Machine Learning

|Sampling Distribution

Prof. Dr. Mostafa Elhosseini

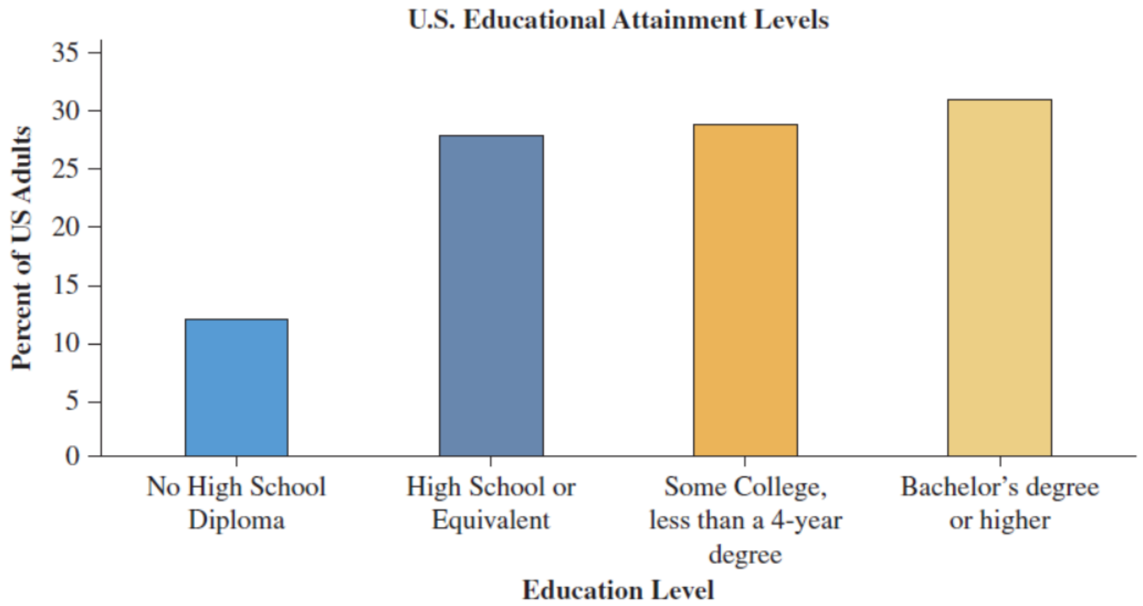
https://youtube.com/@ProfElhosseiniTechVerse

Agenda

- How to choose a suitable **Distribution** for a Hyperparameter
- Distribution
- Frequency Table
- Frequency, Proportion, & Percentage
- Graphical Summaries
- Random Variable
- Probability Mass Function
- Histogram
- Probability Density Function (PDF)

Distribution

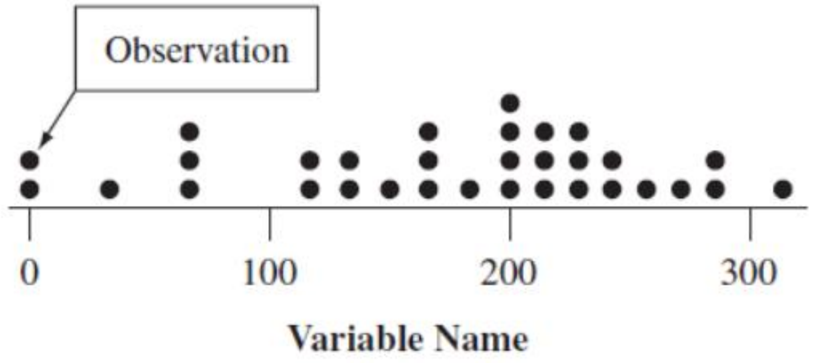
- Describes **how the observations** fall (are **distributed**) across the range of possible values
  - The distribution for a categorical variable then simply shows all possible categories and the number (or proportion) of observations falling into each category

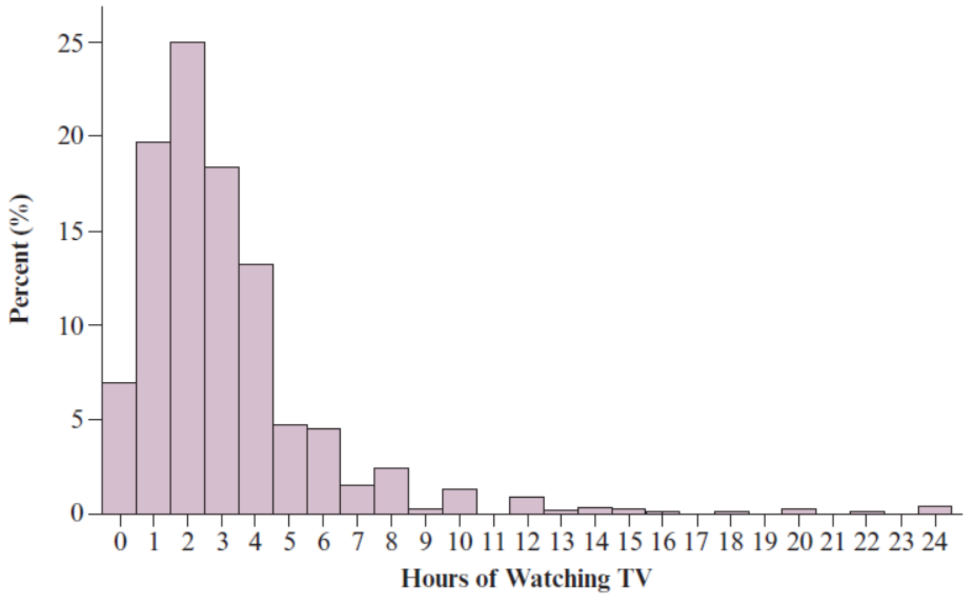


▲ **Figure 1.1** Educational Attainment, Based on a Sample of 78,000 Households in the 2013 Current Population Survey. (*Source:* Data from United States Census Bureau.)

Distribution

- Describes **how the observations** fall (are **distributed**) across the range of possible values
  - For a quantitative variable, the entire range of possible values is split up into separate intervals, and the number (or proportion) of observations falling in each interval is given





▲ **Figure 2.5** Histogram of GSS Responses about Number of Hours Spent Watching TV on an Average Day. *Source:* Data from CSM, UC Berkeley.

Distribution

- Distribution can be displayed by a **graph or a table** (**Frequency Table**)

Young non-citizens in the U.S. The table shows the number of 18- to 24-year-old noncitizens living in the United States between 2010 and 2012.

Noncitizens aged 18 to 24 in the United States

Region of Birth	Number (in Thousands)
Africa	115
Asia	590
Europe	148
Latin America & Caribbean	1,666
Other	49
Total	**2,568**

Source: www.census.gov

Enrollment trends Examples 18 and 19 presented graphs showing the total student enrollment at a U.S. university between 2004 and 2012 and the data for STEM major enrollment during that same time period. The data are repeated here.

Enrollment at the U.S. University

Year	Total Students	STEM Majors
2004	29,404	2,003
2005	29,693	1,906
2006	30,009	1,871
2007	30,912	1,815
2008	31,288	1,856
2009	32,317	1,832
2010	32,941	1,825
2011	33,878	1,897
2012	33,405	1,854

Frequency Table

- Displays the **distribution** of a variable **numerically**

France is most popular holiday spot Which countries are most frequently visited by tourists from other countries? The table shows results according to *Travel and Leisure* magazine (2005).

Most Visited Countries, 2005

Country	Number of Visits (millions)
France	77.0
China	53.4
Spain	51.8
United States	41.9
Italy	39.8
United Kingdom	24.2
Canada	20.1
Mexico	19.7

Source: Data from *Travel and Leisure* magazine, 2005.

Frequency Table

Table 2.1 Frequency of Shark Attacks in Various Regions for 2004–2013*

Region	Frequency	Proportion	Percentage
Florida	203	0.295	29.5
Hawaii	51	0.074	7.4
South Carolina	34	0.049	4.9
California	33	0.048	4.8
North Carolina	23	0.033	3.3
Australia	125	0.181	18.1
South Africa	43	0.062	6.2
Réunion Island	17	0.025	2.5
Brazil	16	0.023	2.3
Bahamas	6	0.009	0.9
Other	138	0.200	20.0
Total	**689**	**1.000**	**100.0**

$$\text{proportion} = \frac{(frequency\ of\ that\ class)}{(sum\ of\ all\ frequencies)}$$

Frequency, Proportion, & Percentage Example

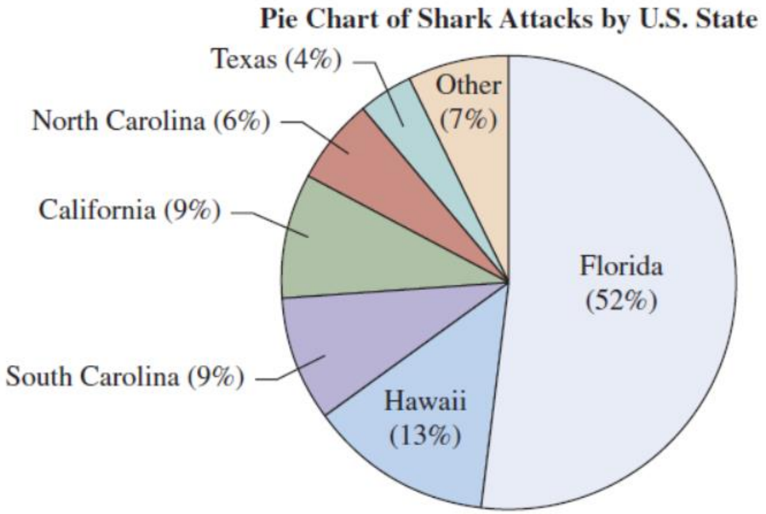
- There were 203 reported shark attacks in Florida between 2004 and 2013. Table 2.1 classifies 689 shark attacks reported from 2004 through 2013, so, for Florida,
  - 203 is the frequency.
  - 203/689 = 0.295 is the proportion and relative frequency.
  - 29.5 is the percentage 0.295 *100= 29.5%.

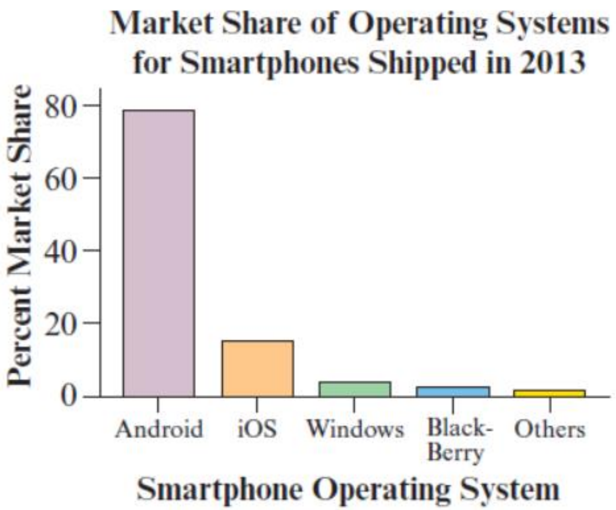
Graphical Summaries of Data

- Looking at a **graph** often gives you **more of a feel** for a variable and its distribution than looking at the raw data or a frequency table
- For the categorical variable
  - the pie chart and
  - the bar graph
- For the quantitative variable displays
  - the dot plot,
  - stem-and-leaf plot, and
  - histogram

Graphical Summaries of Data

- For the categorical variable
  - Pie chart
  - Bar graph

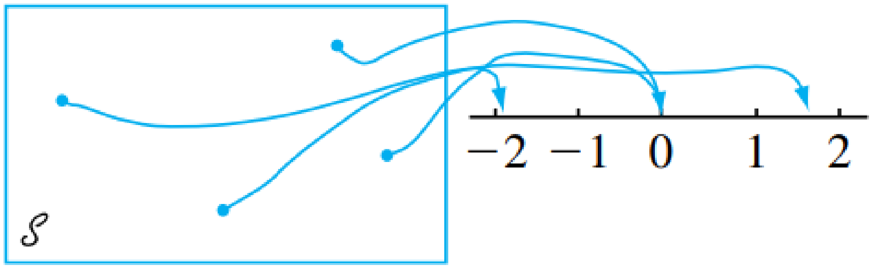




Source: idc.com

Random Variable

- In many situations, it is desirable to **assign a numerical value to each outcome of an experiment**.
- Such an assignment is called a **random variable (rule of association)**
  - The **number** among the sample of ten components that **fail to last 1000 hours**
  - The **total weight** of baggage for a sample of 25 airline passengers



Random Variable

- Random variables are customarily denoted by **uppercase letters**, such as $\boldsymbol{X}$ and $\boldsymbol{Y}$, near the end of our alphabet
- we will now use **lowercase letters** to **represent some particular value** of the corresponding random variable.
- The notation $X(s) = x$ means that x is the value associated with the outcome s by the random variable X

Example - 01

- Suppose that an electrical engineer has on hand **six resistors**. **Three** of them are **labeled $10\ \Omega$** and the **other three** are labeled $\mathbf{20\ \Omega}$. The engineer **wants to connect a $10\ \Omega$ resistor and a $20\ \Omega$ resistor** in series, to **create a resistance of $30\ \Omega$**.
- Now suppose that in fact the **three resistors labeled $10\ \Omega$** have actual resistances of **9,10, and $11\ \Omega$**, and that the **three resistors labeled $20\ \Omega$** have actual resistances of **19, 20, and $21\ \Omega$**.
- The process of **selecting one resistor of each type is an experiment** whose **sample space consists of nine** equally likely outcomes

Example - 01

- Now **what is important** to the engineer in this experiment is the **sum of the two resistances, rather than their individual values**.
- Therefore, we **assign to each outcome a number equal to the sum of the two resistances selected**. This assignment, represented by the letter X

Outcome	Probability
(9, 19)	1/9
(9, 20)	1/9
(9, 21)	1/9
(10, 19)	1/9
(10, 20)	1/9
(10, 21)	1/9
(11, 19)	1/9
(11, 20)	1/9
(11, 21)	1/9

Example - 01

- A **random variable** assigns a numerical value to each outcome in a sample space.
- We can compute probabilities for random variables in an obvious way.
- The **event X = 29** corresponds to the event **{(9, 20), (10, 19)}** of the sample space.
- Therefore $\boldsymbol{P(X = 29)} = P(\{(9,20),(10,19)\}) = \mathbf{2/9}$.

Outcome	X	Probability
(9, 19)	28	1/9
(9, 20)	29	1/9
(9, 21)	30	1/9
(10, 19)	29	1/9
(10, 20)	30	1/9
(10, 21)	31	1/9
(11, 19)	30	1/9
(11, 20)	31	1/9
(11, 21)	32	1/9

Example - 01

- List the possible values of the random variable X, and find the probability of each of them.

x	$P(X = x)$
28	1/9
29	2/9
30	3/9
31	2/9
32	1/9

Outcome	X	Probability
(9, 19)	28	1/9
(9, 20)	29	1/9
(9, 21)	30	1/9
(10, 19)	29	1/9
(10, 20)	30	1/9
(10, 21)	31	1/9
(11, 19)	30	1/9
(11, 20)	31	1/9
(11, 21)	32	1/9

Example - 02

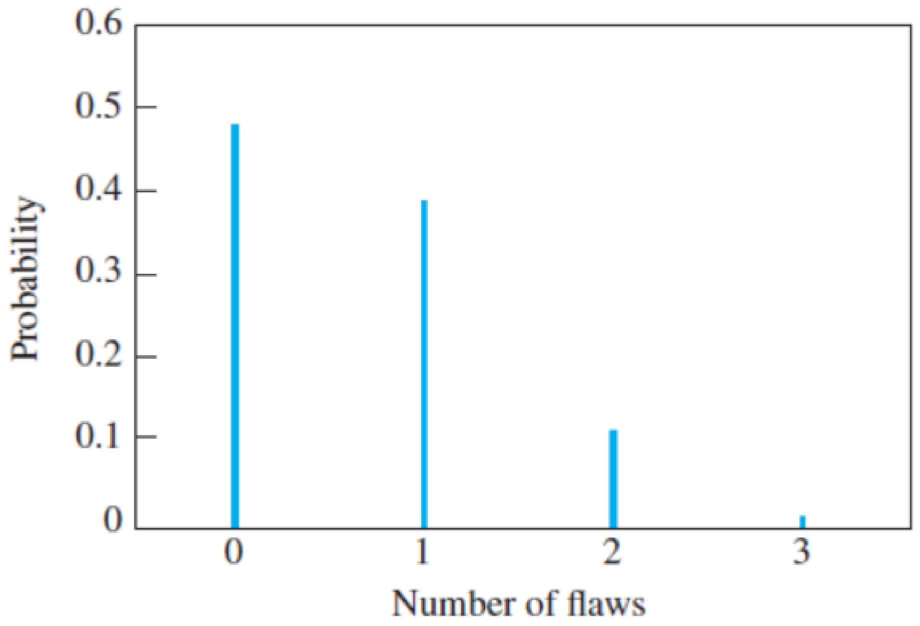
- The experiment of tossing a coin twice. Let X represent number of heads appearing
  - X is a random variable with values 0, 1, 2
  - $X(T,T) = 0$
  - $X(H,T) = 1$

Probability Mass Function

- **The number of flaws in a 1-inch length of copper wire** manufactured by a certain process varies from wire to wire. Overall, **48%** of the wires produced have **no flaws**, **39%** have **one flaw**, **12%** have **two flaws**, and **1%** have **three flaws**. Let X be the number of flaws in a randomly selected piece of wire. Then
  - $P(X = 0) = 0.48 \quad P(X = 1) = 0.39$
  - $P(X = 2) = 0.12 \quad P(X = 3) = 0.01$
- The list of possible values **0, 1, 2, 3**, **along** with the **probabilities** for each, provide a **complete description of the population** from which X is drawn. This description has a name—the **probability mass function**

Probability Mass Function

- The **probability mass function can be represented by a graph** in which a **vertical line** is drawn **at each of the possible values** of the random variable. The **heights** of the lines are **equal to the probabilities of the corresponding values**.
- The physical interpretation of this graph is that **each line** represents a **mass equal to its height**



Example Best of Seven

- Many playoff and championship games in professional sports such as basketball, baseball, or hockey are **decided by a best of seven series**, in which the team that is the first to get four victories wins the series. This often adds drama because the **number of games** that actually need to be played **is not fixed in advance**. Let X represent the **number of games needed to determine a winner** in a best of 7 series, with possible outcomes 4 (when one teams wins the first four games, called a **sweep**), 5, 6, or 7
- we assume that the chances for either team to win a game are 50% for all games played

Example Best of Seven

- What is the probability that a best of seven series will be decided by at least six games?

Number of Games x	Probability $P(x)$
4	$1/8 = 0.125$
5	$1/4 = 0.25$
6	$5/16 = 0.3125$
7	$5/16 = 0.3125$

Number of Games x	Probability $P(x)$
4	$1/8 = 0.125$
5	$1/4 = 0.25$
6	$5/16 = 0.3125$
7	$5/16 = 0.3125$

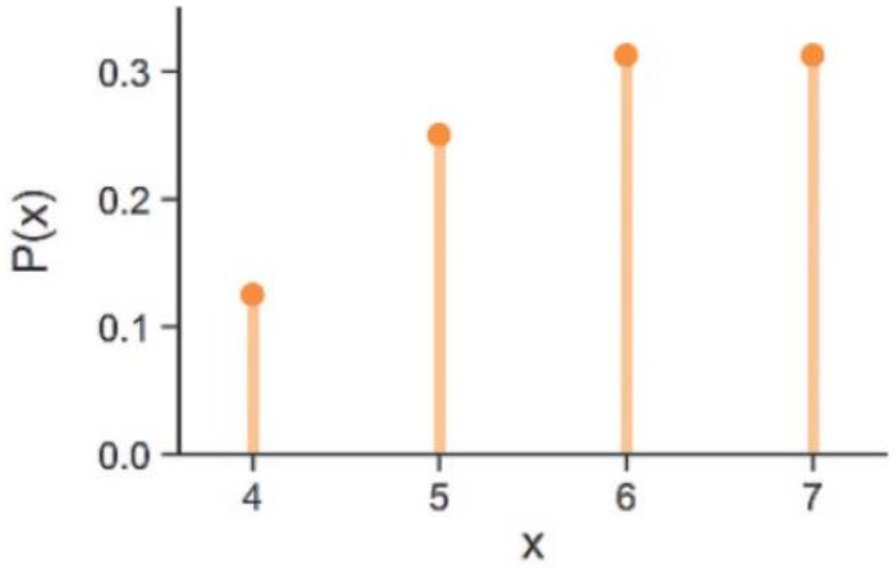
Example Best of Seven

- Why is X a random variable?
- Before the start of the series, **we are uncertain just how many games will be necessary to determine a winner**. This number will be determined by a random phenomenon, making X = number of games a random variable whose values are determined by chance
- What is the probability that a best of seven series will be decided by at least six games?
  - P(6) + P(7) = 5/16 + 5/16 = 10/16 = 0.625.

Example Best of Seven

Number of Games x	Probability $P(x)$
4	$1/8 = 0.125$
5	$1/4 = 0.25$
6	$5/16 = 0.3125$
7	$5/16 = 0.3125$

- P(6) denotes the probability that the random variable takes the value 6; that is, the probability that the number of games needed to be played to determine a winner is 6

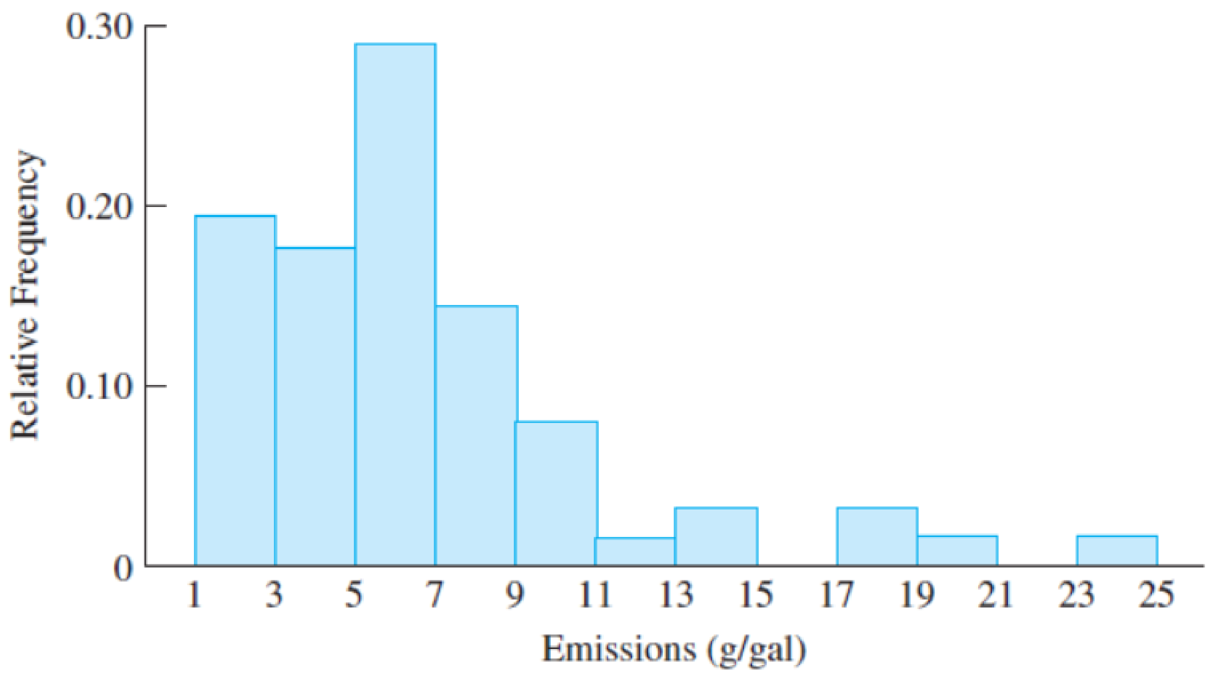


Continuous Random Variable

Histogram

- A **histogram** is a graphic that gives an idea of the **shape** of a sample
- **Frequency** presents the **numbers of data points** that fall into each of the class intervals
- **Relative Frequency** presents the frequencies divided by the total number of data points
- **Density** presents the **relative frequency divided by the class width**

Histogram



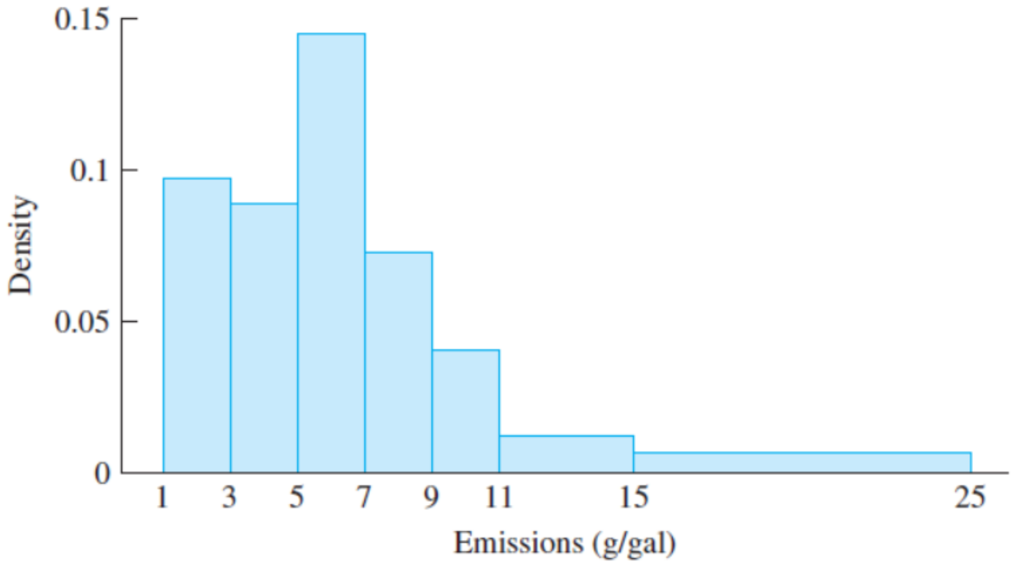
- **In some cases**, histograms are drawn with **class intervals** of **differing** widths.

Histogram – unequal class width

Class Interval (g/gal)	Frequency	Relative Frequency	Density
1–< 3	12	0.1935	0.0968
3–< 5	11	0.1774	0.0887
5–< 7	18	0.2903	0.1452
7–< 9	9	0.1452	0.0726
9–< 11	5	0.0806	0.0403
11–< 15	3	0.0484	0.0121
15–< 25	4	0.0645	0.0065

- Because the **class widths vary in size**, the **densities** are **no** longer **proportional to the relative frequencies**. Instead, the densities adjust the relative frequency for the width of the class.

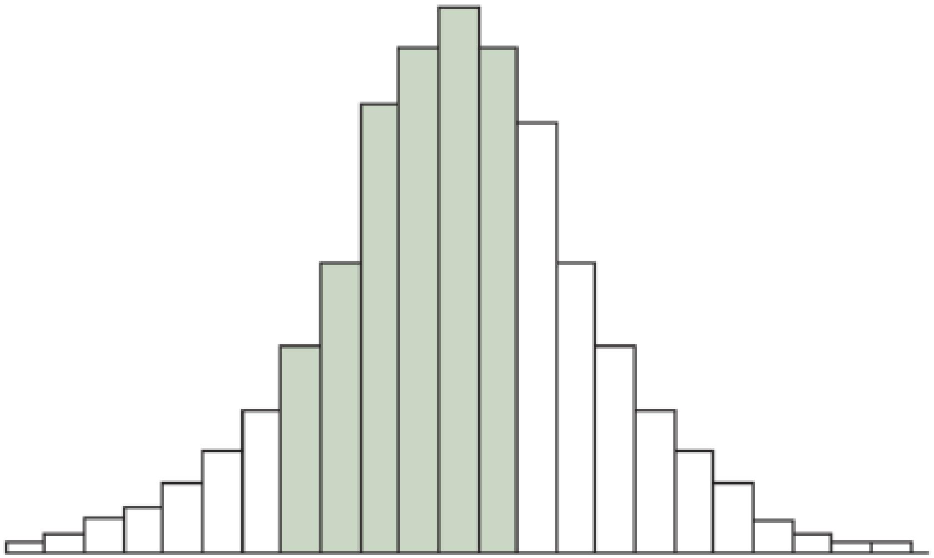
Histogram – unequal class width



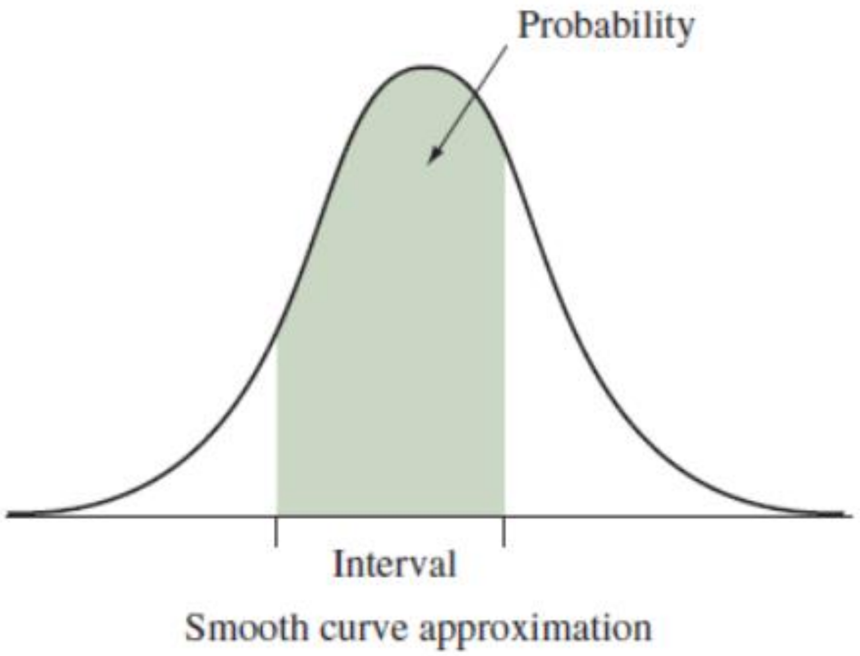
- when the **classes have unequal widths**, the **heights** of the rectangles must be **set equal to the densities**.
- The **areas** of the rectangles then **represent the relative frequencies**

Probability Density Function (PDF)

- As the **number of intervals increases**, with their **width narrowing**, the **shape** of the histogram **gradually approaches a smooth curve**.
- We'll use such curves to portray probability distributions of continuous random variables


Many intervals

Probability Density Function (PDF)



- The **probability that X falls between any two values a and b** is equal to the **area** under the histogram between a and b
- Because the histogram in this case is represented by a **curve**, the **probability** would be found by computing an **integral**

Probability Density Function (PDF)

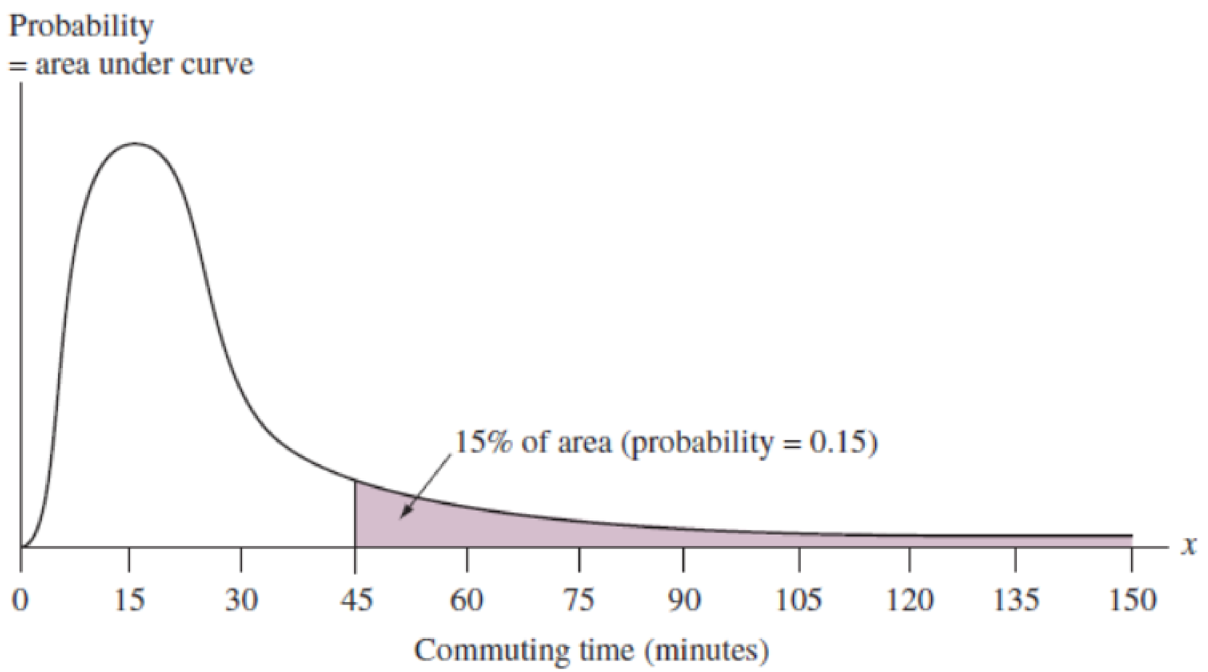
- The proportion of the population whose values of X lie between a and b is given by

$$\int_{a}^{b} f(x)\, dx$$

- Probability that the random variable X takes on a value between a and b.
- Note that the **area under the curve does not depend on** whether the endpoints a and b are included in the interval

Example

- Probability distribution of X = **commuting time** to work for workers in the United States, based on information from the 2009 American Community Survey



▲ **Figure 6.2 Probability Distribution of Commuting Time.** The area under the curve for values higher than 45 is 0.15. **Question** Identify the area under the curve representing the probability of a commute less than 15 minutes, which equals 0.27.