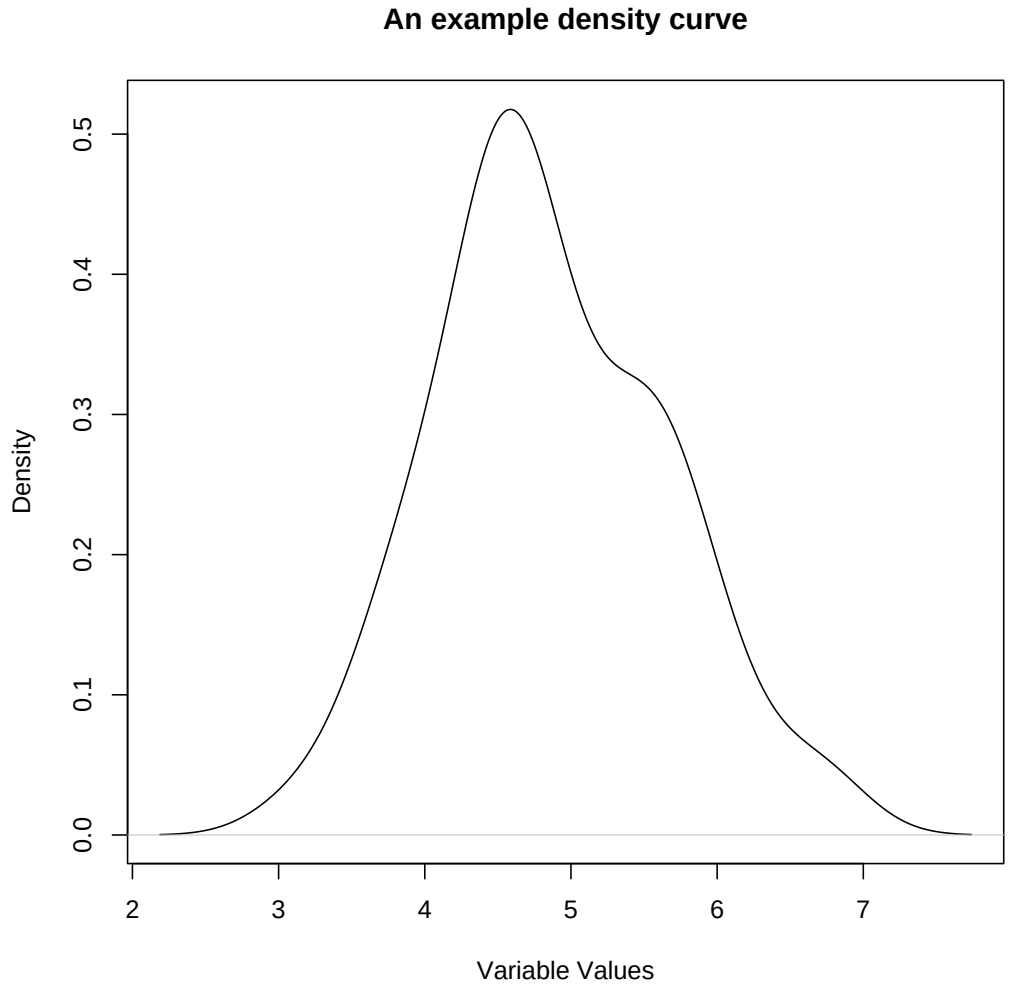
Density Curves and Normal Distributions

- **Recall: For data on a quantitative variable, the histogram gives a graphical picture of the *distribution*.**
- **Histogram will show us approximate shape, center, spread, and any outliers**
- **In addition, numerical measures (like 5-number summary) can describe the distribution.**
- ***Note*: A histogram is designed to summarize a *sample* of data.**

Density Curves

- **A *density curve* is a graphical picture of the *population distribution* of a variable.**
- **Resembles a histogram, but a histogram looks "blocky" (because of the bars)**
- **Density curve is usually a *smooth* curve.**
- **Density curve is like a smoothed-out, idealized version of a histogram.**

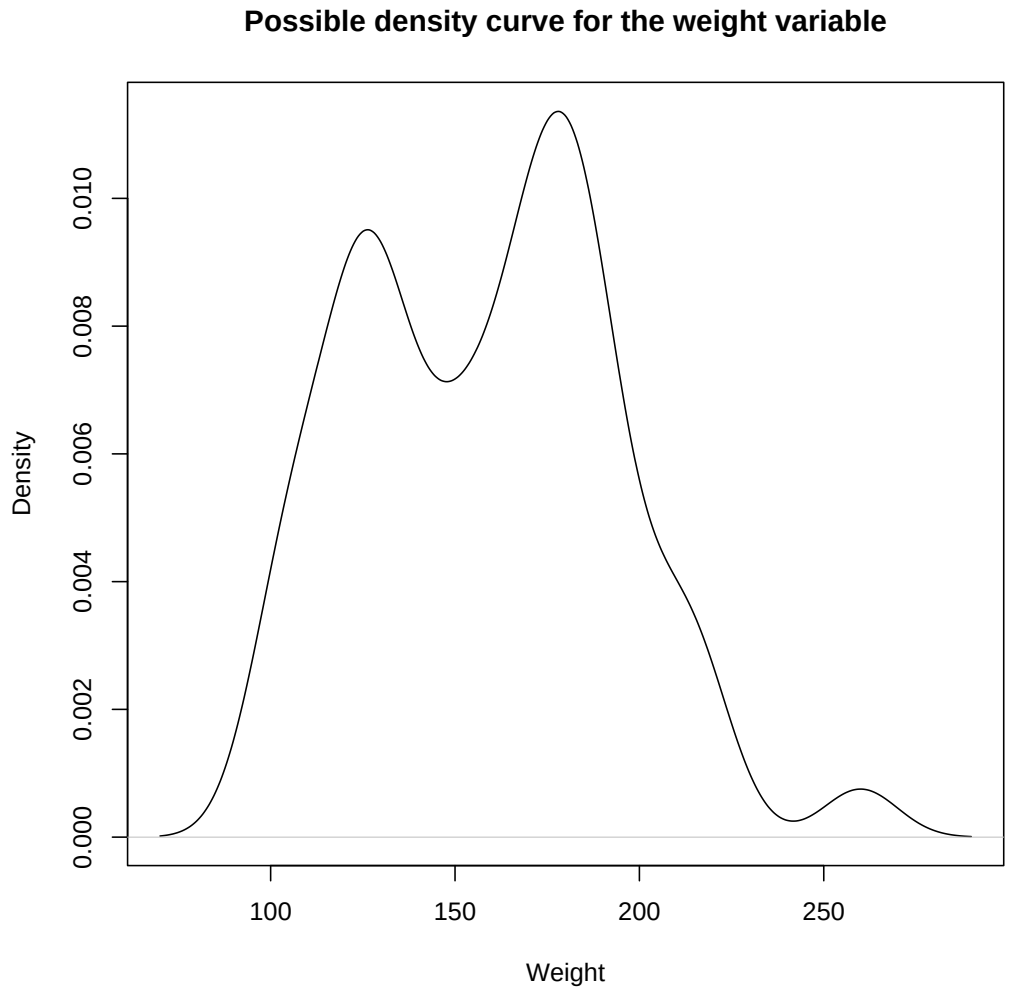
Example of a Density Curve:



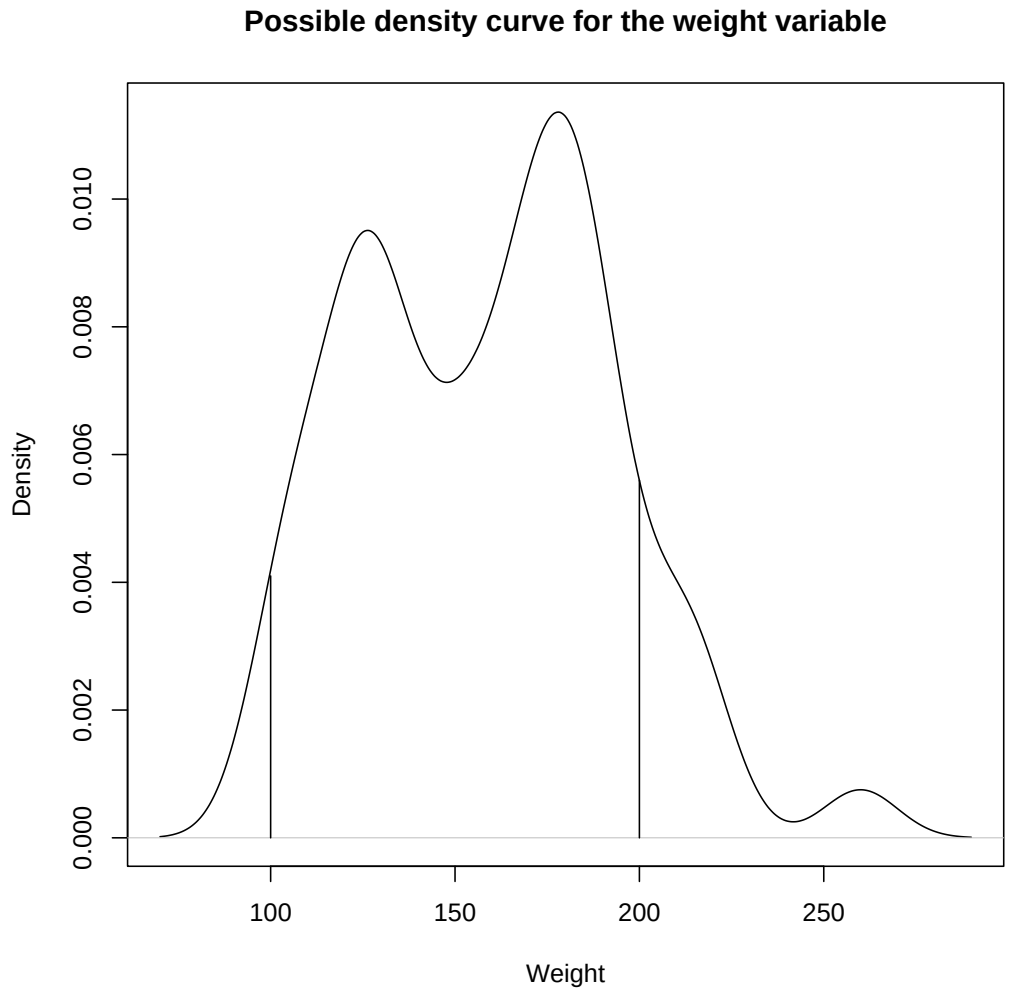
More on Density Curves

- **Histograms may plot counts or proportions on vertical axis**
- **Density curves always deal with *proportions of data*, but the plotting is done a bit differently.**
- **A proportion of data is represented by a certain *area* under the density curve.**
- **Total area under a density curve is always 1 (Total proportion = 1).**

Recall the Student Weight Variable:



What is the area under the curve between 100 and 200 pounds?



This area represents the proportion of all weights (in the population) that are between 100 and 200 pounds.

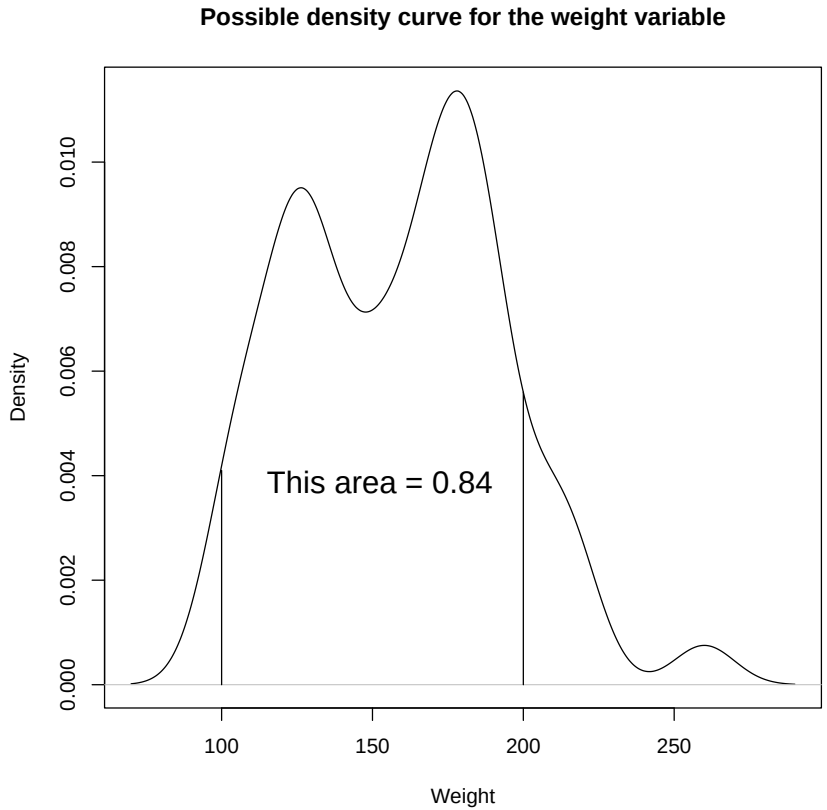


Figure 1: The area under the curve between 100 and 200 pounds is around 0.80 (it's 0.84, to be exact).

Shapes of Density Curves

- **Like histograms, density curves could be symmetric or skewed.**
- **A density curve is *symmetric* if the left and right sides of the density curve are approximately mirror images.**
- **A density curve is *skewed to the right* if the right side of the density curve extends much farther out than the left side ("long right tail")**
- **A density curve is *skewed to the left* if the left side of the density curve extends much farther out than the right side ("long left tail")**
- **A density curve is *unimodal* if it has only one prominent peak.**
- **A density curve is *bimodal* if it has two separated peaks.**

Clicker Quiz 1

What is the best description of the *shape* of the Student Weight density curve?

A. Symmetric and unimodal

B. Somewhat skewed to the left and bimodal

C. Somewhat skewed to the right and bimodal

D. Skewed and unimodal

Center and Spread of a Density Curve

- **The *median* of a density curve is the value with half the area to the left of it and half the area to the right. (“equal-areas point”)**
- **What is the area under the density curve to the *left* of the median?**
- **The *first quartile* of a density curve is the value with 0.25 area under the curve to the left of it.**
- **The *third quartile* of a density curve is the value with 0.75 area under the curve to the left of it.**
- **It’s easy to approximate the median / quartiles by eye: Divide the area under density curve into 4 equal parts.**

Center and Spread of a Density Curve (Continued)

- **The *mean* of a density curve is the "balancing point" of the density curve if it were solid.**
- **For a *symmetric* density curve, both the mean and the median will be the same (exactly at center)**
- **For a *skewed* density curve, the mean will be *farther out in the long tail* than the median.**
- **See example pictures:**

A symmetric density curve:

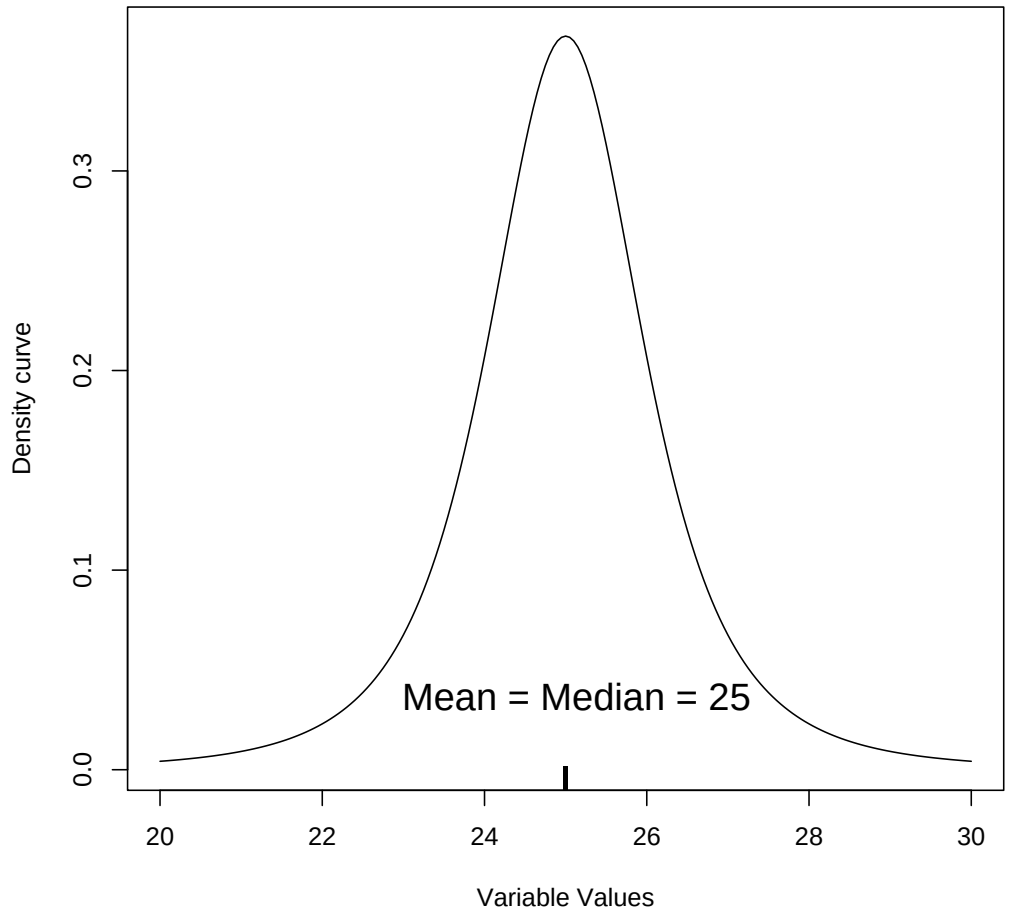


Figure 2: Mean and Median both at the same place, 25.

A skewed density curve:

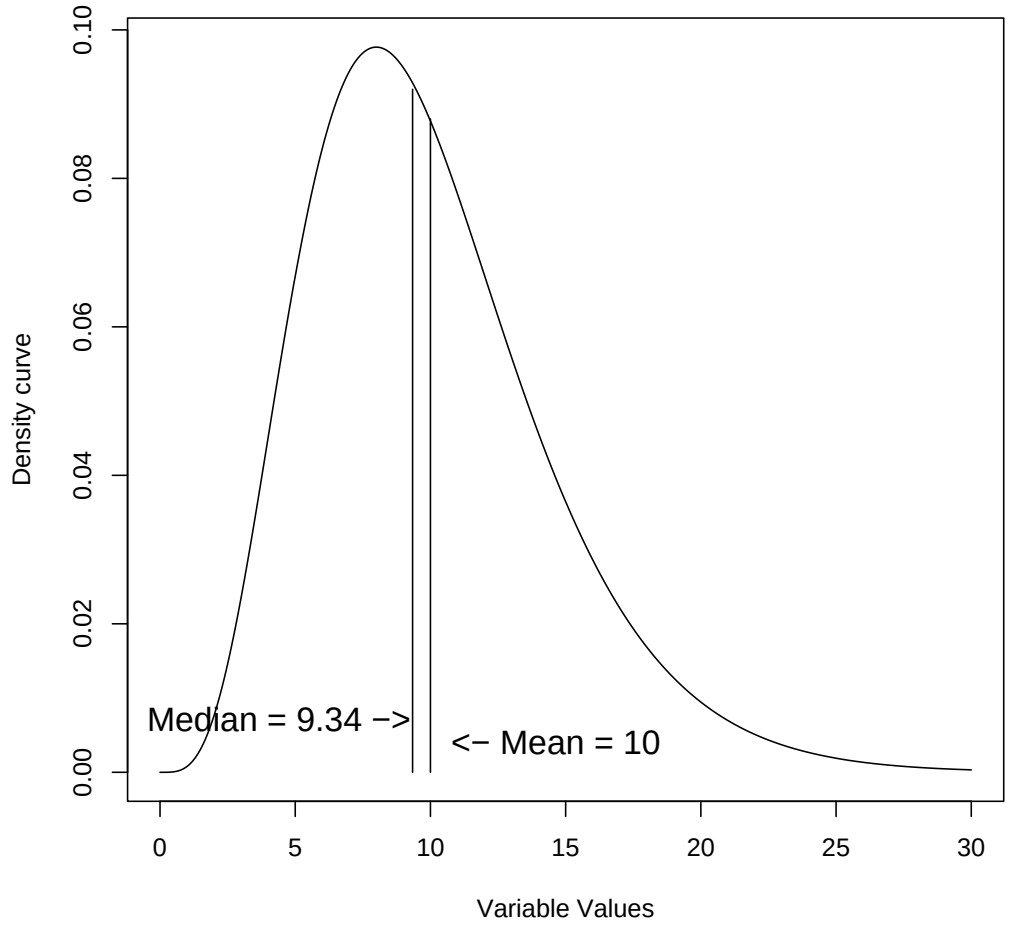


Figure 3: Mean (10) farther out toward the long tail than Median (9.34).

Clicker Quiz 2

If the density curve is *skewed to the left*, what are possible values of the mean and median?

A. Mean = 70, Median = 70

B. Mean = 65, Median = 70

C. Mean = 70, Median = 65

Normal Distributions

- **Normal density curves are a specific class of density curves.**
- **They are *symmetric*, *unimodal*, and *bell-shaped*.**
- **Many real data sets have distributions that are approximately normal.**
- **They are also useful in certain statistical methods. (more later)**

An example normal density curve:

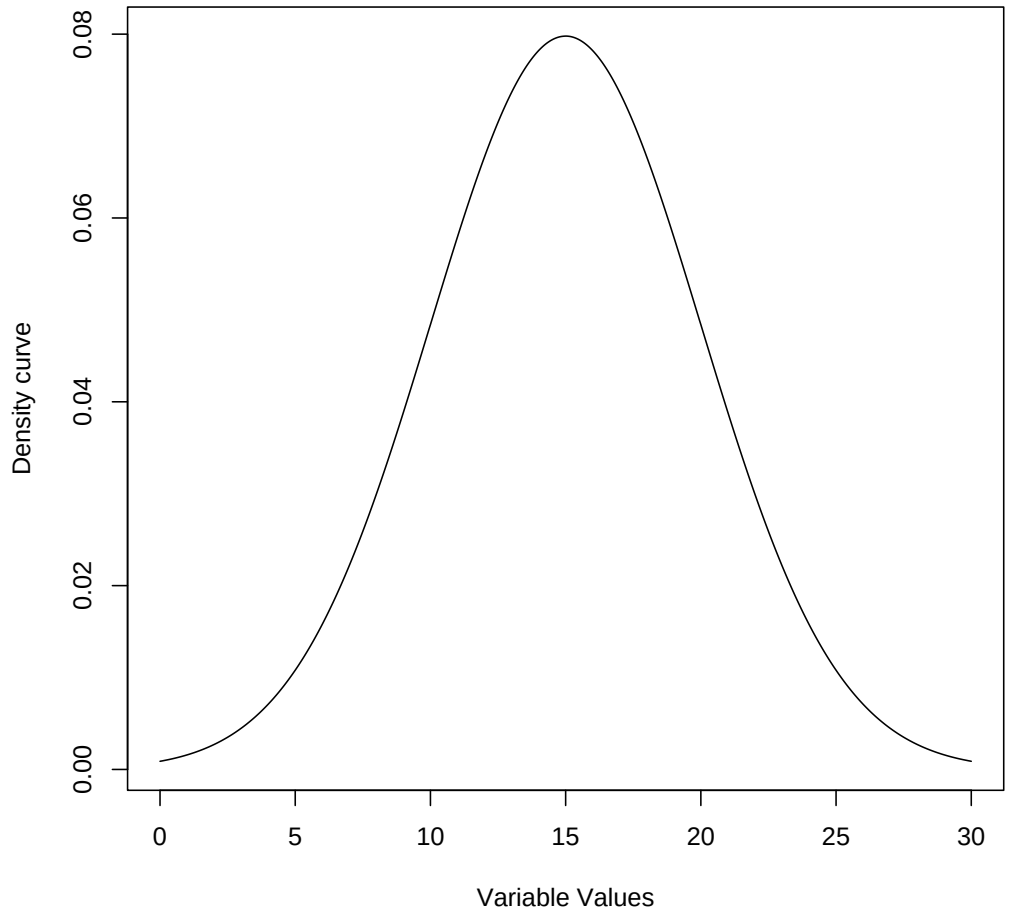


Figure 4: A normal density curve with mean 15.

Mean and Standard deviation of Normal Curve

- **We can completely characterize a normal distribution by giving its mean and standard deviation.**
- **The mean describes the center of the distribution.**
- **The standard deviation describes the spread of the distribution.**
- **For a normal density curve, the *standard deviation* is the *distance* between the center and the "inflection point"**
- ***Inflection point* = the place where the curve changes from "opening down" to "opening up"**
- **What is the standard deviation of the previously shown normal density curve?**

An example normal density curve:

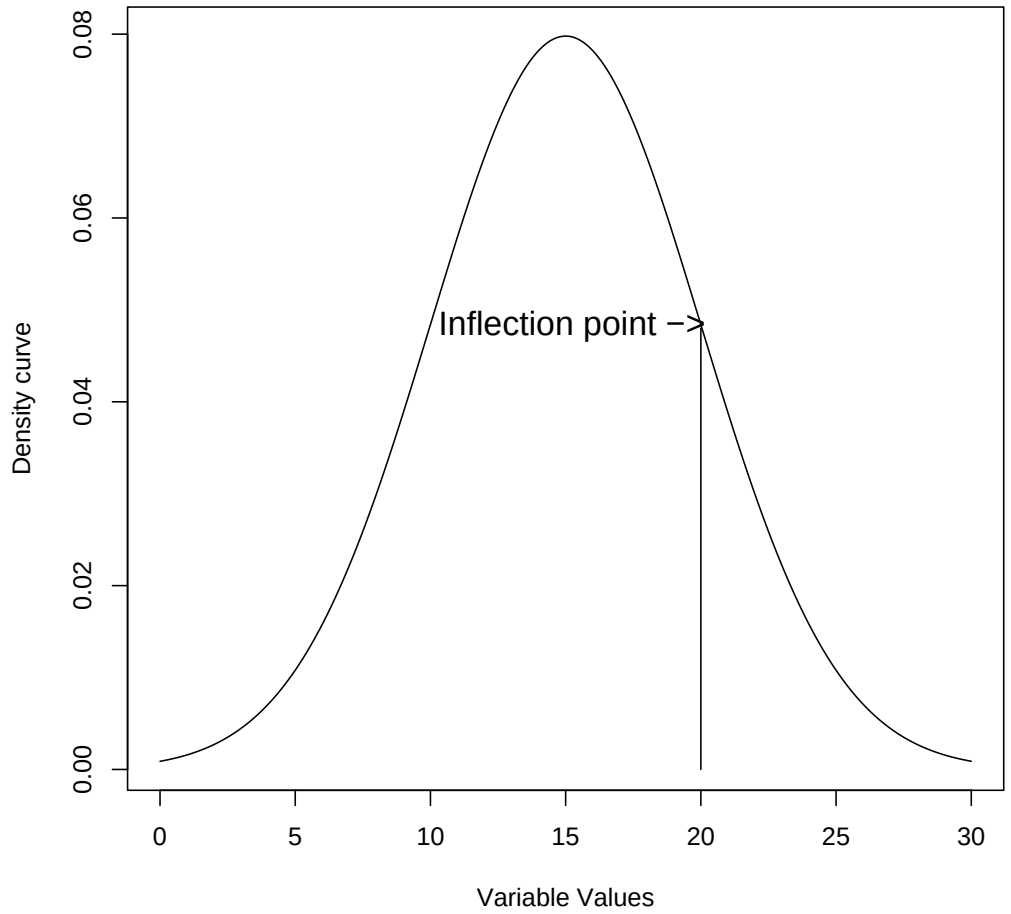


Figure 5: A normal density curve with mean 15 and standard deviation 5.

Clicker Quiz 3

What is the mean and standard deviation of the following normal curve?

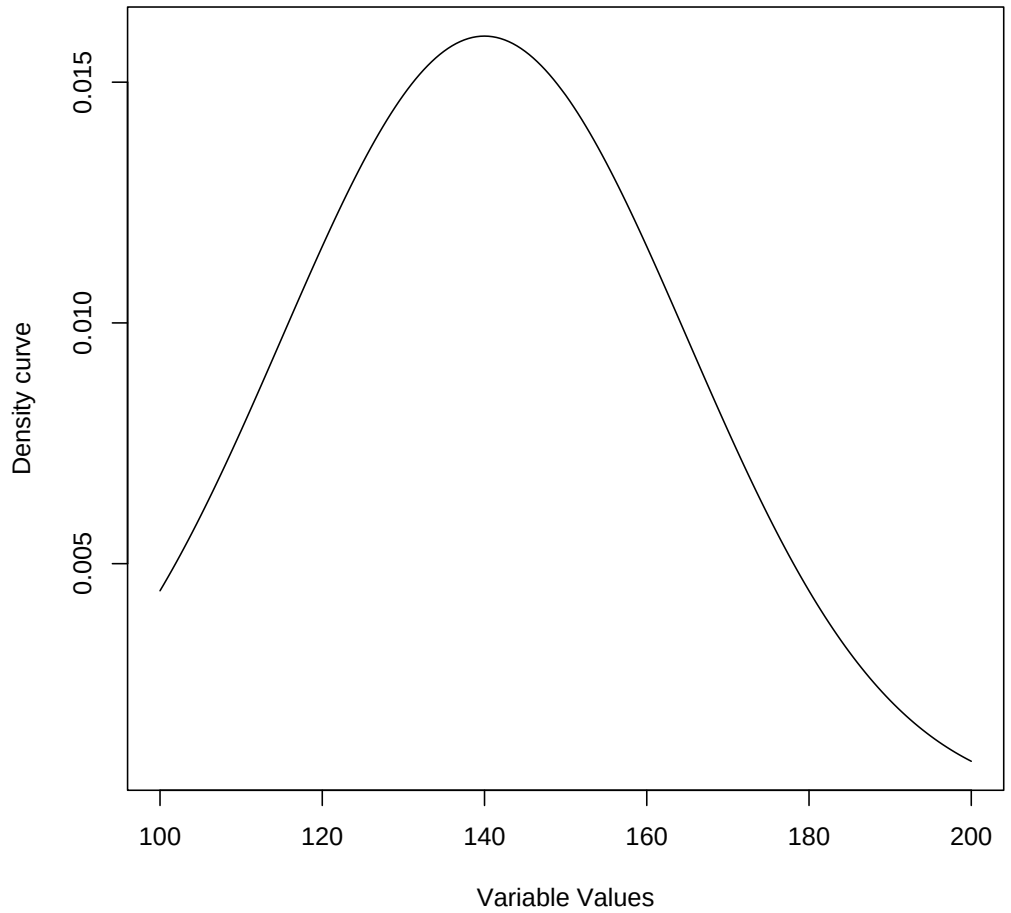
A. Mean = 160, sd = 40

B. Mean = 145, sd = 60

C. Mean = 100, sd = 35

D. Mean = 140, sd = 25

An example normal density curve (shown partially):



Usefulness of Normal Curves

- **Some distributions for real data may be well approximated by normal distributions.**
- ***Example 1:*** **Replicated measurements of the same quantity may follow a normal distribution. (Example: Gauss & astronomy data)**
- ***Example 2:*** **Measurements on many biological and psychological variables may follow a normal distribution. (Example: Galton & human forearm and height data)**

Usefulness of Normal Curves (Continued)

- **In addition, the sampling distributions of *sample proportions* and *sample means* are approximately normal distributions, when the sample size is large.**
- ***Caution:*** **Some data (examples: income data, house price data) tend to be skewed and therefore not normal.**

The 68-95-99.7 Rule

- **This rule is also known as the *Empirical Rule*.**
- **For a normal distribution, 68% of the data fall within *one* standard deviation of the mean.**
- **For a normal distribution, 95% of the data fall within *two* standard deviations of the mean.**
- **For a normal distribution, 99.7% (almost all!) of the data fall within *three* standard deviations of the mean.**
- **Why? Think about the *area under the normal curve* between the values *one sd below* and *one sd above* the mean. (Recall examples)**
- **Many data sets are only *approximately* normal, so these percentages would be *approximately correct* for such data.**

Clicker Quiz 4

Suppose newborn babies' weights follow an approximately normal distribution with mean 3.5 kg and standard deviation 0.5 kg. About what percentage of babies will have weights between 2.5 and 4.5 kg?

A. 95%

B. 99.7%

C. 34%

D. 68%

Clicker Quiz 5

Suppose newborn babies' weights follow an approximately normal distribution with mean 3.5 kg and standard deviation 0.5 kg. For this normal density curve, what is the area between 3.5 and 4.0?

A. 0.95

B. 0.34

C. 0.68

D. 0.475

Standard Scores

- **For an observation from a normal distribution, its *standard score* is found by subtracting off the mean *and then* dividing by the standard deviation of the distribution.**
- **Standard score = (observation - mean) / sd**
- **This tells us how many standard deviations *above the mean* or *below the mean* an observation is.**
- **This allows us to compare values coming from different distributions.**

Standard Scores (Example)

- ***Example:*** **Suppose SAT Math scores are normally distributed with a mean of 500 and a standard deviation of 100.**
- **Suppose ACT Math scores are normally distributed with a mean of 18 and a standard deviation of 6.**
- **Sarah got a 550 on the SAT math exam.**
- **Bill got a 15 and Julie got a 24 on the ACT math exam.**

Standard Scores (Example continued)

- **Sarah's standard score was (550 - 500) / 100 = 0.5, so Sarah was 0.5 standard deviations *above* the mean.**
- **Bill's standard score was (15 - 18) / 6 = -0.5, so Bill was 0.5 standard deviations *below* the mean.**
- **Julie's standard score was (24 - 18) / 6 = 1.0, so Julie was 1 standard deviation *above* the mean.**

Clicker Quiz 6

Based on their standard scores, who did best among Bill, Sarah, and Julie, relative to their own testing group?

A. Bill best, Sarah next, Julie worst

B. Julie best, Bill and Sarah tied for worst

C. Julie best, Sarah next, Bill worst

D. Sarah best, Julie next, Bill worst

Percentiles

- **A *percentile* of a distribution is a value such that a given percentage of the data lies below that value.**
- **The median is the 50th percentile of a distribution, because 50% of the data lie below the median.**
- **The first quartile is the 25th percentile of a distribution; the third quartile is the 75th percentile of a distribution.**
- **In the ACT distribution, we know Julie's score was at the 84th percentile (84% of test-takers did worse than Julie)**
- **Draw the normal density to see why!**
- **Can we approximate the percentiles for Sarah's score and Bill's score?**