NUMERICAL SUMMARIES OF DATA:

How do you measure the center of the data?

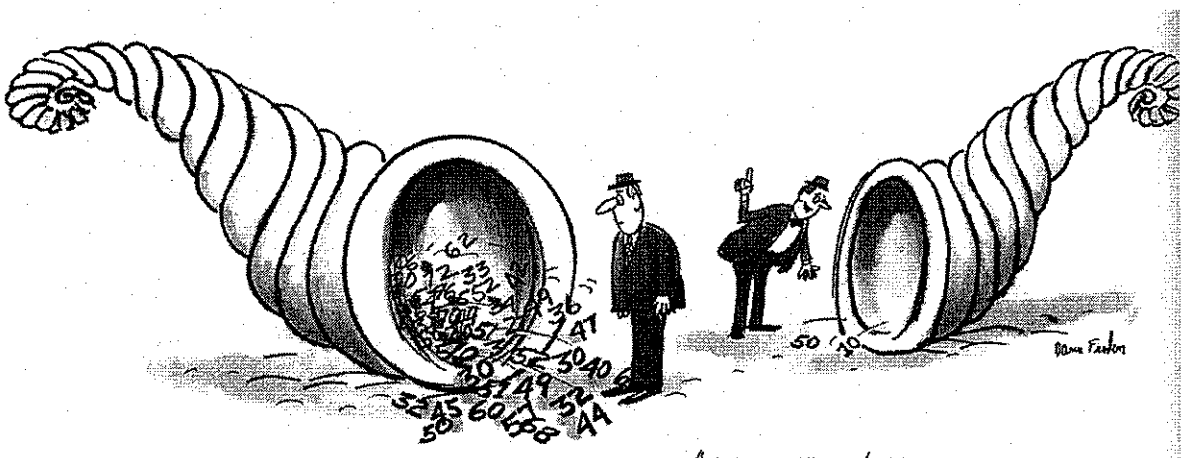
Average

Median

How do you measure the overall size of a list of numbers?

How do you measure the spread in the data?

Chapter 4: The Average and the SD



"LOOK, FRED! THIS SEEMS TO BE THE SAME THING, SUMMARIZED."

Summarizing Data

There are a number of aspects of a list of numbers we may want to summarize, including:

The middle or center of the data, and

The amount of spread in the data.

The Average

The Average is a measure of the center or middle of the data.

$$\text{Average} = \frac{\text{Sum of Data Values}}{\text{Number of Data Values}}$$

Example: Find the average age of the first 5 Presidents

President	Age
Washington	67
Adams	90
Jefferson	83
Madison	85
Monroe	73

$$\text{Average age} = \frac{67 + 90 + 83 + 85 + 73}{5}$$

$$= 79.6 \text{ years}$$

The Average

The average is the most commonly used summary statistic for analyzing experimental and observational data.

Example: The National Health and Nutrition Examination Survey (NHANES) is a representative survey of Americans carried out by the Public Health Service.

NHANES I was conducted between 1971 and 1975 and involved approximately 28,000 people.

NHANES II was conducted between 1976 and 1980 and involved approximately 28,000 people.

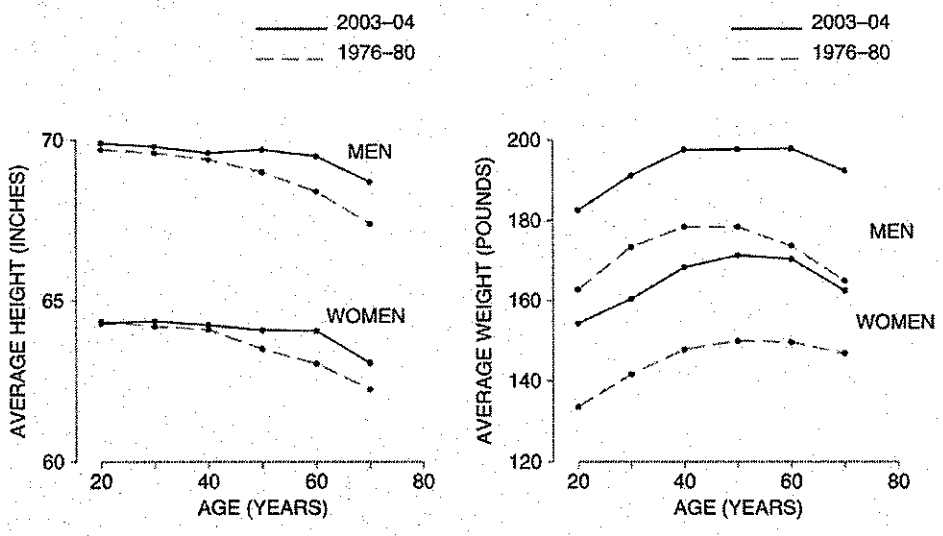
The Hispanic Health and Nutrition Examination Survey (HHANES) was conducted between 1982 and 1984 and involved about 16,000 people.

NHANES III was conducted between 1988 and 1994 and involved about 40,000 people.

NHANES included data on:

- Demographic variables, like age, education, and income
- Physiologic variables, like height, weight, blood pressure, and serum cholesterol levels
- Dietary habits
- Levels of lead and pesticides in blood
- Prevalence of various diseases.

Age-specific Heights and Weights for Men and Women Aged 18—74 in NHANES



For the 1976-80 data, do men and women decrease in height by more than 2 inches as they go from about 20 to 70 years of age?!?

Cross-sectional and Longitudinal Studies

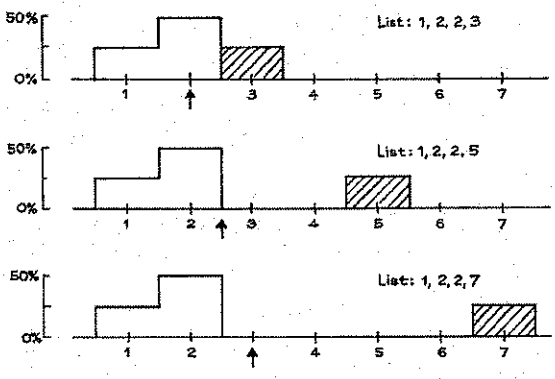
NHANES is a **cross-sectional** study because it consists of a cross-section of the U.S. population at some time point.

A **longitudinal** or **cohort** study is one in which we follow a group of people over time. For example,

- The Framingham Heart Study in Framingham, Massachusetts.
- The Cache County Study of Memory and Aging.

If you want to draw conclusions about what happens over time, you must have a longitudinal study!

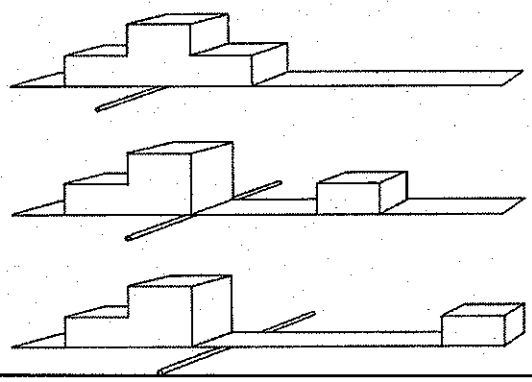
The average is the balance point of the histogram.



ave = 2.0, median = 2

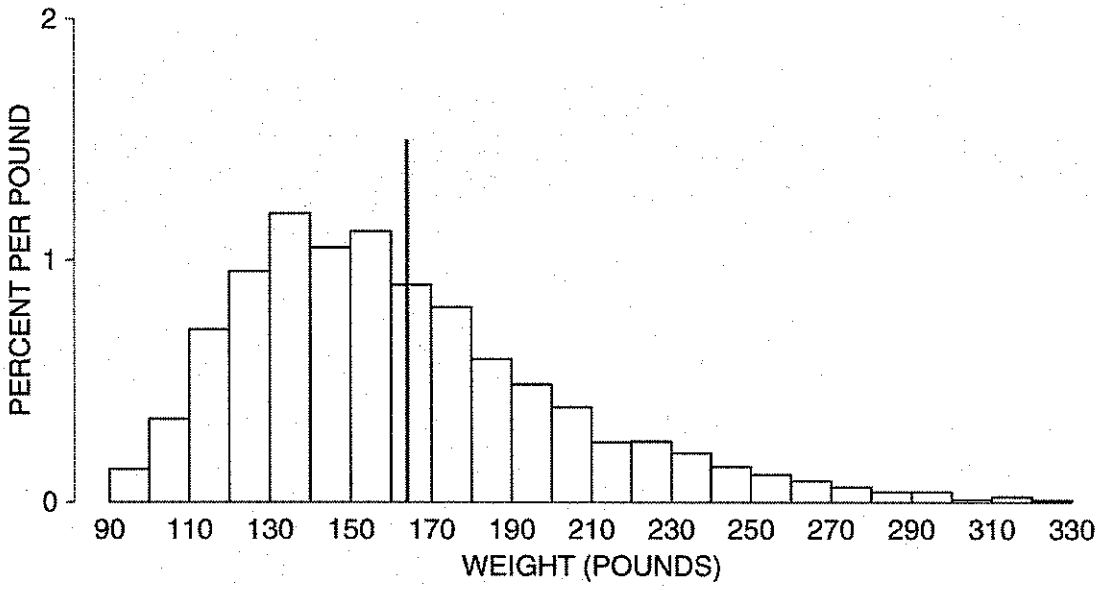
ave = 2.5, median = 2

ave = 3.0, median = 2



The average is the balance point of the histogram.

Figure 4. Histogram for the weights of the 2,696 women in the HANES5 sample. The average is marked by a vertical line. Only 41% of the women were above average in weight.

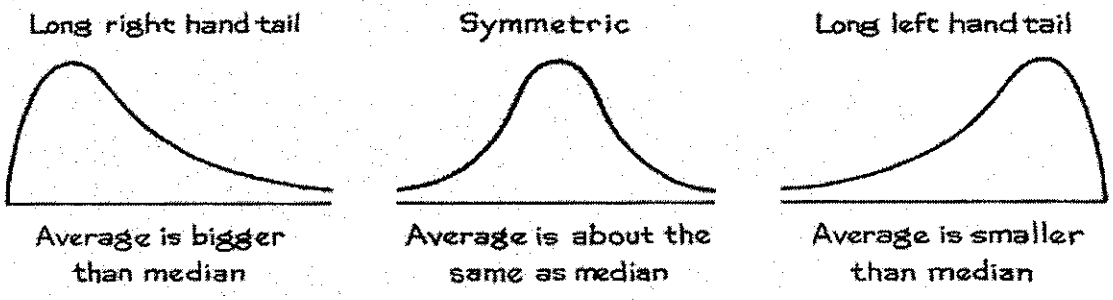


The Median

Half the data values in a list are less than the median and half the data values are greater than the median.

Half the area of a histogram is to the left of the median, and half the area is to the right.

For histograms that are not symmetric, statisticians sometimes prefer to use the **median** to measure the center of the data instead of the average.



The Median

For a list of data values,

if the number of values is ODD, the **median** is the **middle number** when they are arranged from smallest to largest

if the number of values is EVEN, the **median** is the or the **average of the two middle numbers**

Example: Median age of Presidents

President	Age
Washington	67
Adams	90
Jefferson	83
Madison	85
Monroe	73

Age of first 3 Presidents:

- 67, 83, 90

median is ?

Age of first 4 Presidents:

- 67, 83, 85, 90

median is ?

Income, Housing Costs, ooo

The Standard Deviation (SD)

The standard deviation measures the **spread** of the data, or the amount of **variability** in the data values.

For example, the list

50, 52, 49, 50, 51, 48

has a much smaller SD than the list

50, 65, 48, 37, 59, 41

The Standard Deviation (SD)

The standard deviation is the r.m.s. of the **deviations** of the data values from the average.

What is the r.m.s.?

The Root-Mean-Square (r.m.s.)

The r.m.s. is a measure of the average **magnitude** of a list of numbers that can take on both positive and negative values.

To calculate the r.m.s. we

- SQUARE all the entries in the list, getting rid of the negative signs.
- Take the MEAN (average) of the squared values.
- Take the square ROOT of the average of the squared values.

An example: List = 6, 7, -9, -7, 8, -5. Average = 0.
The r.m.s. of this list is calculated as follows:

- $6^2 = 36$, $7^2 = 49$, $(-9)^2 = 81$, $(-7)^2 = 49$, $8^2 = 64$, $(-5)^2 = 25$
- $(36 + 49 + 81 + 49 + 64 + 25) \div 6 = 50.67$
- $\sqrt{50.67} \approx 7.12$ = r.m.s.

The Standard Deviation (SD)

The standard deviation is the r.m.s. of the **deviations** of the data values from the average.

Example: List of numbers = 8, 3, 1, 1, 4, 6, 5

Data value	Deviation	Squared deviation
8	$8 - 4 = 4$	$4^2 = 16$
3	$3 - 4 = -1$	$(-1)^2 = 1$
1	$1 - 4 = -3$	$(-3)^2 = 9$
1	$1 - 4 = -3$	$(-3)^2 = 9$
4	$4 - 4 = 0$	$0^2 = 0$
6	$6 - 4 = 2$	$2^2 = 4$
5	$5 - 4 = 1$	$1^2 = 1$
ave = 4	ave = 0	ave = 40/7 = 5.71

SD = $\sqrt{5.71} = 2.39$

Calculating the SD

Example 2: Ages at death of the first 5 U.S. presidents:

President	Age	Deviation	Squared Deviation
Washington	67	67 – 79.6 = -12.6	$(-12.6)^2 = 158.76$
Adams	90	90 – 79.6 = 10.4	$10.4^2 = 108.16$
Jefferson	83	83 – 79.6 = 3.4	$3.4^2 = 11.56$
Madison	85	85 – 79.6 = 5.4	$5.4^2 = 29.16$
Monroe	73	73 – 79.6 = -6.6	$(-6.6)^2 = 43.56$
Average	79.6	0.0	70.24

SD = $\sqrt{70.24}$ = **8.38**

Standard Units

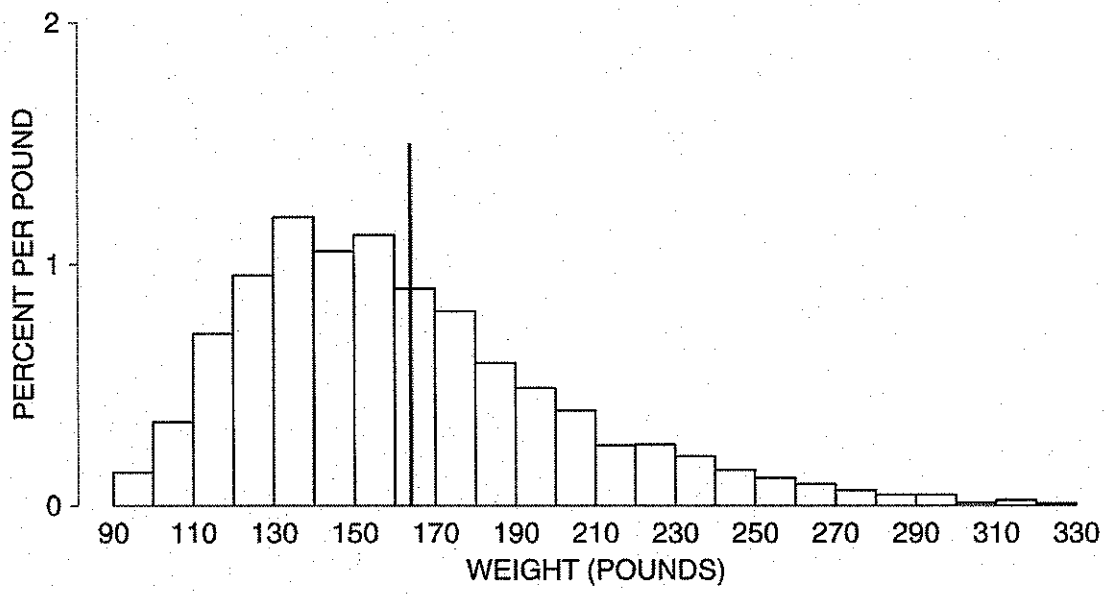
Which is better: A score of 650 on the SAT or a score of 28 on the ACT?

SAT scores have a mean of 500 and a standard deviation of 100.

ACT scores have a mean of 21 and a standard deviation of 5.

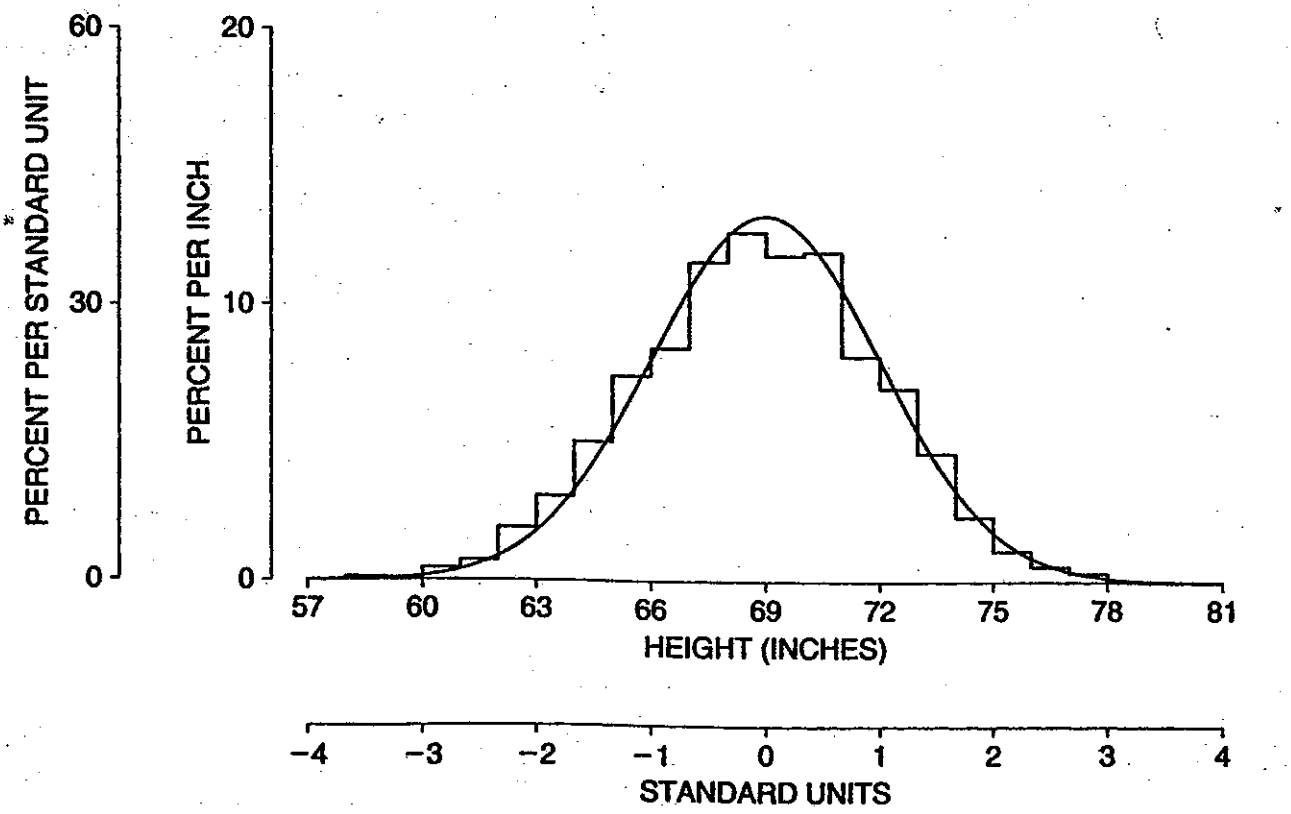
The average is the balance point of the histogram.

Figure 4. Histogram for the weights of the 2,696 women in the HANES5 sample. The average is marked by a vertical line. Only 41% of the women were above average in weight.



What happens if we combine the data for women + men?

HEIGHTS OF MEN



AV = 69

SD = 3

Interpreting the Standard Deviation

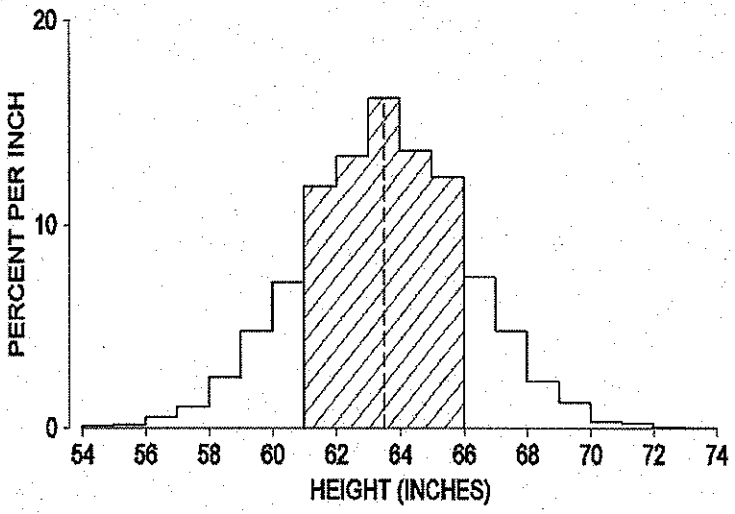
The standard deviation is a measure of the **spread** of the data.

For many "bell-shaped" histograms

- Approximately 68% of the data values lie within 1 standard deviation of the average.
- Approximately 95% of the data values lie within 2 standard deviations of the average.

Example: The histogram of heights of 6,588 women aged 18-74 in NHANES II.

Average = 63.5"
SD = 2.5"



Average – 1 SD = 61″
Average + 1 SD = 66″
} 67% of the women were between 61″ and 66″ tall (shaded)

Average – 2 SDs = 58.5″
Average + 2 SDs = 68.5″
} 94% of the women were between 58.5″ and 68.5″ tall

Example 1.

Heights of 200 men have an average of 69.5″ and an SD of 3″.

Approximately what percentage of the men are between 66.5″ and 72.5″ tall? How many men is this?

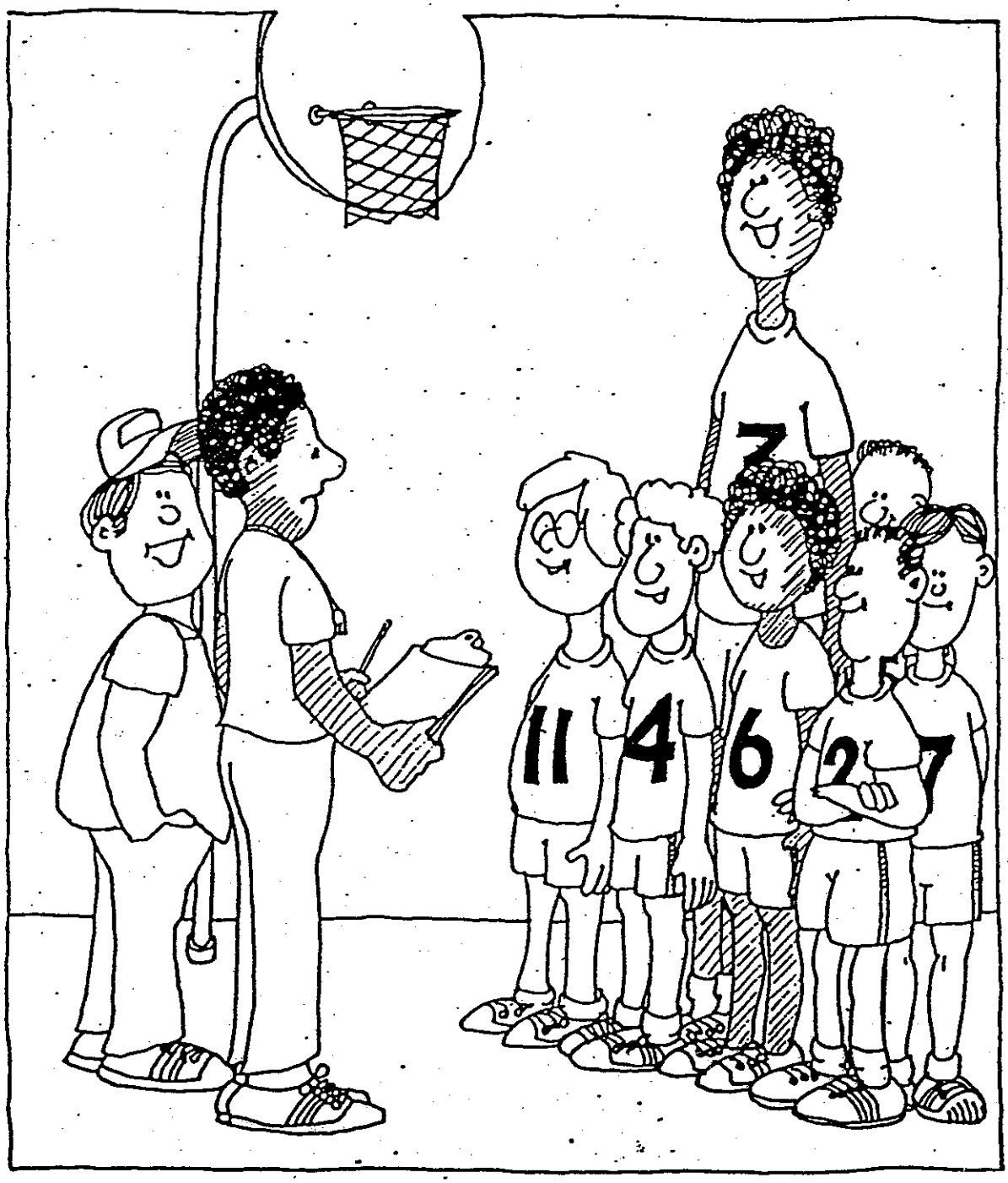
Approximately what percentage of the men are between 63.5″ and 75.5″ tall? How many men is this?

Facts about averages, medians and SDs

- Units for the average, median and SD are the same as for the data values.

- Adding the same number to each data value adds that same number to the average and the median, but THE SD DOES NOT CHANGE!

- Multiplying each data value by a positive number multiplies the average, the median, and the SD by that number.



"Should we scare the opposition by announcing our mean height or lull them by announcing our median height?"