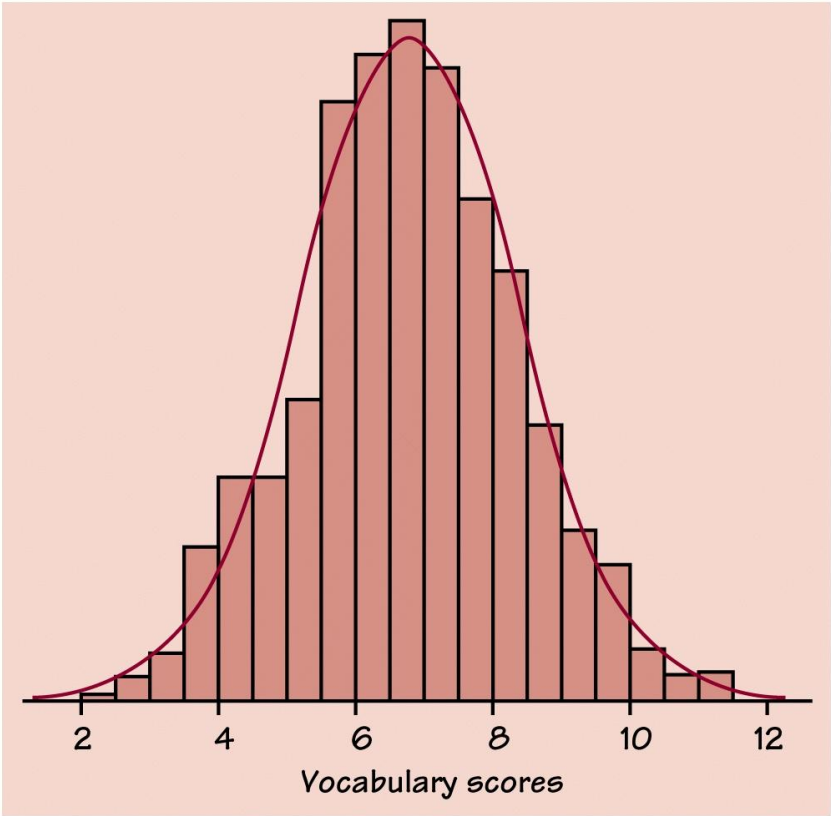
Chapter 3
Modeling Distributions of Data

Section 3.1
Measuring Locations in Distribution

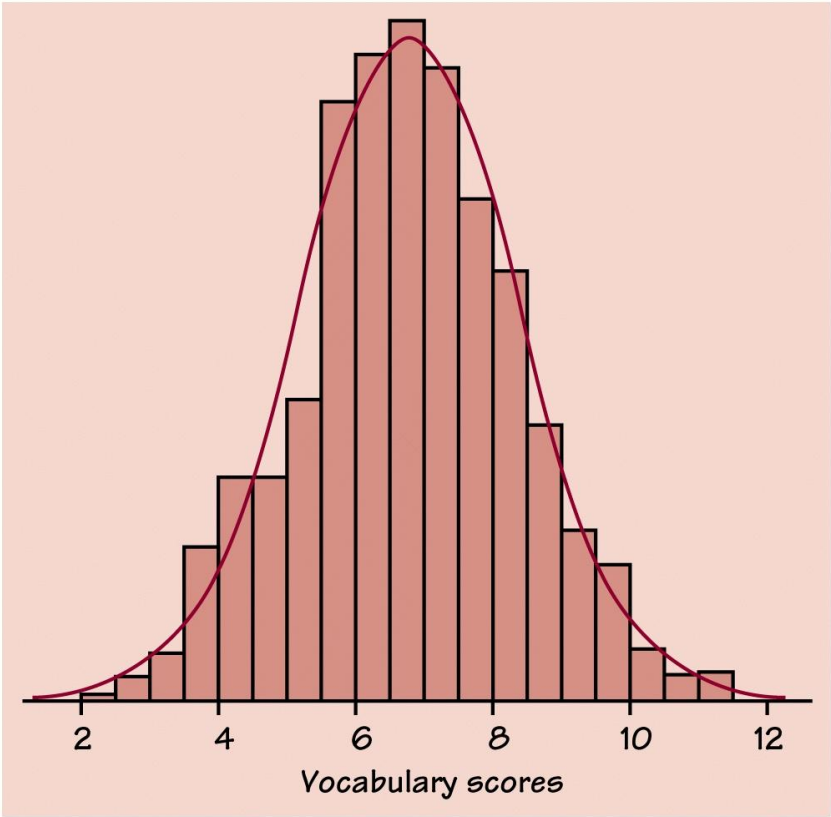
Histograms and Smooth Curves

- This is a histogram of the scores of 947 seventh-grade students on a standardized vocabulary test.
- Scores on this test have a very regular distribution.



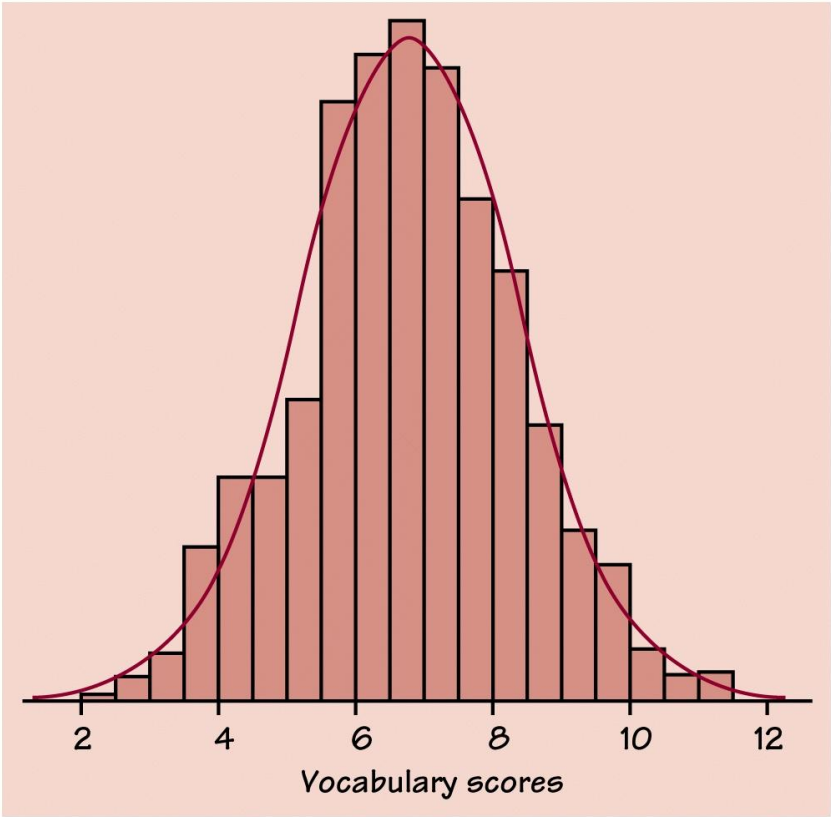
Histograms and Smooth Curves

- Let's describe this histogram
  - Shape
  - Peak
  - Gaps
  - Outliers



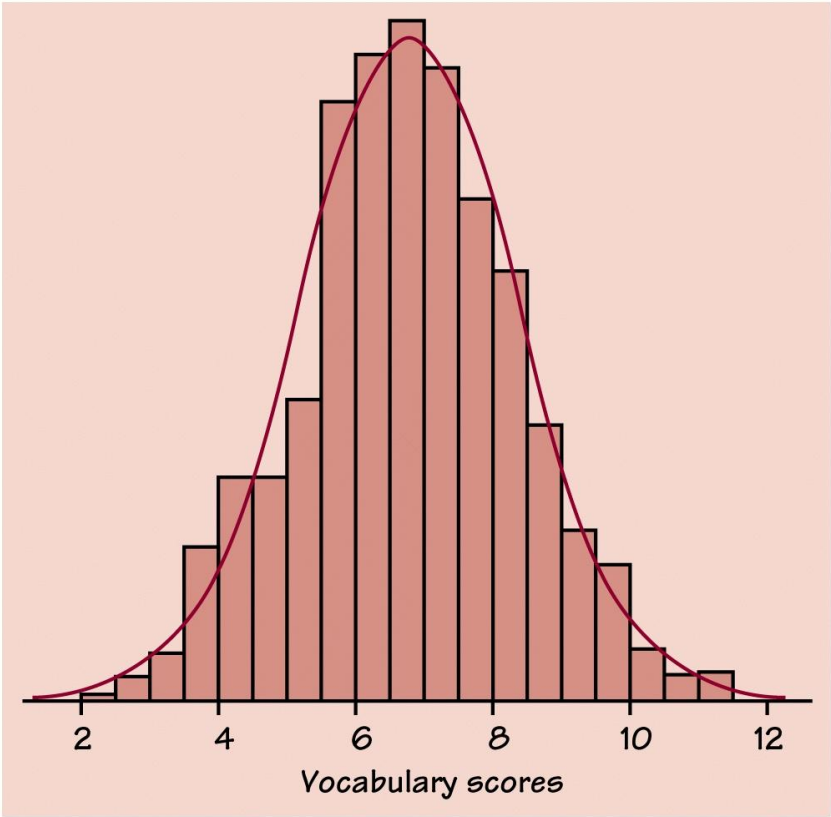
Histograms and Smooth Curves

- The histogram is:
  - Symmetric
  - Single center peak
  - ‘Tails’ fall off smoothly
  - No gaps
  - No outliers



Histograms and Smooth Curves

- The smooth curve drawn through the tops of the histogram bars is a good description of the overall pattern of the data.
- You can think of drawing a curve through the tops of the histogram and smoothing out the irregular ups and downs of the bars.
- This is called a density curve.



Density Curves

- One important distinction between histograms and density curves is most histograms show the counts of observations in each class by the height of their bars and therefore by the areas of their bars.

- We set up curves to show the proportion of observations in any region by areas under the curve.

- To do that, we choose the scale so that the total area under the curve is exactly 1.

Density Curves

Why we use it:

- Sometimes histograms give us more information than we need to analyze our data.

- The density curve is a model of a histogram with an infinite number of observations and infinitely many classes.

Density Curves

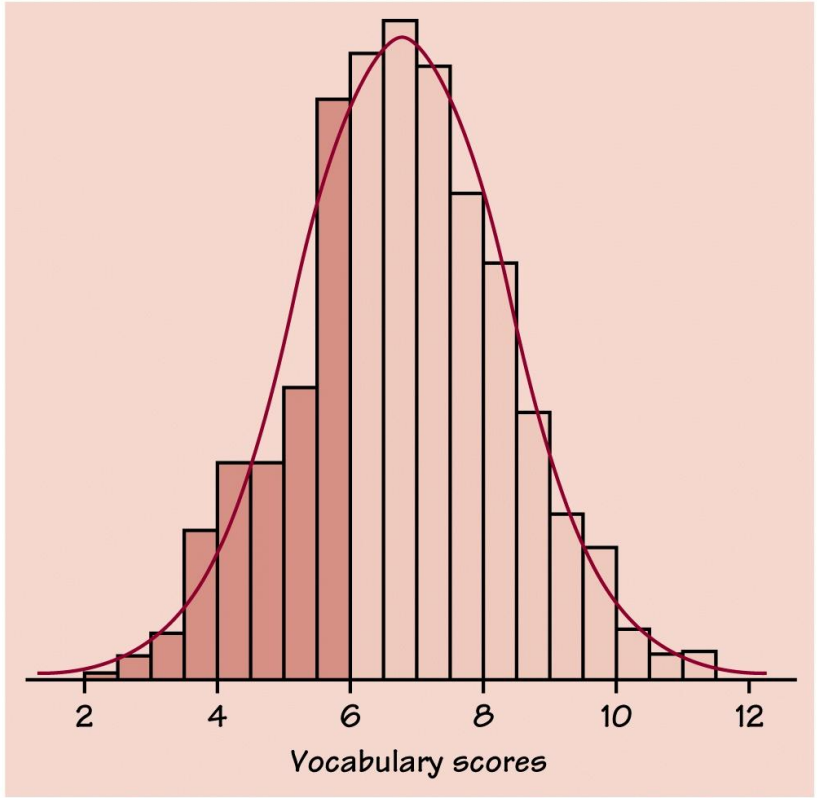
- A density curve is always on or above the horizontal axis.
- The vertical axis of the density curve is always in percents.
- The total area under any density curve is set equal to 1.
- Represents 100% of the population

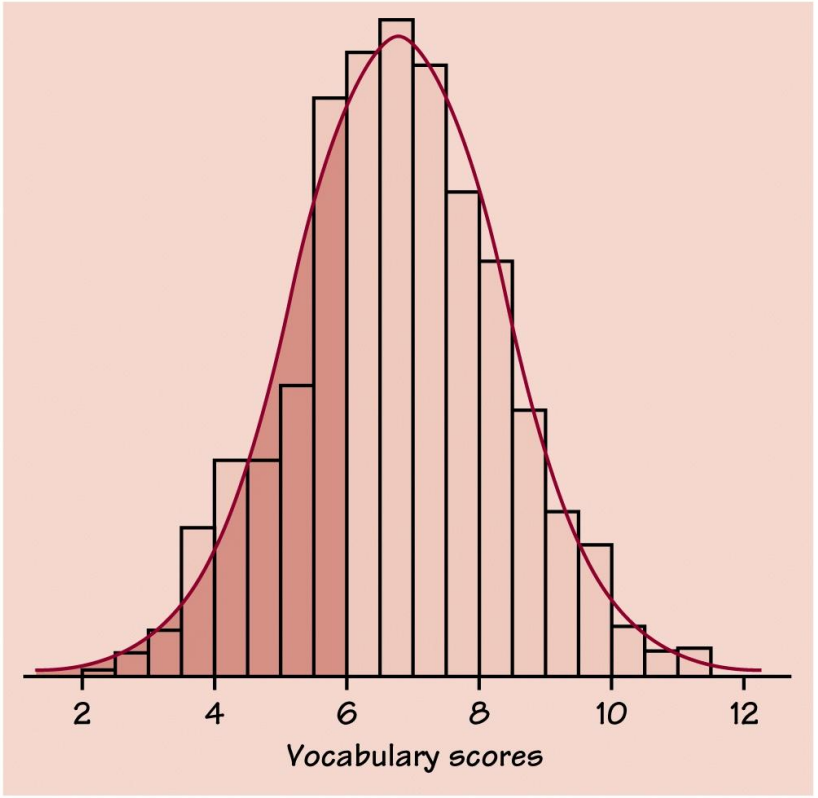
Percentiles

- The p^{th} percentile of a distribution is the value with p percent of the observations less than or equal to it.

Histogram vs. Density Curve

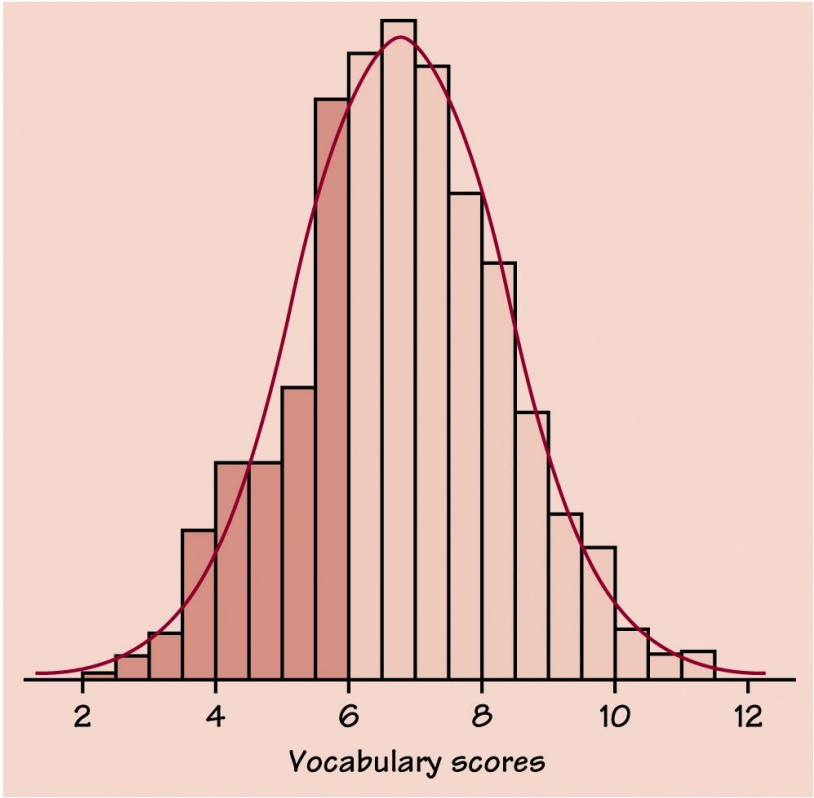
To find percentages from a density curve, we will find the area under the curve.





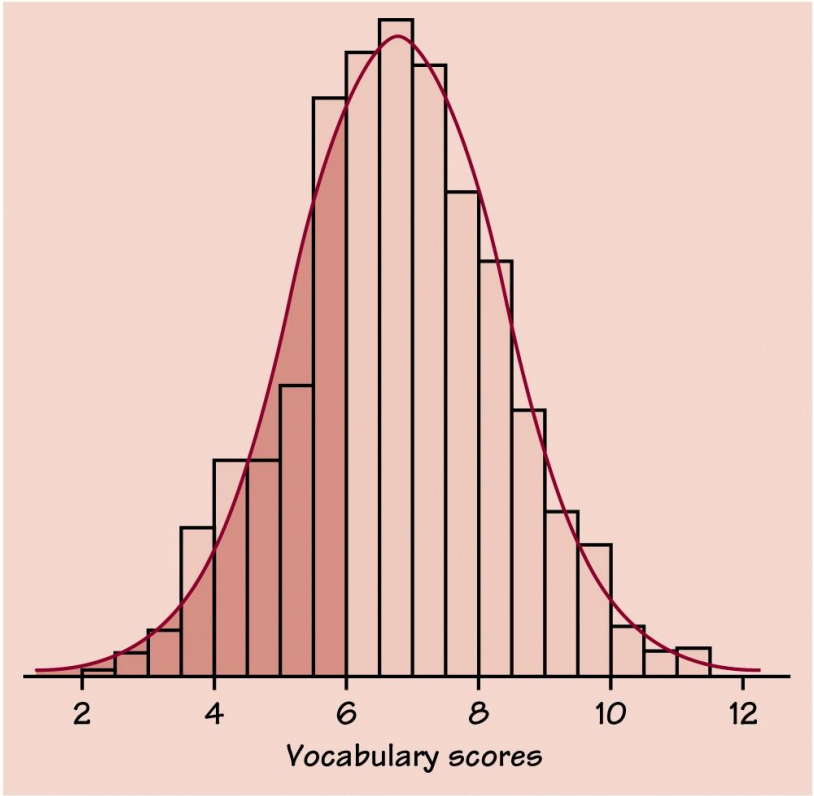
Histogram vs Density Curve

- The area of the shaded bars represent the students with vocabulary scores lower than 6.0.
- There are 287 such students, who make up the proportion 287/947 = 0.303.
- A score of 6.0 corresponds to about the 30th percentile.



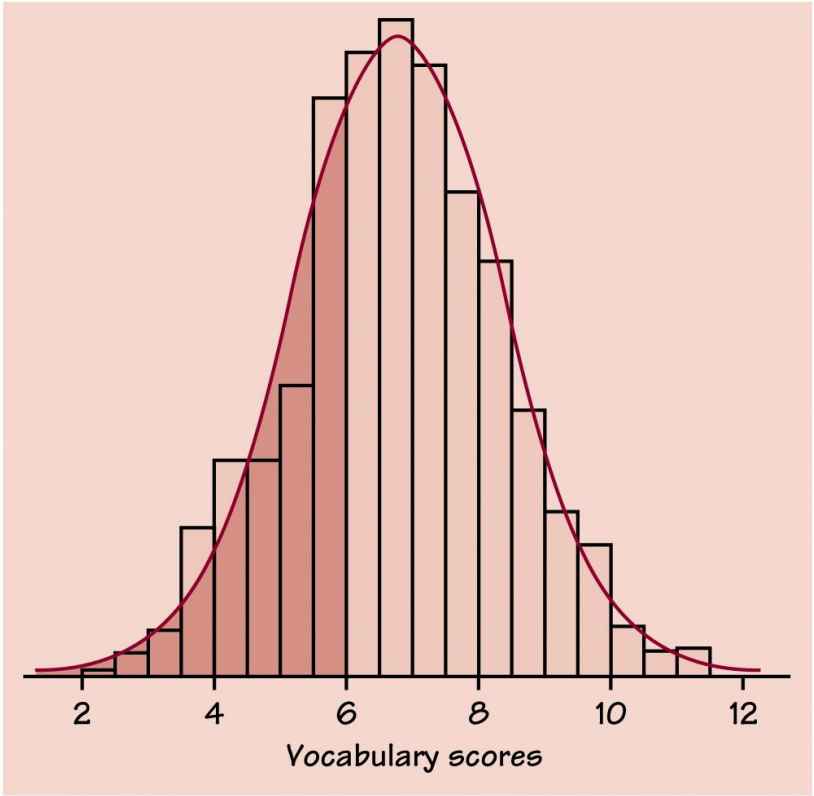
Histogram vs Density Curve

- Adjust the scale of the graph so that the total area under the curve is exactly 1. This area represents the proportion 1, that is, all observations.
- Areas under the curve then represent proportions of the observations.
- The curve is now a density curve.



Histogram vs Density Curve

- The shaded area under the density curve represents the proportion of students with scores lower than 6.0.
- This area is 0.293, only 0.010 away from the histogram result.
- This estimate corresponds to about the 29th percentile.
- You can see that this is a good approximation of the area given by the histogram.



The Center and Spread of a Density Curve

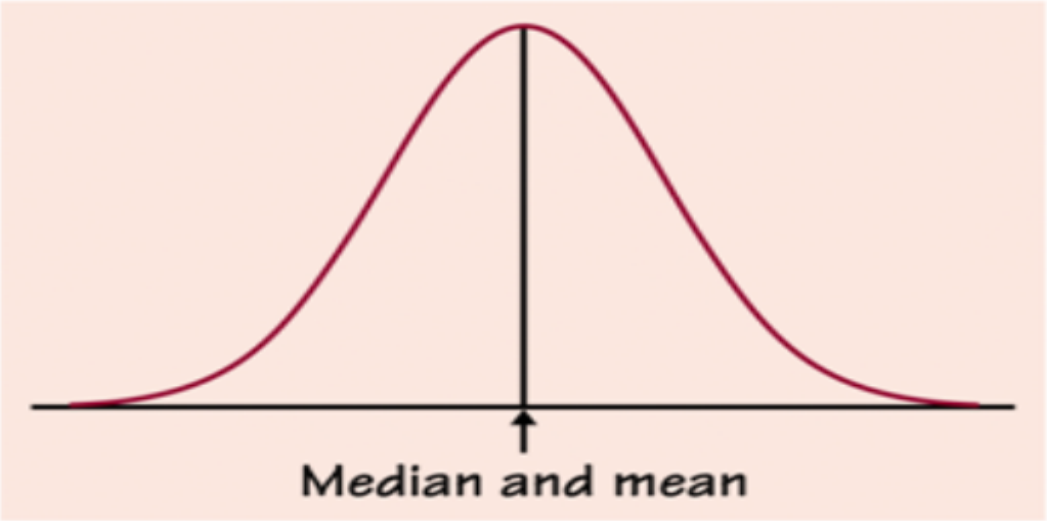
- The **median** is the point with half the observations on either side.
  - The median of a density curve is the **equal-areas point**
  - This is the point with half the area under the curve to its left and the other remaining half of the area to its right.
- The quartiles divide the area under the curve into quarters.
  - One-fourth of the area under the curve is to the left of the first quartile.
  - Three-fourths of the area is to the left of the third quartile.
- You can roughly locate the median and quartiles of any density curve by eye by dividing the area under the curve into four equal parts.

The Center and Spread of a Density Curve

- The **mean** of a density curve is the **balance point**.
  - It is the point at which the curve would balance if made of solid material.
- The mean and median of a symmetric density curve are equal. They both lie at the center of the curve.
- The mean of a skewed distribution is pulled toward the long tail.

The Center and Spread of a Density Curve

If the density curve is perfectly symmetric, the mean is equal to the median.

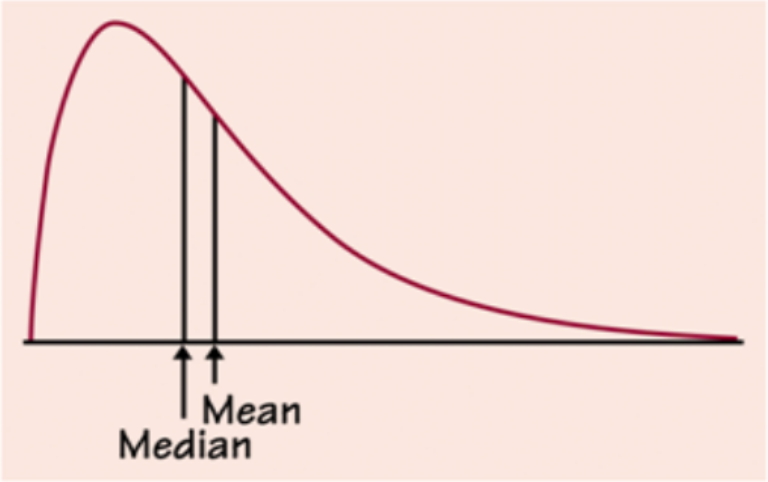


The Center and Spread of a Density Curve

The median divides the curves area in half.

The mean is the balance point; where the curve would balance if it's made of solid material.

The mean is pulled towards the skew or long tail.



New Symbols

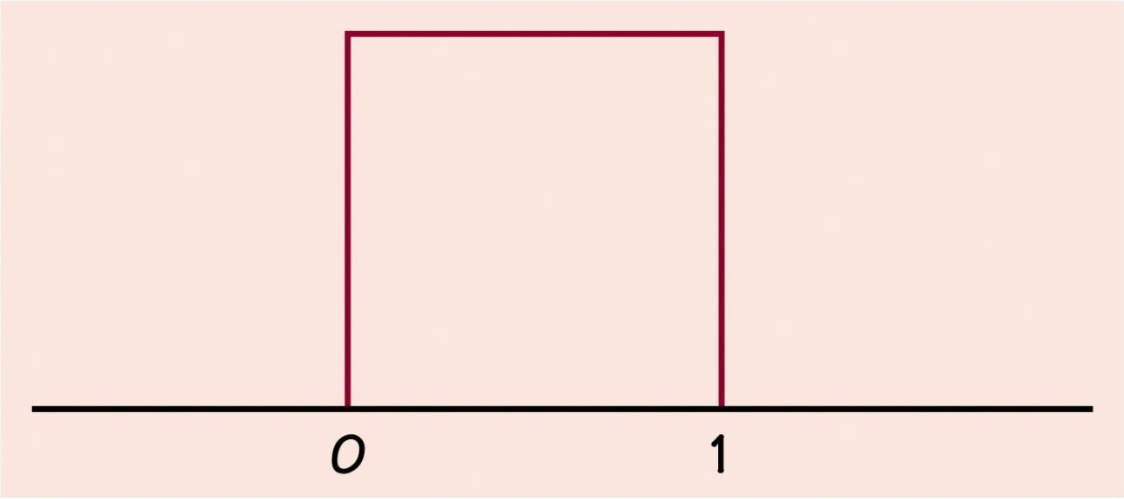
μ represents the mean of a density curve (pronounced ‘mu’)

S represents the sample population

σ represents the standard deviation of a density curve (pronounced ‘sigma’)

Example 1: A uniform Distribution

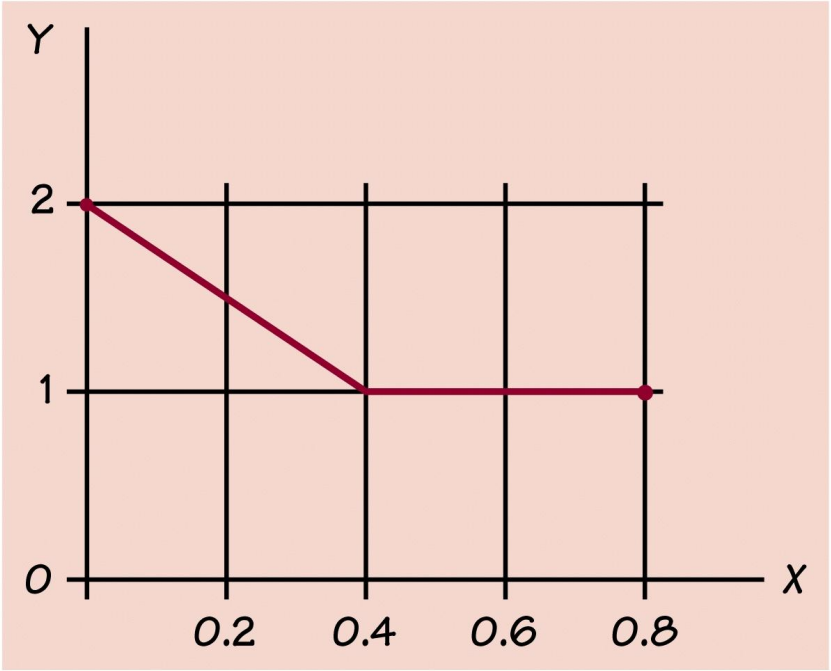
a. Why is the total area under this curve equal to 1?
b. What percent of the observations lie above 0.8?
c. What percent of the observations lie below 0.6?
d. What percent of the observations lie between 0.25 and 0.75?
e. What is the mean, μ of this distribution?



Example 2

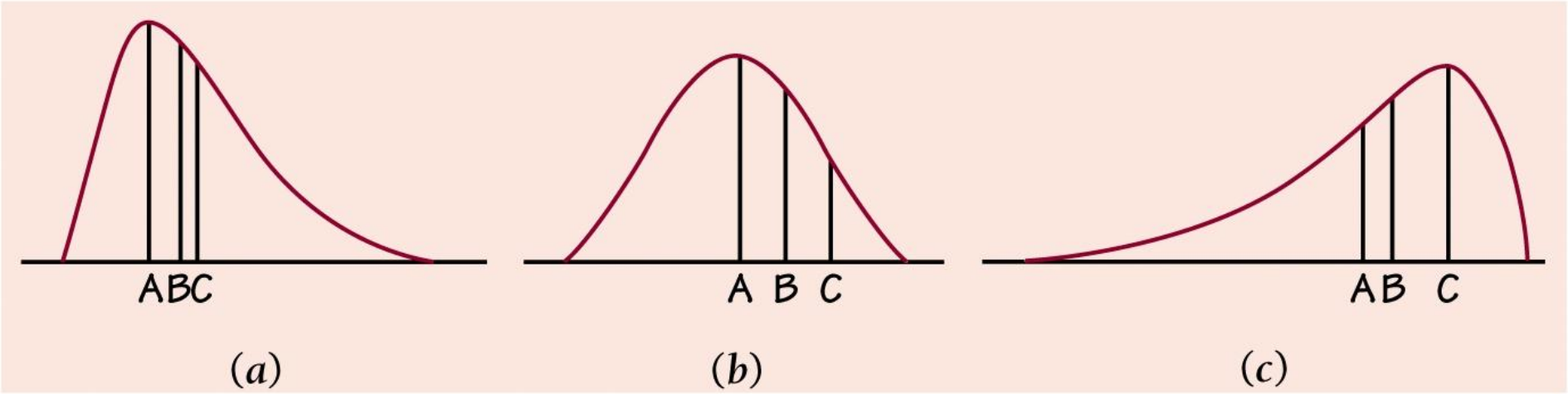
An unusual ‘broken line’ density curve. Use areas under this density curve to find the proportion of observations within the given interval.

a. $0.6 \le x \le 0.8$
b. $0 \le x \le 0.4$
c. $0 \le x \le 0.2$
d. The median of this density curve is a point between $x = 0.2$ and $x = 0.4$. Explain why.



Example 3

At which of these points on each curve do the mean and the median fall?



Example 4

Consider a uniform distribution that outputs numbers between 0 and 2.

a. Find P ($x < 0.5$)
b. Find P ($x > 0.7$)
c. Find P ($0.2 < x < 1.2$)