HH, HT, TH\}$

# Probability Rules

Consider a pop quiz with three multiple-choice questions. Each question has five options, and a student's answer is either correct(C) or incorrect(I). What is the sample space for the possible answers on this pop quiz?

# Probability Rules

Consider a pop quiz with three multiple-choice questions. Each question has five options, and a student's answer is either correct(C) or incorrect(I). What is the sample space for the possible answers on this pop quiz?

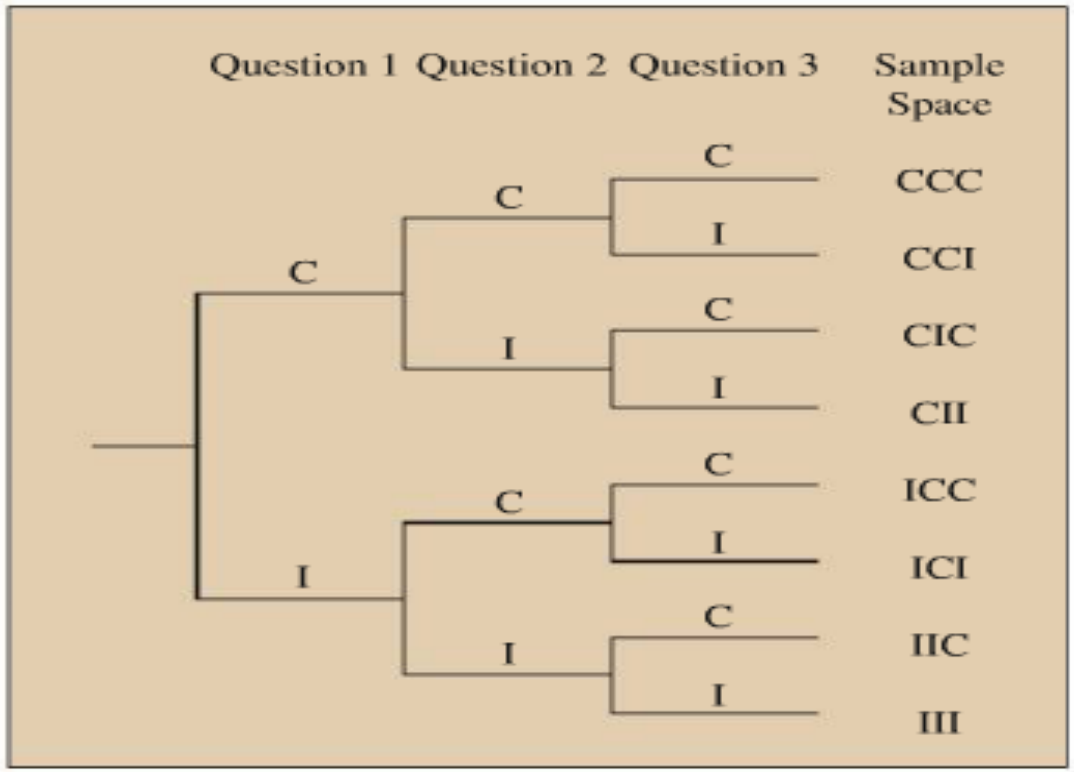



Figure 2: Tree diagram for student performance on a three-question pop quiz. The sample space $\Omega$={CCC, CCI, CIC, CII, ICC, ICI, IIC, III}

# Continue . . .

## Probability of an Event

When all the possible outcomes are equally likely,

$$P(\text{A}) = \frac{\text{number of outcomes in event A}}{\text{number of outcomes in the sample space}}$$

**Complement of A** Consists of all outcomes that are not in A.

$$P(A^c) = 1 - P(A)$$

**Union of A and B** Consists of outcomes that are in A or B or both.

$$P(\text{A or B}) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

**Intersection of A and B** Consists of outcomes that are in both A and B. When the two events are independent,

$$P(A \cap B) = P(A) \times P(B)$$

# Complement of an Event

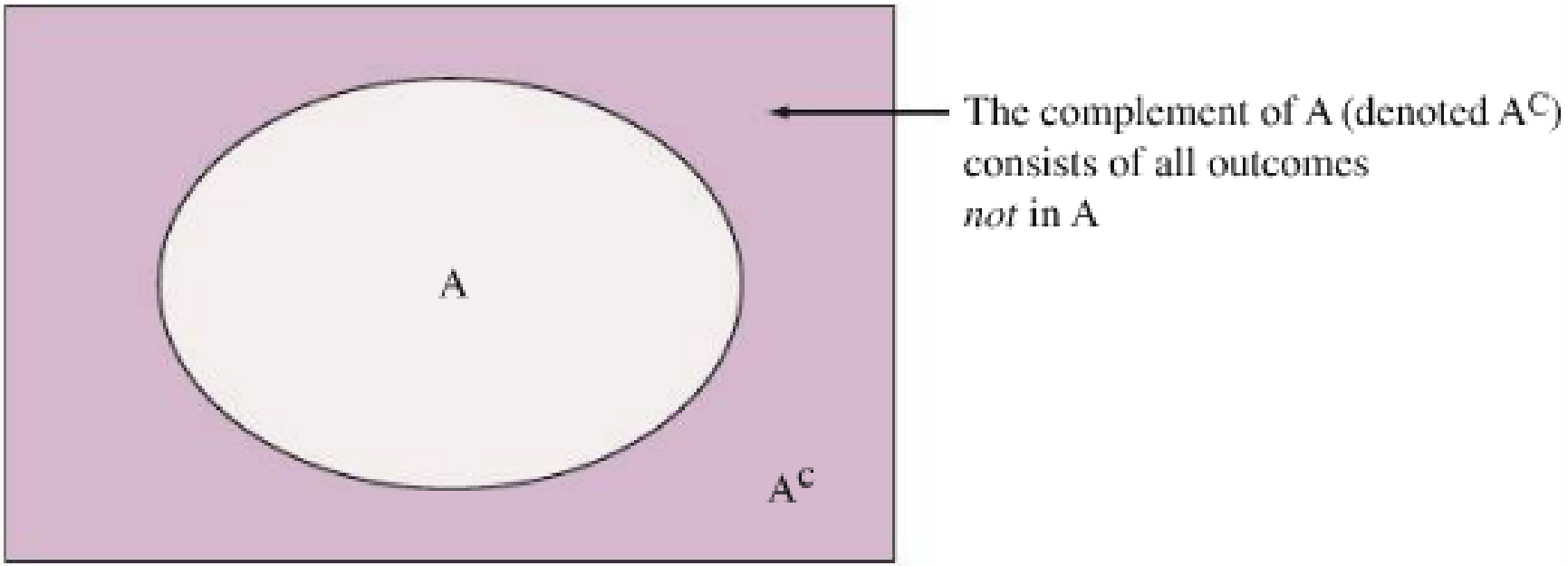



Figure 3: Venn diagram illustrating an event A and its complement $A^c$. Note that the probability of the whole sample space is 1, i.e., $P(\Omega) = 1$, then we have $P(A^c) = 1 - P(A)$

# Intersection and Union

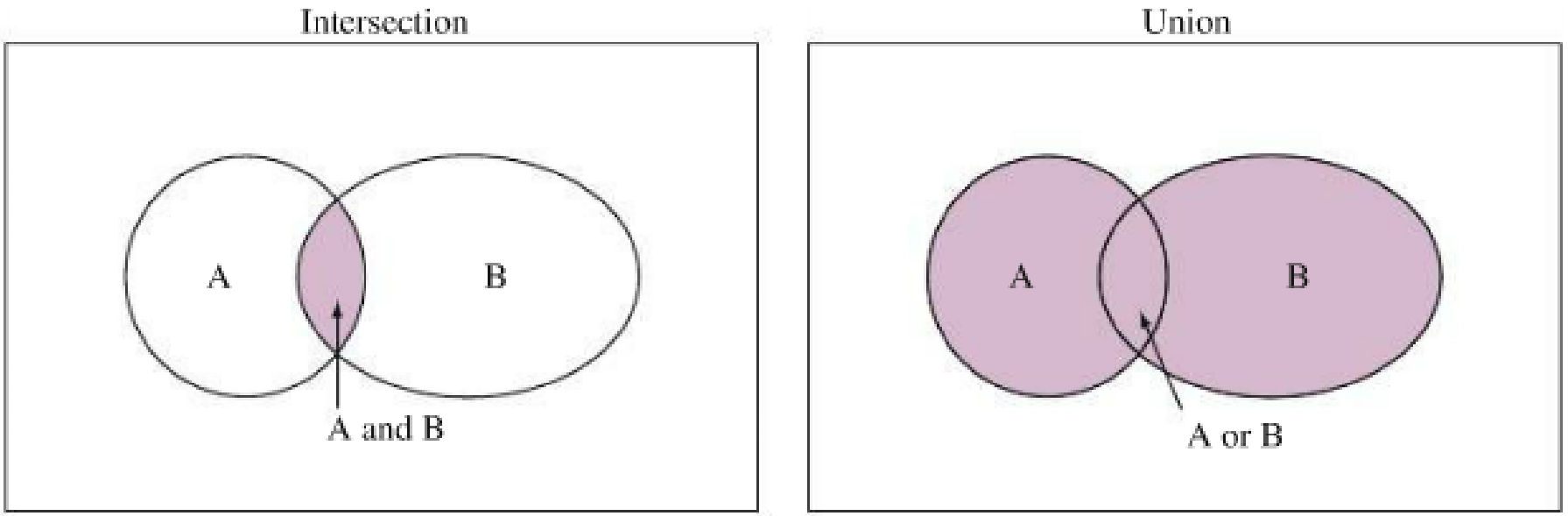



Figure 4: The left plot illustrates the intersection of events A and B; the right shows the union of events A and B. In particular, when events A and B are independent, $P(\text{A} \cap \text{B}) = P(\text{A}) \times P(\text{B})$.

# Independent versus Disjoint

**Independent Trials** Different trails of a random phenomenon are independent if the outcome of any one trial is not affected by the outcome of any other trial.

**Disjoint Events** Two events A and B are disjoint if they do not have any common outcomes.

**Example**

A = the student answers exactly one question correctly = {CII, ICI, IIC}
B = the student answers exactly two question correctly = {CCI, ICC, CIC}
Are events A and B disjoint? Yes!

A = Emily answers exactly one question correctly = {CII, ICI, IIC}
B = Kate answers exactly two question correctly = {CCI, ICC, CIC}
Are events A and B disjoint? Yes!
Are events A and B independent? It depends!
**In conclusion, independence has nothing to do with disjoint!**

# Additional Rule

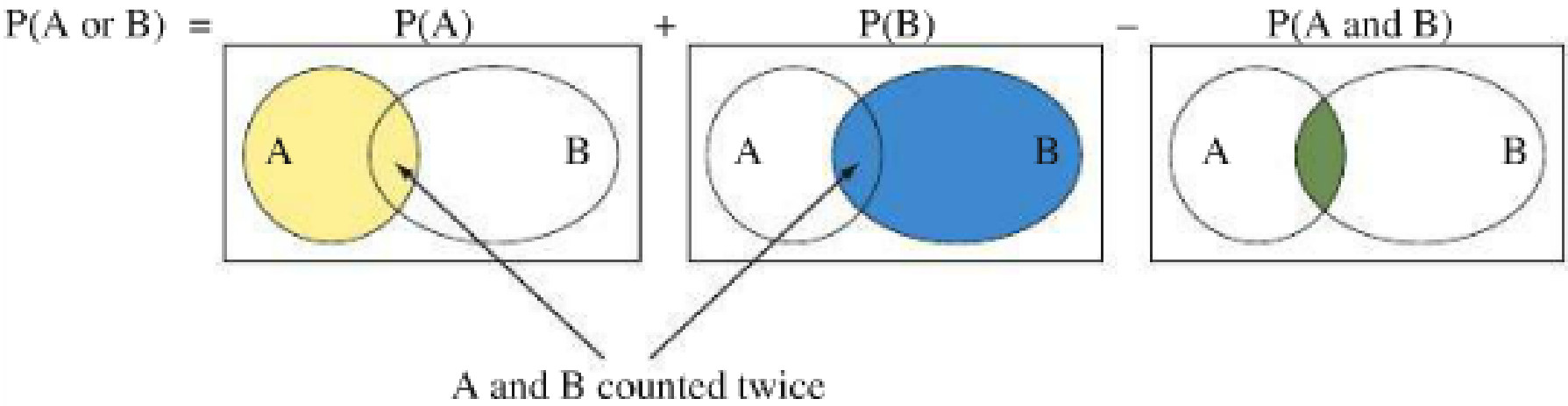



Figure 5: Sets algebra explains the union of two events. When events A and B are disjoint, i.e., $P(\mathrm{A} \cap \mathrm{B}) = 0$, then $P(\mathrm{A} \cup \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B})$.

A=student answers the first question correctly={CCC, CCI, CIC, CII}
B=student answers at least two questions correctly={CCI, CIC, ICC, CCC}
The probability of students answering at least two questions including the first one is

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{4}{8} + \frac{4}{8} - \frac{2}{8} = \frac{3}{4}$$

# Multiplication Rule

Consider the three-question multiple-choice pop quiz, a student is totally unprepared and randomly guesses the answer to each question. If each question has five options, then the probability of selecting the correct answer for any given question is 1/5 or 0.2. With guessing, the response on one question is not influenced by the response on another question. Thus, whether one question is answered correctly is independent of whether another question is answered correctly.

Note that $P(C) = 0.2$ and $P(I) = 0.8$. The probability that the student answers all three questions correctly is

$$P(CCC) = P(C) \times P(C) \times P(C) = 0.2 \times 0.2 \times 0.2 = 0.008$$

The probability of answering the first two questions correctly and the third question incorrectly is

$$P(CCI) = P(C) \times P(C) \times P(I) = 0.2 \times 0.2 \times 0.8 = 0.032$$

# Multiplication Rule

Consider the three-question multiple-choice pop quiz, a student is totally unprepared and randomly guesses the answer to each question. If each question has five options, then the probability of selecting the correct answer for any given question is 1/5 or 0.2. With guessing, the response on one question is not influenced by the response on another question. Thus, whether one question is answered correctly is independent of whether another question is answered correctly.

Note that $P(C) = 0.2$ and $P(I) = 0.8$. The probability that the student answers all three questions correctly is

$$P(CCC) = P(C) \times P(C) \times P(C) = 0.2 \times 0.2 \times 0.2 = 0.008$$

The probability of answering the first two questions correctly and the third question incorrectly is

$$P(CCI) = P(C) \times P(C) \times P(I) = 0.2 \times 0.2 \times 0.8 = 0.032$$

# Multiplication Rules

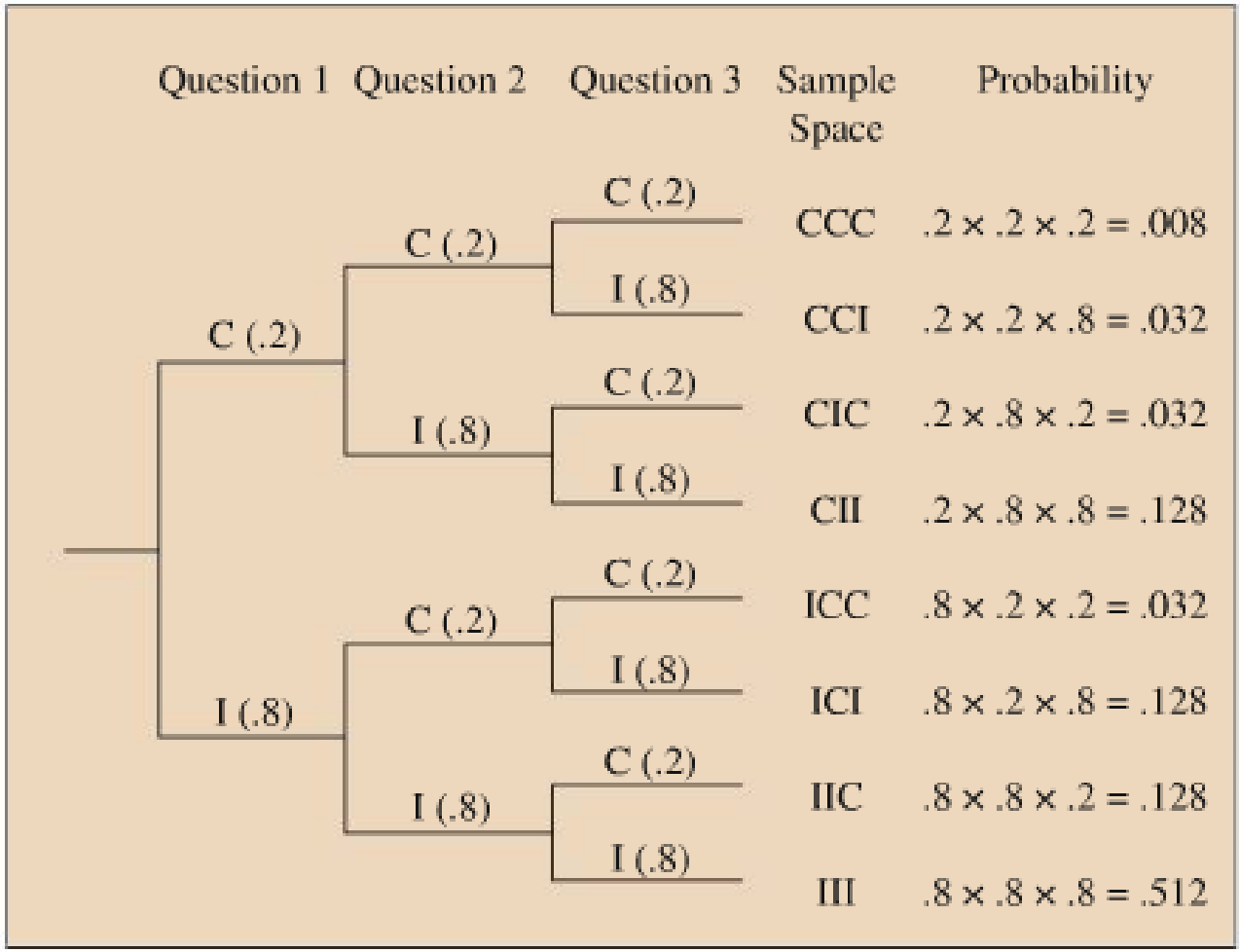



Figure 6: Tree diagram for guessing on a three-question pop quiz, given the independence assumption of correctly answering each question.

# Outline

# Conditional Probability

**Conditional probability** deals with finding the probability of an event when you know that the outcome was in some particular part of the sample space. Most commonly, it is used to find a probability about a category for one variable, when we know the outcome on another variable.

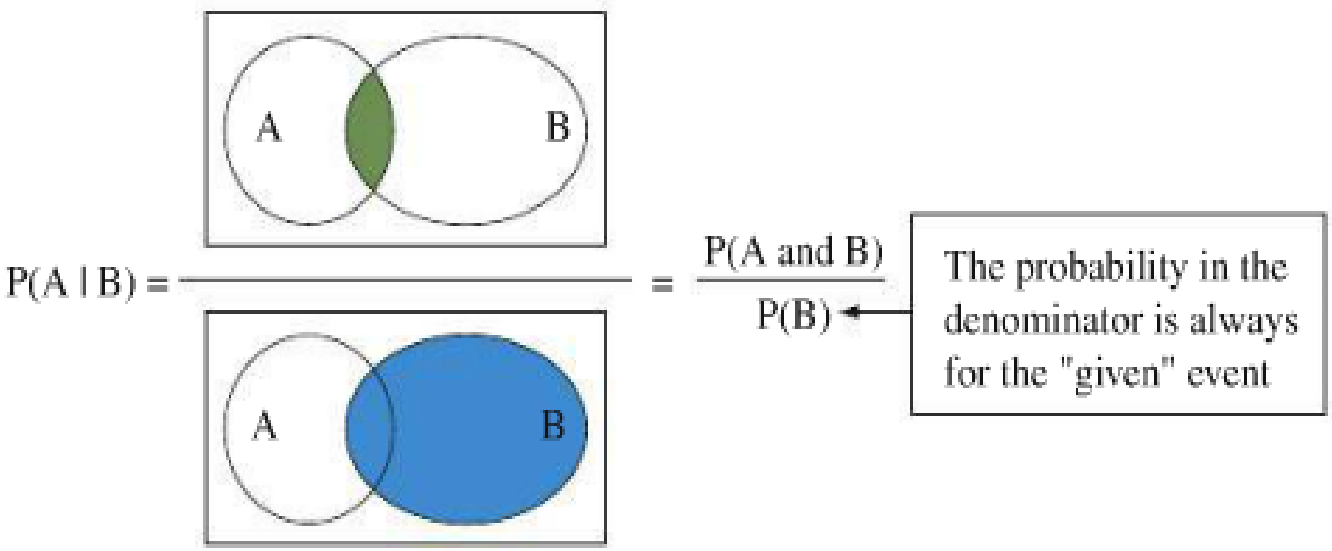



Figure 7: Venn diagram of conditional probability of event A given event B.

# Multiplication Rules for P(A and B)

For events A and B, the probability that A and B occur equals

$$P(\text{A} \cap \text{B}) = P(\text{B}) \times P(\text{A}|\text{B}) = P(\text{A}) \times P(\text{B}|\text{A})$$

## Checking for Independence

Three ways to determine if events A and B are independent

- $P(\text{A}|\text{B}) = P(\text{A})$
- $P(\text{B}|\text{A}) = P(\text{B})$
- $P(\text{A} \cap \text{B}) = P(\text{A}) \times P(\text{B})$

If any of these is true, then the others are also true and the events A and B are independent.

# Dianostic Testing

Specificity $P(NEG|S^c)$ and sensitivity $P(POS|S)$ are often reported on medical journal articles. However, what's more relevant to us once we take a diagnostic test are the conditional probabilities that condition on the test result. For instance, if a diagnostic test for Down syndrome is positive, you want to know that probability $P(S|POS)$ that Down Syndromw is truly present. We can use the tree diagram to find $P(S|POS)$ given the sensitivity and specificity.

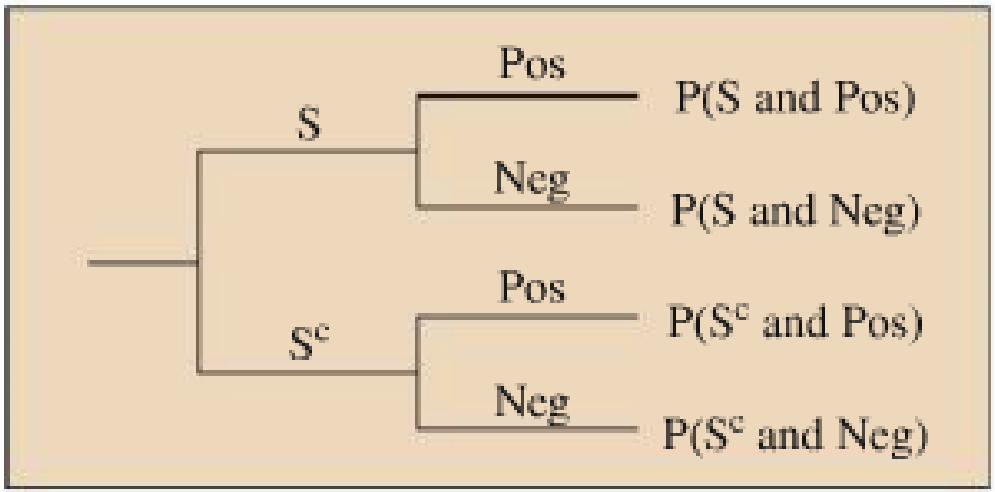



Figure 8: Tree diagram to get the conditional probability $P(S|POS)$.

# Down Syndrome

The Triple Blood Test screens a pregnant woman and provides an estimated risk of her baby being born with the genetic disorder Down syndrome. A study of 5282 women aged 35 or over analyzed the Triple Blood Test to test its accuracy.

| | **Blood Test** | | |
|---|---|---|---|
| **Down Syndrome Status** | **POS** | **NEG** | **Total** |
| D(Down) | 48 | 6 | 54 |
| $D^c$(unaffected) | 1307 | 3921 | 5228 |
| Total | 1355 | 3927 | 5282 |

Table 1: Contingency table for Triple Blood Test of Down Syndrome

Calculate the estimated cell probabilities based on the frequencies table.

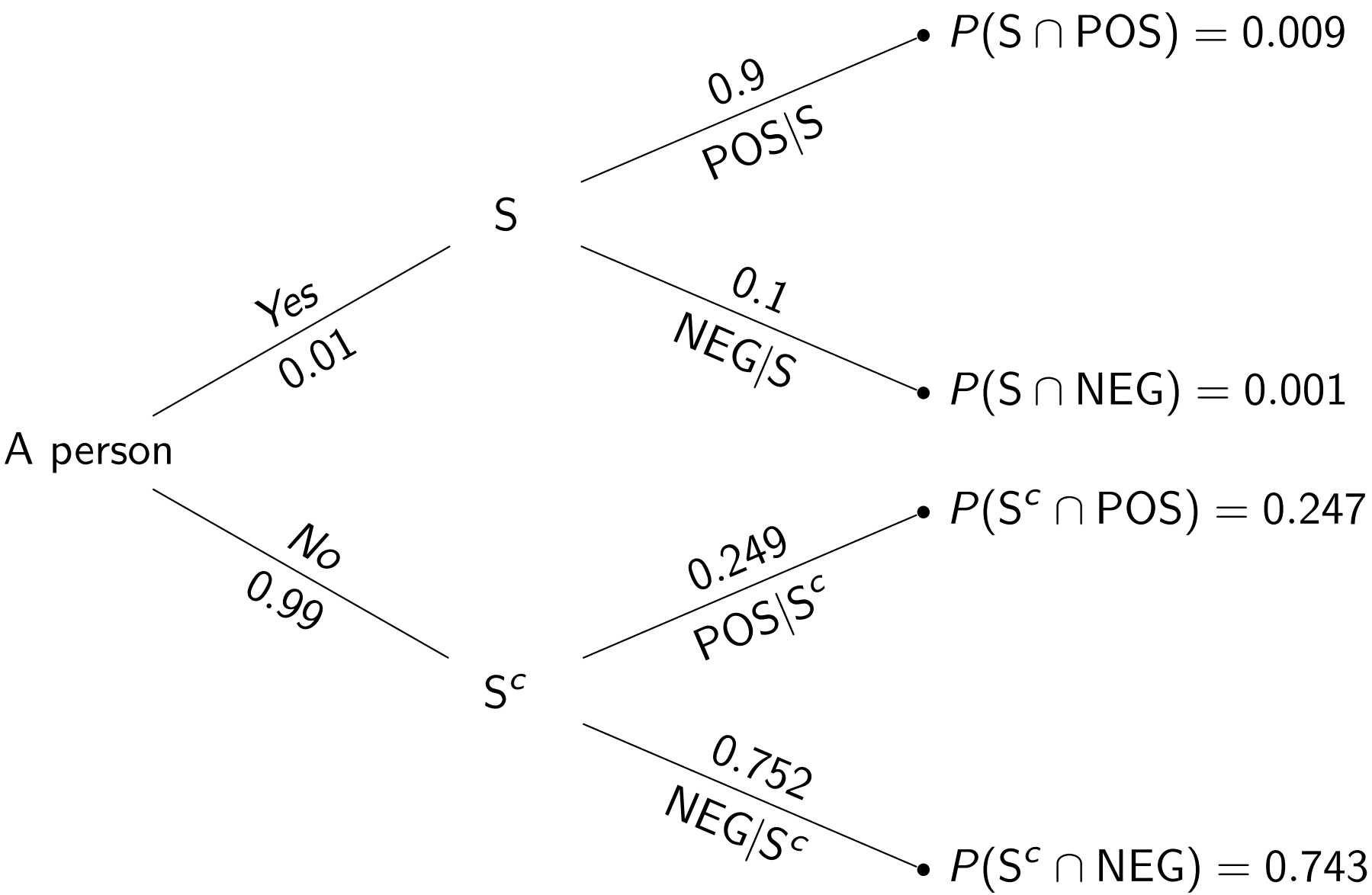
# Down Syndrome

| | Blood Test | | |
|---|---|---|---|
| **Down Syndrome Status** | **POS** | **NEG** | **Total** |
| D(Down) | 0.009 | 0.001 | 0.01 |
| $D^c$(unaffected) | 0.247 | 0.743 | 0.99 |
| Total | 0.256 | 0.744 | 1.00 |

Table 2: Probability contingency table for Triple Blood Test of Down Syndrome

1. Are the events POS and D independent?
   Note $P(POS) = 0.256$ and $P(\text{POS}|\text{D}) = \frac{P(\text{POS and D})}{P(\text{D})} = \frac{0.009}{0.01} = 0.9$.
   Since $P(\text{POS}|\text{D})$ differs from $P(\text{POS})$, the events POS and D are dependent.
2. Are the events POS and $D^c$ independent?
   Note $P(\text{POS}|\text{D}^c) = \frac{P(\text{POS and D})}{P(\text{D}^c)} = \frac{0.247}{0.99} = 0.25$. Since $P(\text{POS}|\text{D}^c)$ differs from $P(\text{POS})$, the events POS and D are dependent.

# Probability Tree for Down Syndrome



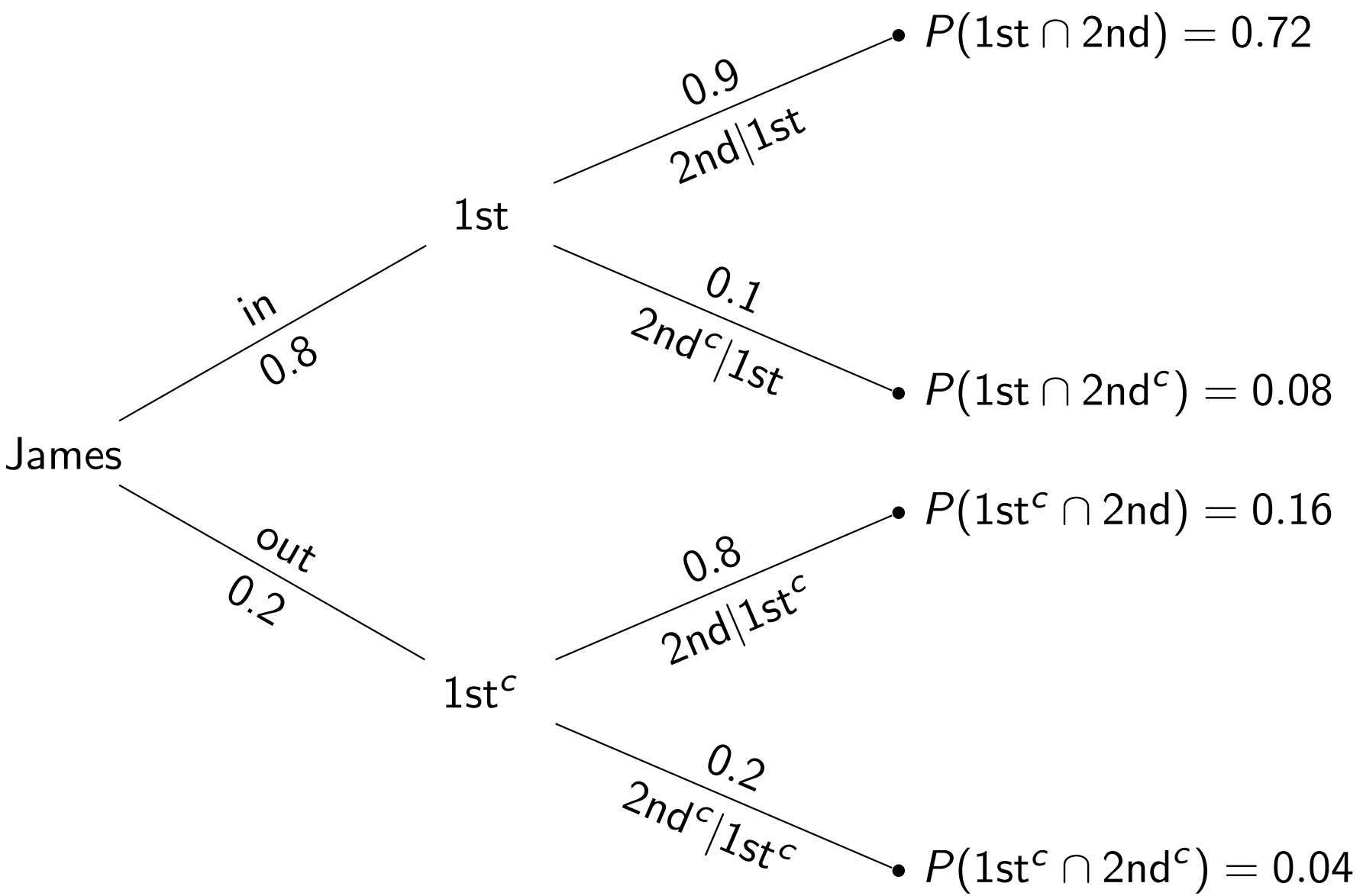
# Free Throws

Say James makes two free throws. Assume that the probability of making the 1st free throw is 0.8. If he successfully made the first one, he would successfully make the 2nd with probability 0.9; otherwise, the probability for making the 2nd successfully is still 0.8.

**Question**

1. Sketch a probability tree for the outcome of the two-free throws.
2. Draw a contingency table for the outcome of the two-free throws.
3. What's the probability of James making the 2nd successfully conditioning on he successfully made the 1st one.
4. Assume you noted James successfully made the 2nd one, but you missed the 1st shoot. What's the probability of James making the first one successfully given James successfully made the 2nd one.

# Free Throws-Probability Tree



# Free Throws - Contingence Table

| | 2nd | | |
|---|---|---|---|
| 1st | In | Out | Total |
| In | 0.72 | 0.08 | 0.8 |
| Out | 0.16 | 0.04 | 0.2 |
| Total | 0.88 | 0.12 | 1 |

Table 3: Contingency Table for the outcomes of two free-throws.

## Solution

(3) $P(\text{2nd In}|\text{1st In}) = \frac{0.72}{0.8} = 0.9$

(4) $P(\text{1st In}|\text{2nd In}) = \frac{0.72}{0.88} = 0.82$

# Outline

# Random Variable

## Random Variable(r.v.)

A **random variable** is a numerical measurement of the outcome of a random phenomenon.

Discrete r.v. The possible outcomes are a set of separate numbers. For example, $X =$ number of heads in three flips of a coin, whereas $x = 2$ is one of its possible values.

Continuous r.v. Can take any value in an interval. For example, $X =$ the change of Apple's stock price between two transaction days, whereas $x = 1.45$ or $x = -3.29$ are both possible values for $X$.

# Probability Distribution of Discrete r.v.

A discrete random variable $X$ takes a set of separate values such as $0, 1, 2, 3, \ldots$. Its probability distribution assigns a probability $P(X)$ to each possible value $x$.

- $0 \leq P(X = x) \leq 1$
- $\sum_x P(X = x) = 1$

For example, $X =$ number of games needed to determine a winner in a best of 7 series with possible values $4, 5, 6,$ or $7$. As follows a probability distribution of the random variable $X$.

| **Number of Games x** | **Probability P(x)** |
|---|---|
| 4 | 0.125 |
| 5 | 0.25 |
| 6 | 0.3125 |
| 7 | 0.3125 |

Table 4: Since $\sum_x P(x) = P(X = 4) + P(X = 5) + P(X = 6) + P(X = 7) = 1$, it is a legitimate probability distribution.

# Probability Distribution of Discrete r.v.

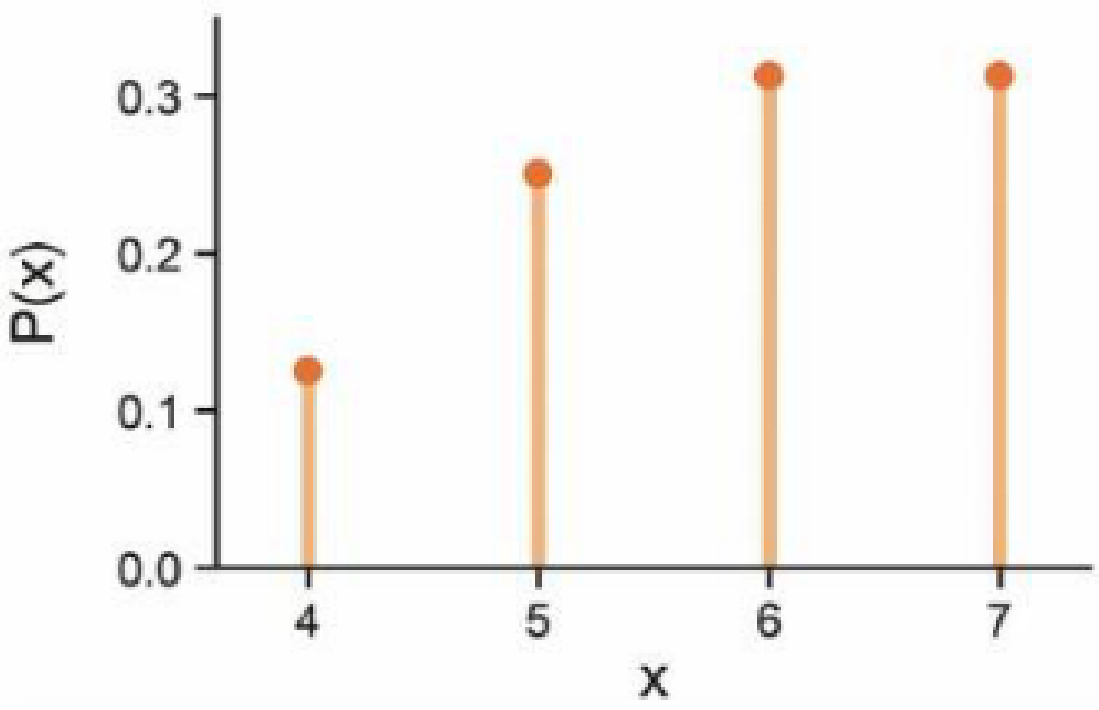



Figure 9: Graph of Probability Distribution of Number of Games needed to Determine a Winner in a best of Seven Series.

# Probability Distribution of Discrete r.v.

## mean($\mu$) and standard deviation($\sigma$)

For a discrete random variable, the mean $\mu$ and standard deviation $\sigma$ of its probability distribution are

- $\mu = \sum_x xP(x)$
- $\sigma = \sqrt{\sum_x (x-\mu)^2 P(x)}$

For example, you are given \$1,000 to invest and must choose between (i) a sure gain of \$500 and (ii) a 0.50 chance of a gain of \$1,000 and a 0.50 chance to gain nothing.

i $\mu = \sum xP(x) = 500 \times 1.0 = 500$, $\sigma = 0$

ii $\mu = \sum xP(x) = 1000 \times 0.5 + 0 \times 0.5 = 500$
$\sigma = \sqrt{(1000-500)^2 \times 0.5 + (0-500)^2 \times 0.5} = 500$

# Probability Distribution of Continuous r.v.

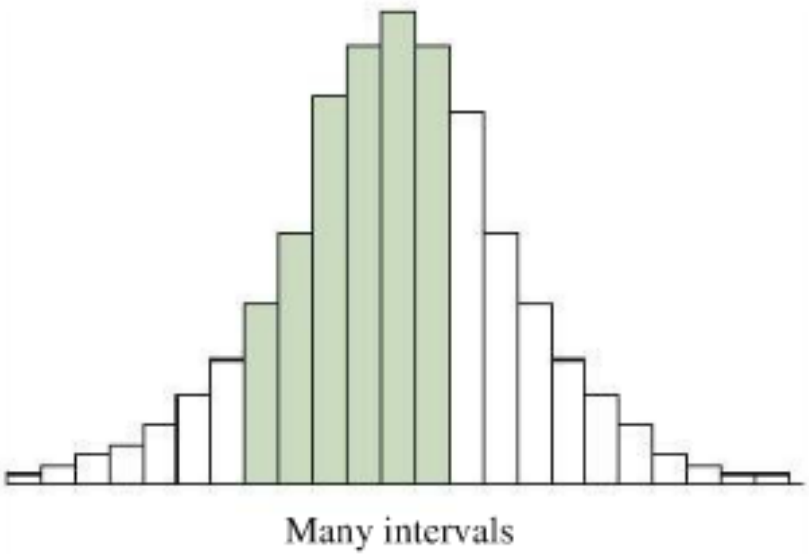
A **continuous** random variable $X$ has possible values that form an interval. Its probability distribution is specified by a curve that determines the probability that the random variable falls in any particular interval of values.

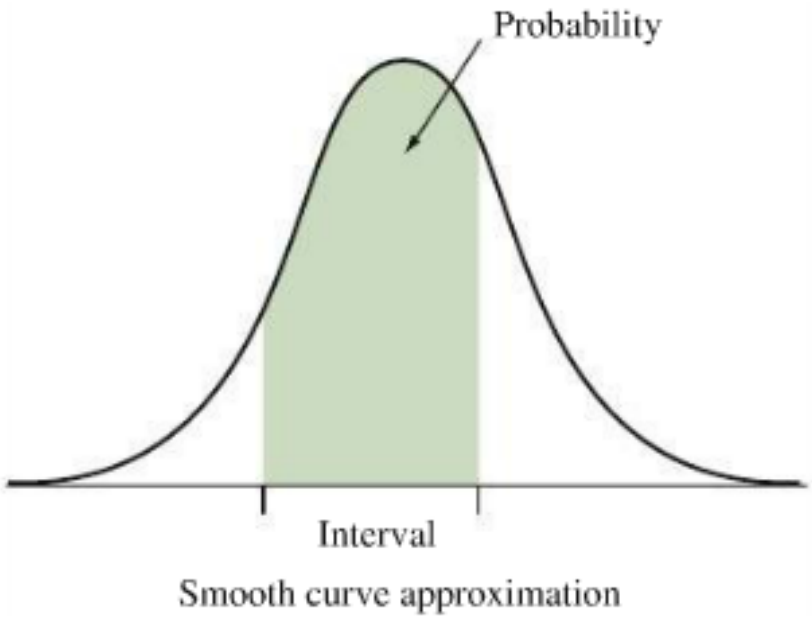
- $0 \leq P(X \leq x) \leq 1$
- $\int_x P(x)\,dx = 1$

Similarly, the mean $\mu$ and the standard deviation $\sigma$ of its probability distribution are

- $\mu = \int_x xP(x)\,dx$
- $\sigma = \sqrt{\int_x (x-\mu)^2 P(x)\,dx}$

# Probability Distribution of Continuous r.v.




**Remark**: The left is a histogram that approximates the probability distribution of a continuous r.v., and the right is the desired portray of the theoretical probability distribution.

# Normal Distribution

The **normal distribution** is symmetric, bell-shaped and characterized by its mean $\mu$ and standard deviation $\sigma$.

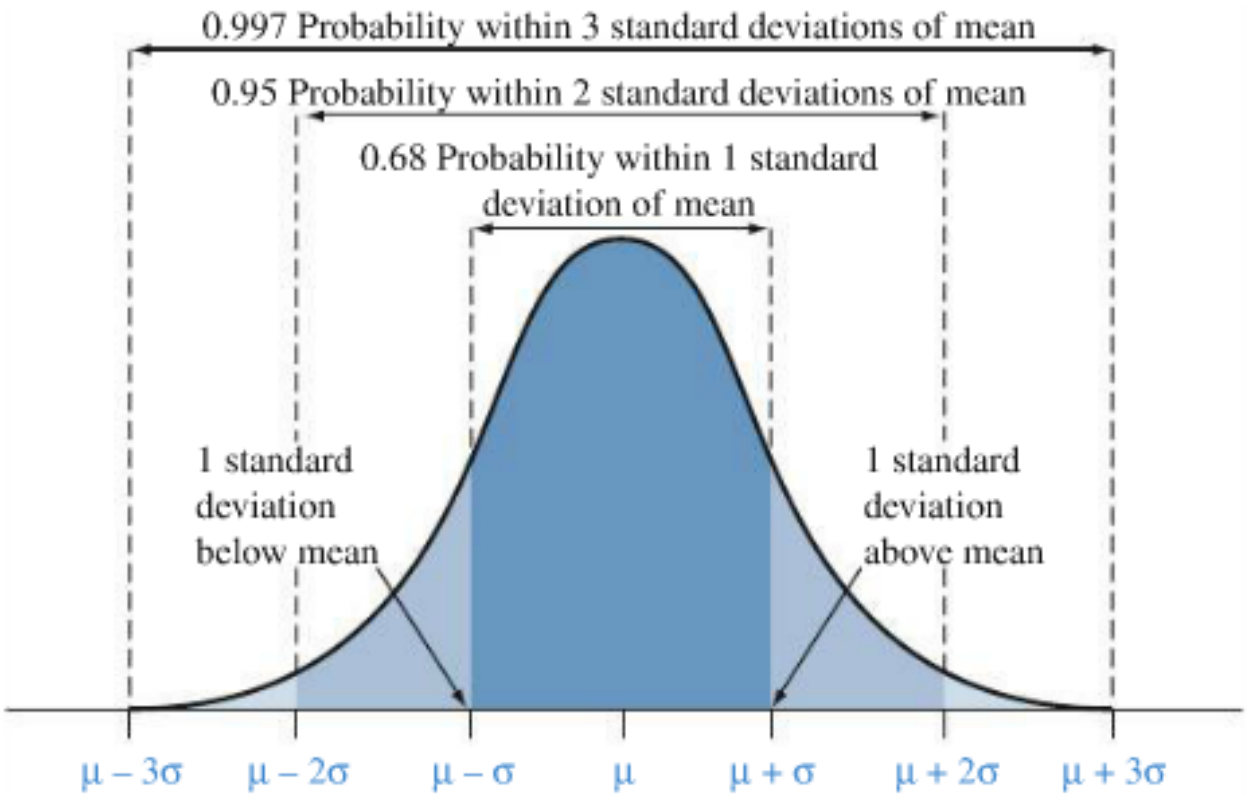



Figure 10: $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$, $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95$, $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$

# Normal Distribution

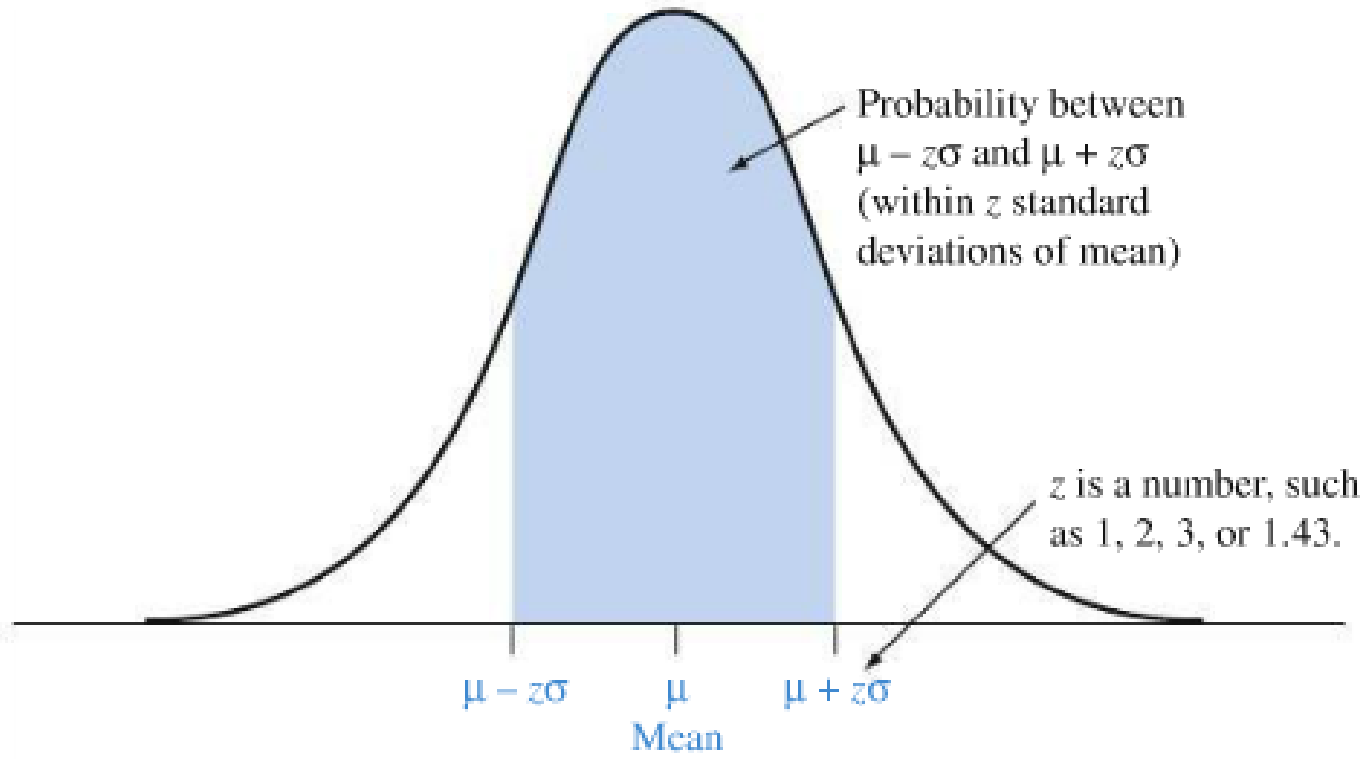



Figure 11: The probability between $\mu - z\sigma$ and $\mu + z\sigma$ is the same for every normal distribution and depends only on the value of z.

# Normal Distribution - Z table

For a normal random variable, using Z table, we can calculate probabilities for any given interval and any percentiles.

- $P(X \leq a) = P(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}) = P(Z \leq z_a)$
- For $p^{th}$ percentile, we can first find its corresponding $z$-value via Z-table. Then using the formula $z = \frac{x-\mu}{\sigma}$, the $p^{th}$ percentile is $x = \mu + z\sigma$.

In particular, the probability of any given interval can be calculated using

$$\begin{aligned} P(a \leq X \leq b) &= P(X \leq b) - P(X \leq a) \\ &= P(Z \leq z_b) - P(Z \leq z_a) \end{aligned}$$

# Normal Distribution-Example

The exam I scores have approximately a normal distribution with mean $\mu = 80$ and standard deviation $\sigma = 5$.

(a) What's the probability of a student's score less than 87?

(b) What's the probability of a student's score between 77 and 87?

(c) What's the $90^{th}$ percentile score?

**Solution**

(a) $P(X \leq 87) = P(Z \leq \frac{87-80}{5}) = P(Z \leq 1.4) = 0.9192$

(b) $P(X \leq 77) = P(Z \leq \frac{77-80}{5}) = P(Z \leq -0.6) = 0.2743$

$$P(77 \leq X \leq 87) = P(X \leq 87) - P(X \leq 77)$$
$$= 0.9192 - 0.2743 = 0.6449$$

(c) $Z_{0.9} = 1.28$, thus $X_{0.9} = \mu + Z_{0.9}\sigma = 80 + 1.28 \times 5 = 86.4$

# Normal Distribution-Example

The exam I scores have approximately a normal distribution with mean $\mu = 80$ and standard deviation $\sigma = 5$.

(a) What's the probability of a student's score less than 87?

(b) What's the probability of a student's score between 77 and 87?

(c) What's the $90^{th}$ percentile score?

**Solution**

(a) $P(X \leq 87) = P(Z \leq \frac{87-80}{5}) = P(Z \leq 1.4) = 0.9192$

(b) $P(X \leq 77) = P(Z \leq \frac{77-80}{5}) = P(Z \leq -0.6) = 0.2743$

$$P(77 \leq X \leq 87) = P(X \leq 87) - P(X \leq 77)$$
$$= 0.9192 - 0.2743 = 0.6449$$

(c) $Z_{0.9} = 1.28$, thus $X_{0.9} = \mu + Z_{0.9}\sigma = 80 + 1.28 \times 5 = 86.4$

# Binomial Distribution

The **binomial random variable** $X$ is the **number of successes** in the $n$ trials, which satisfies

- Each of $n$ trials has two possible outcomes. The outcome of interest is called a success and the other outcome called a failure.
- Each trial has the same probability of a success denoted by $p$.
- The $n$ trials are independent. That is, the result for one trial does not dependent on the results of other trails.

The probability distribution of a **binomial random variable** is called **binomial distribution**. For example, $X =$ number of heads in 3 flips of a coin is a binomial r.v. denoted by $B(3, 0.5)$ which follows the assumptions

- Each trial is a flip of the coin. Let's say head is a success (arbitrarily).
- $P(H) = 0.5$ and the success probability is the same for each flip.
- The flips are independent.

# Binomial Distribution

Denote the probability of success on a trial is $p$. For $n$ independent trials, the probability of $x$ successes equals

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

Where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is called binomial coefficient and $n!$ is called "$n$ factorial" such that $n! = n \times (n-1) \times \ldots \times 2 \times 1$. For example, $3! = 3 \times 2 \times 1 = 6$.

## Binomial Mean and Standard Deviation

The binomial probability distribution for $n$ trials with probability $p$ of success on each trial has mean $\mu$ and standard deviation $\sigma$ given by

$$\mu = np, \sigma = \sqrt{np(1-p)}$$

# Binomial Distribution

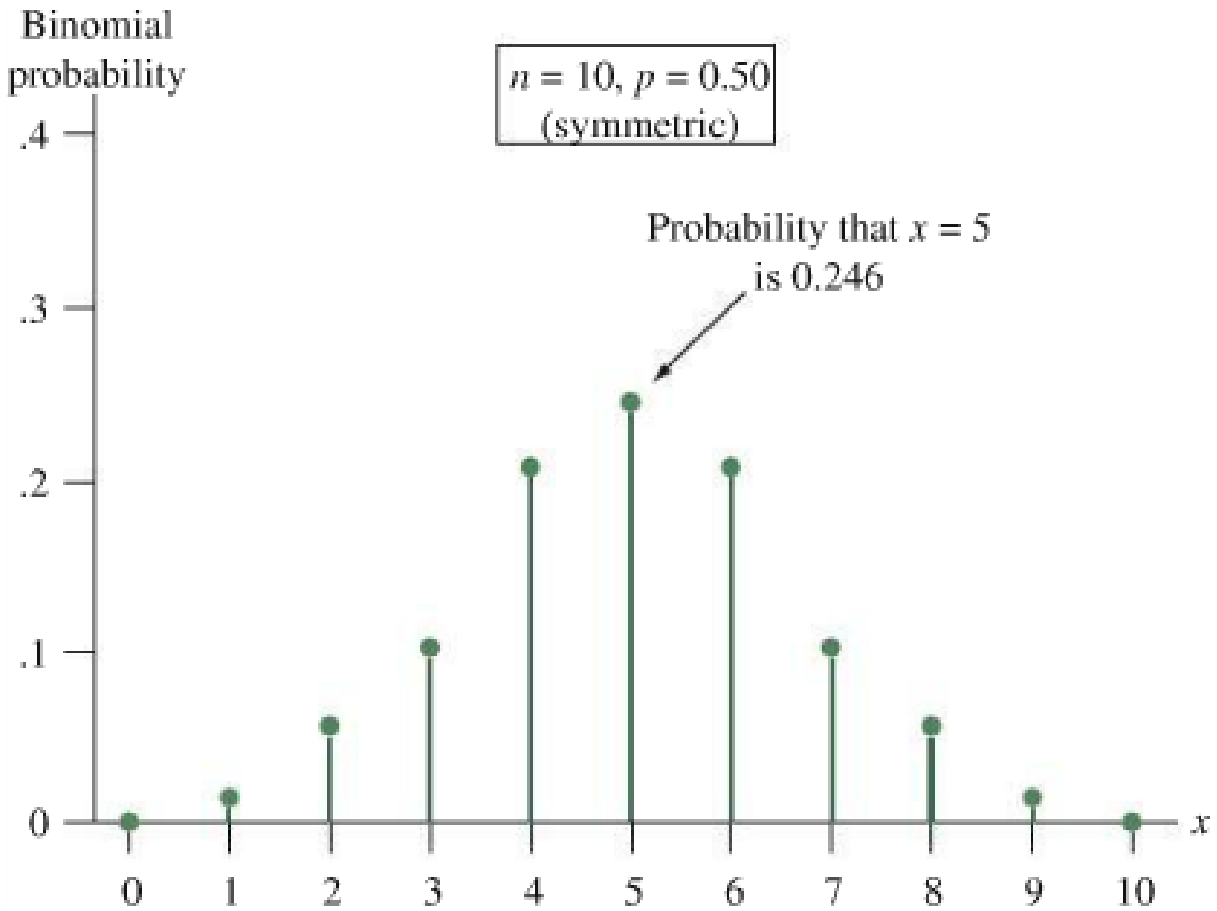



Figure 12: Binomial Distribution when $n = 10$ for $p = 0.5$

# Binomial Distribution

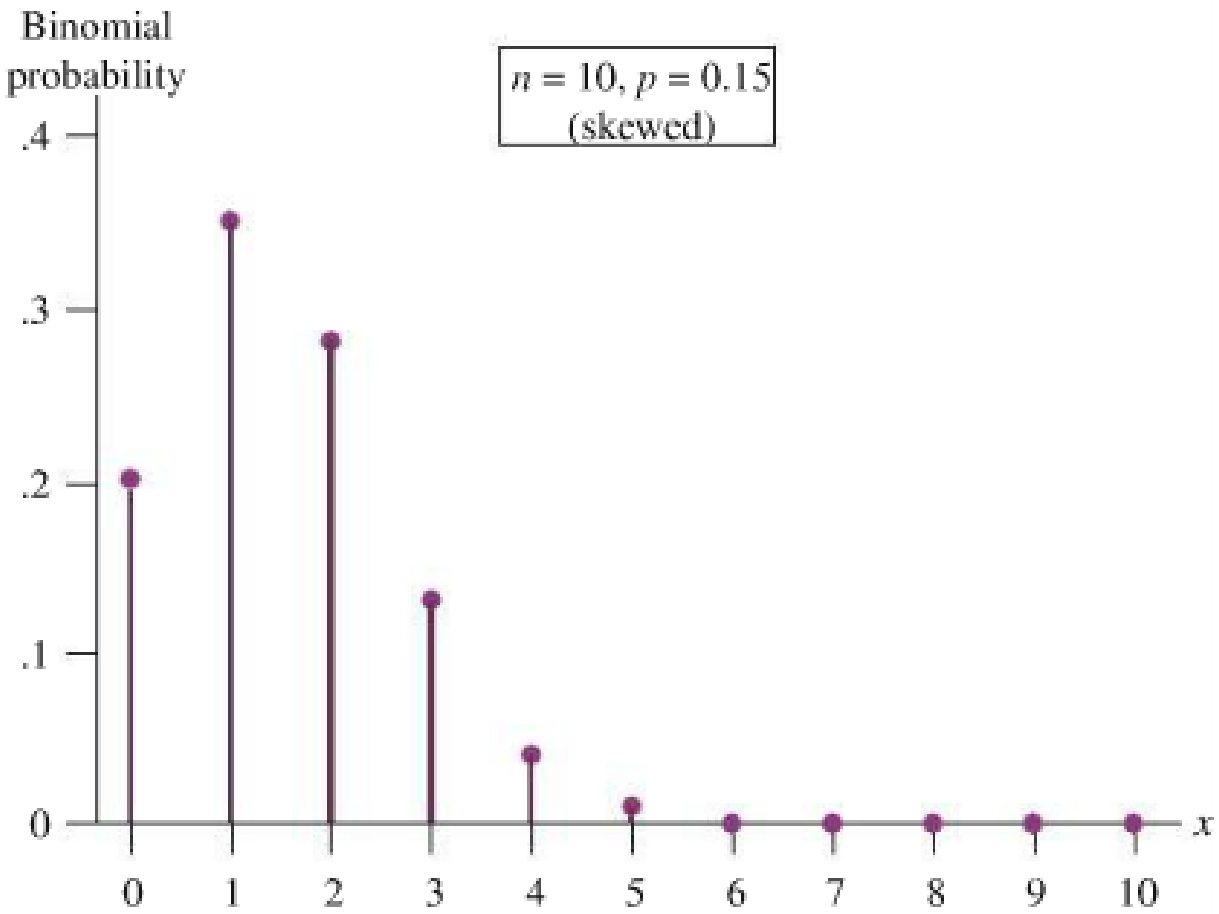



Figure 13: Binomial Distribution when $n = 10$ for $p = 0.15$

# Bidding on eBay

Assume you are bidding on four items available on eBay. For each bid, you have a 25% chance of winning it, and the outcome of the four bids are independent events. Let $X$ denote the number of winning bids out of the four items you bid on.

(a) Explain why the distribution of $X$ can be modeled by the binomial distribution.

(b) Find the probability that you win exactly 2 bids.

(c) Find the probability that you win 2 bids or fewer.

(d) Find the probability that you win more than 2 bids.

# Bidding on eBay

(a)
- Each bid is a binary random variable.
- Each bid has the same success probability $p = .25$.
- The bids are independent.

(b) $P(X = 2) = \binom{4}{2} 0.25^2 0.75^2 = 0.21$

(c)

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \binom{4}{0} 0.25^0 0.75^4 + \binom{4}{1} 0.25^1 0.75^3 + \binom{4}{2} 0.25^2 0.75^2 \\ &= 0.32 + 0.42 + 0.21 \\ &= 0.95 \end{aligned}$$

(d) $P(X > 2) = 1 - P(X \leq 2) = 1 - 0.95 = 0.05$

# Exam II Practice Problems

Feel free to do the following problems as practice for the Exam II. You can access the answers to these problems via the etext in Pearson.

**Page 231-232** 5.33, 5.38, 5.39, 5.41, 5.42

**Page 249-250** 5.78, 5.79, 5.88, 5.89, 5.91, 5.92

**Page 265-266** 6.3, 6.7, 6.10, 6.11

**Page 278-279** 6.25, 6.29, 6.32, 6.31

**Page 287-289** 6.35, 6.45, 6.46, 6.49