Lecture 5: Chapter 4, Sections 3-4 Quantitative Variables (Summaries, Begin Normal)

- Mean vs. Median
- Standard Deviation
- Normally Shaped Distributions
- 68-95-99.7 Rule for Normal Distributions
- Normal Histogram Approximated by Curve
- *z*-scores

Looking Back: *Review*

- **4 Stages of Statistics**
  - Data Production (discussed in Lectures 1-3)
  - Displaying and Summarizing
    - Single variables: 1 categorical, 1 quantitative
    - Relationships between 2 variables
  - Probability
  - Statistical Inference

Ways to Measure Center and Spread

- Five Number Summary *(already discussed)*
- Mean and Standard Deviation

Definition

- **Mean**: the arithmetic average of values. For *n* sampled values, the mean is called "x-bar":

$$\bar{x} = \frac{x_1 + \cdots + x_n}{n}$$

- *The mean of a population, to be discussed later, is denoted "* μ *" and called "mu".*

Example: *Calculating the Mean*

- **Background**: Credits taken by 14 "other" students:

  4 7 11 11 11 13 13 14 14 15 17 17 17 18

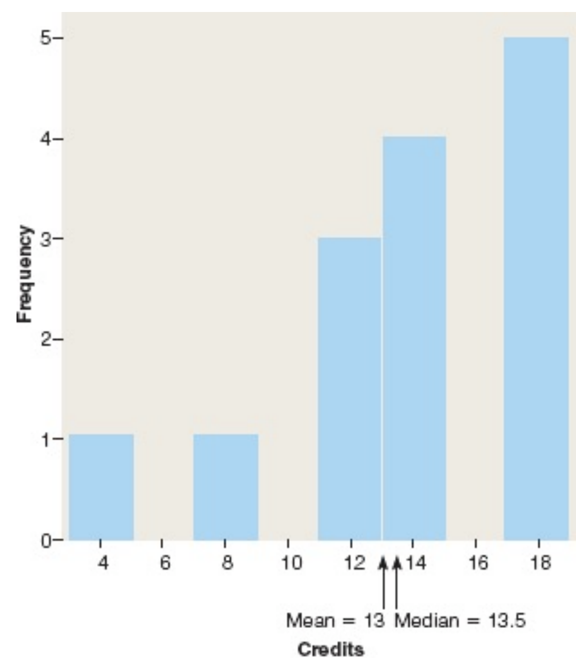
- **Question:** How do we find the mean number of credits?

- **Response:**

Example: *Mean vs. Median (Skewed Left)*

- **Background**: Credits taken by 14 "other" students:

  4 7 11 11 11 13 13 14 14 15 17 17 17 18
- **Question:** Why is the mean (13) less than the median (13.5)?
- **Response:**



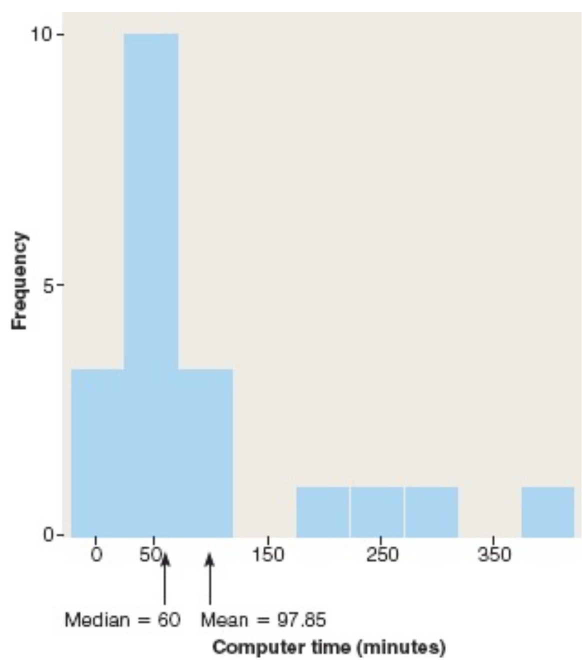
Example: *Mean vs. Median (Skewed Right)*

- **Background**: Output for students' computer times:

0 10 20 30 30 30 30 45 45 60 60 60 67 90 100 120 200 240 300 420

Variable	N	Mean	Median	TrMean	StDev	SE Mean
computer	20	97.9	60.0	85.4	109.7	24.5

- **Question:** Why is the mean (97.9) more than the median (60)?
- **Response:**



Practice: 4.30b p.104

Role of Shape in Mean vs. Median

- **Symmetric:**

  mean approximately equals median

- **Skewed left / low outliers:**

  mean less than median

- **Skewed right / high outliers:**

  mean greater than median

Mean vs. Median as Summary of Center

- **Pronounced skewness / outliers→**
  Report median.
- **Otherwise, in general→**
  Report mean (contains more information).

Ways to Measure Center and Spread

- Five Number Summary
- Mean and Standard Deviation

Definition

- **Standard deviation**: square root of "average" squared distance from mean $\bar{x}$. For *n* sampled values the standard deviation is

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{\boxed{n-1}}}$$

Looking Ahead: *Ultimately, squared deviation from a sample is used as estimate for squared deviation for the population. It does a better job as an estimate if we divide by n-1 instead of n.*

s refers to sample,
σ (sigma) refers to population

Interpreting Mean and Standard Deviation

- **Mean:** typical value
- **Standard deviation:** typical distance of values from their mean

(Having a feel for how standard deviation measures spread is much more important than being able to calculate it by hand.)

Example: *Guessing Standard Deviation*

- **Background:** Household size in U.S. has mean approximately 2.5 people.
- **Question:** Which is the standard deviation?

  (a) 0.014 (b) 0.14 (c) 1.4 (d) 14.0
- **Response:**

  Sizes vary; they differ from ____ by about ____

Practice: 4.36d p.106

Example: *Standard Deviations from Mean*

- **Background:** Household size in U.S. has mean 2.5 people, standard deviation 1.4.
- **Question:** About how many standard deviations above the mean is a household with 4 people?
- **Response:**
  - ___________________
  - ___________
  - ____

Looking Ahead: *For performing inference, it will be useful to identify how many standard deviations a value is below or above the mean, a process known as "standardizing".*

Practice: 4.38b p.107

Example: *Estimating Standard Deviation*

- **Background:** Consider ages of students…
- **Question:** Guess the standard deviation of…

1. Ages of all students in a high school (mean about 16)
2. Ages of high school seniors (mean about 18)
3. Ages of all students at a university (mean about 20.5)

- **Responses:**

1. standard deviation _______________
2. standard deviation _______________
3. standard deviation _______________

Looking Back: *What distinguishes this style of question from an earlier one that asked us to choose the most reasonable standard deviation for household size? Which type of question is more challenging?*

Example: *Calculating a Standard Deviation*

- **Background**: Hts (in inches) 64, 66, 67, 67, 68, 70 have mean 67.
- **Question:** What is their standard deviation?
- **Response:** Standard deviation *s* is sq. root of "average" squared deviation from mean:

mean=67

deviations=

squared deviations=

"average" sq. deviation=

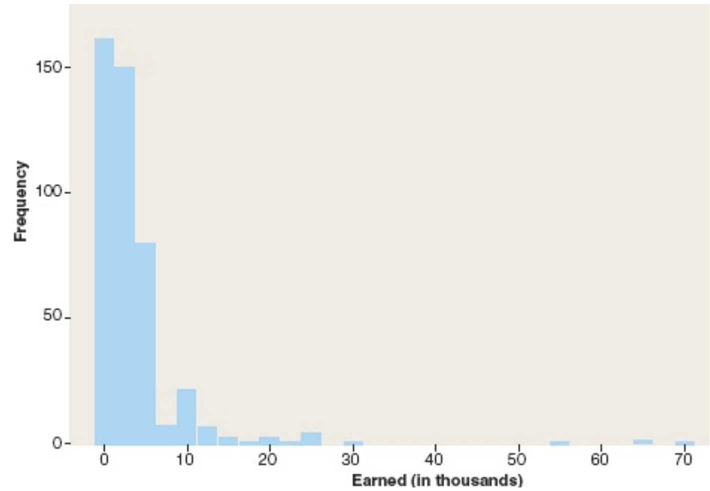
s=sq. root of "average" sq. deviation =

(This is the typical distance from the average height 67.)

Example: *How Shape Affects Standard Deviation*

- **Background**:Output, histogram for student earnings:

Variable	N	Mean	Median	TrMean	StDev	SE Mean
Earned	446	3.776	2.000	2.823	6.503	0.308



In fact, most are within ________ of ________

(Better to report ________________)

- **Question:** Should we say students averaged $3776, and earnings differed from this by about $6500? If not, do these values seem too high or too low?
- **Response:**

Focus on Particular Shape: Normal

- **Symmetric:** just as likely for a value to occur a certain distance below as above the mean.

  Note: if shape is normal, mean equals median

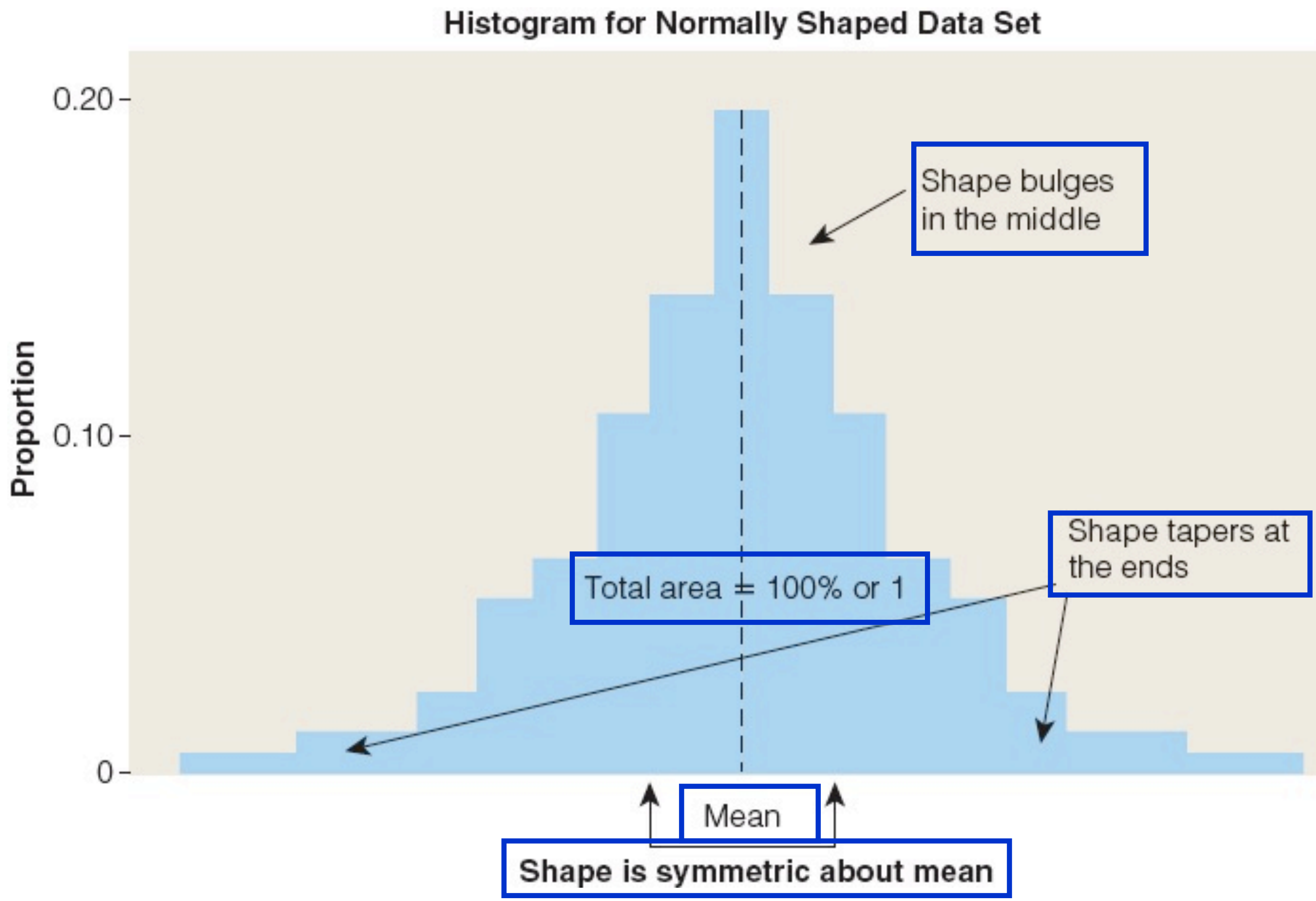
- **Bell-shaped:** values closest to mean are most common; increasingly less common for values to occur farther from mean

Focus on Area of Histogram

Can adjust vertical scale of any histogram so it shows percentage by areas instead of heights.

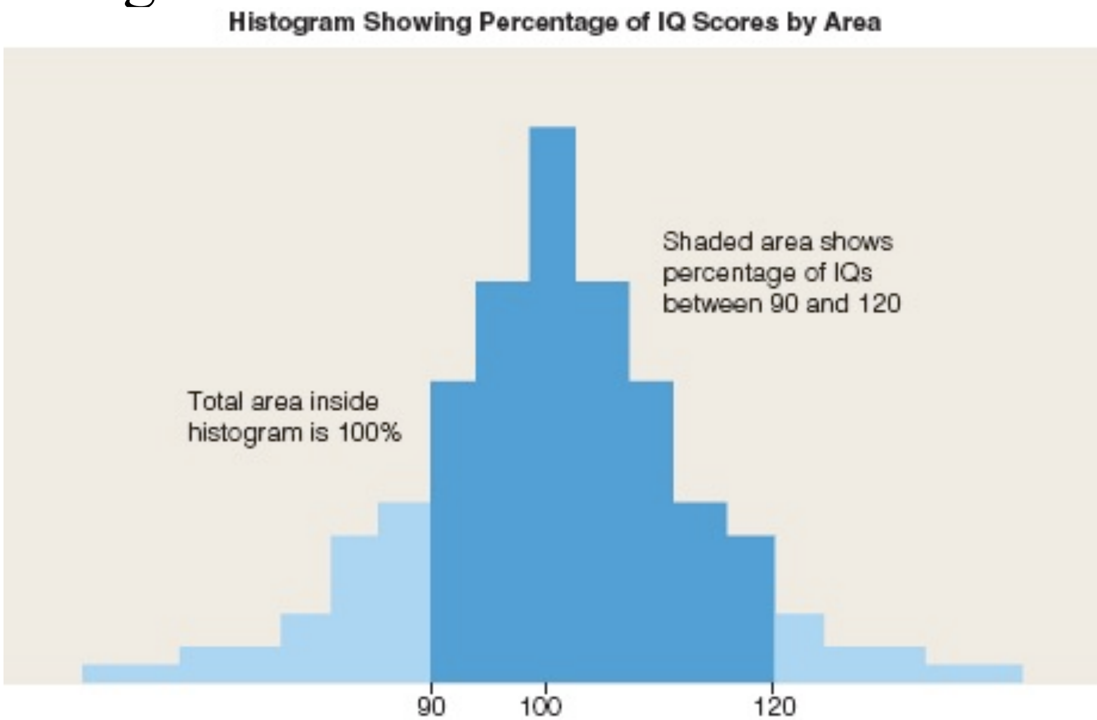
Then total area enclosed is 1 or 100%.

Histogram of Normal Data



Example: *Percentages on a Normal Histogram*

- **Background**: IQs are normal with a mean of 100, as shown in this histogram.



- **Question:** About what percentage are between 90 and 120?
- **Response:**

What We Know About Normal Data

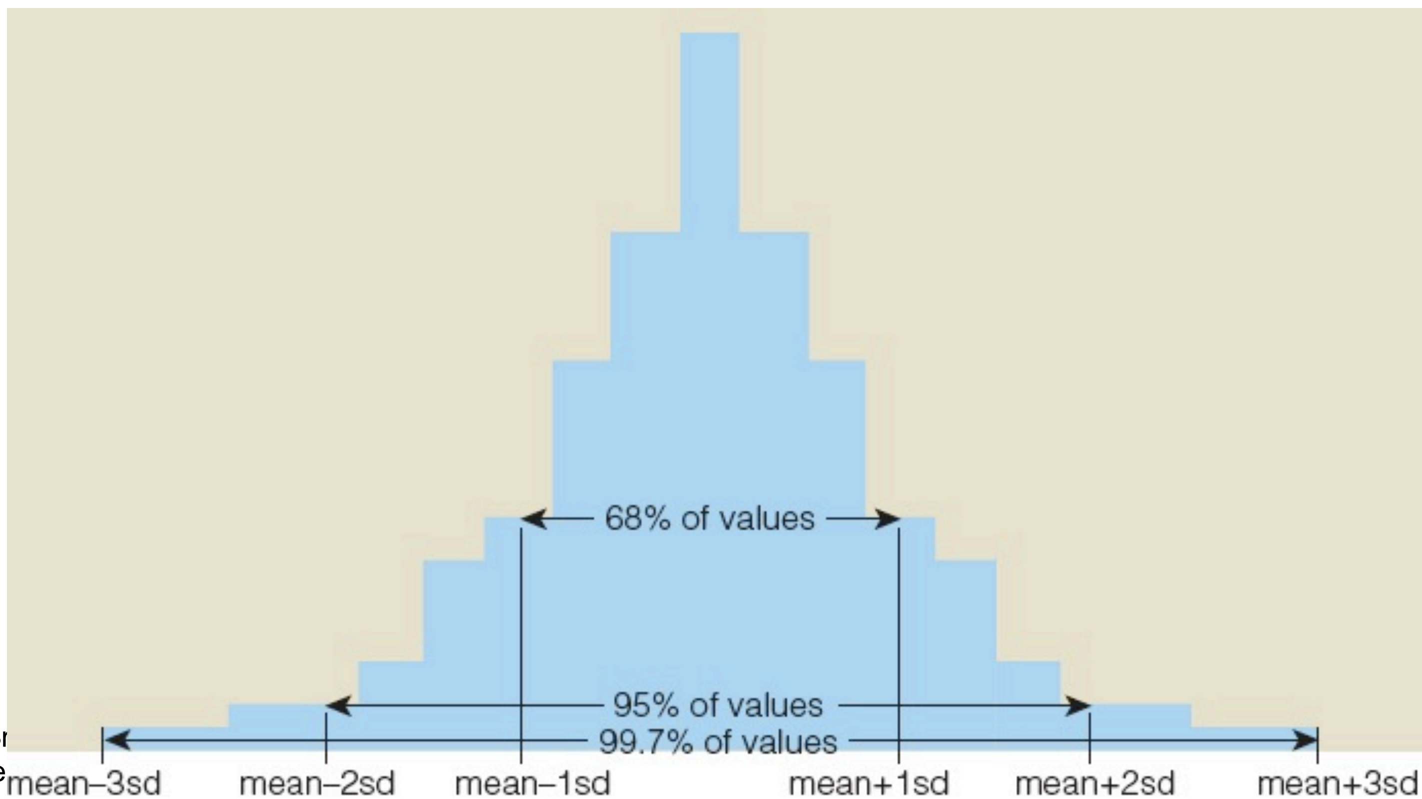
If we know a data set is normal (shape) with given mean (center) and standard deviation (spread), then it is known what percentage of values occur in ***any*** interval.

Following rule presents "tip of the iceberg", gives general feel for data values:

68-95-99.7 Rule for Normal Data

Values of a **normal** data set have

- 68% within 1 standard deviation of mean
- 95% within 2 standard deviations of mean
- 99.7% within 3 standard deviations of mean



68-95-99.7 Rule for Normal Data

If we denote mean $\bar{x}$ and standard deviation s then values of a **normal** data set have

- 68% in $(\bar{x} - 1s, \bar{x} + 1s)$

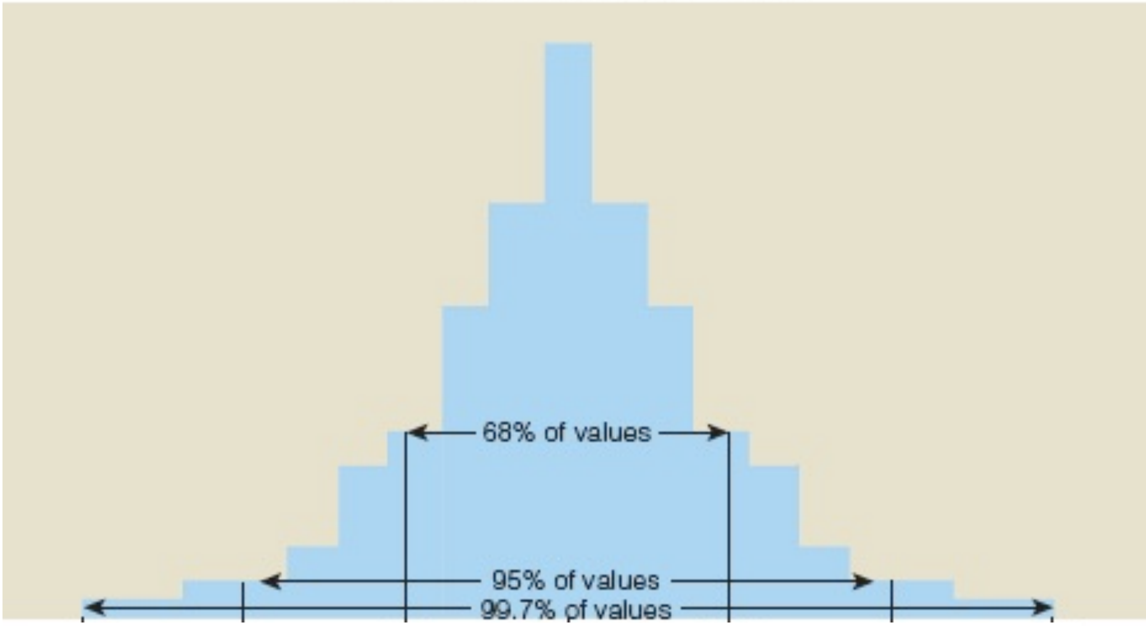
- 95% in $(\bar{x} - 2s, \bar{x} + 2s)$

- 99.7% in $(\bar{x} - 3s, \bar{x} + 3s)$

Example: *Using Rule to Sketch Histogram*

- **Background**: Shoe sizes for 163 adult males normal with mean 11, standard deviation 1.5.
- **Question:** How would the histogram appear?
- **Response:**

68-95-99.7 Rule for Male Shoe Sizes



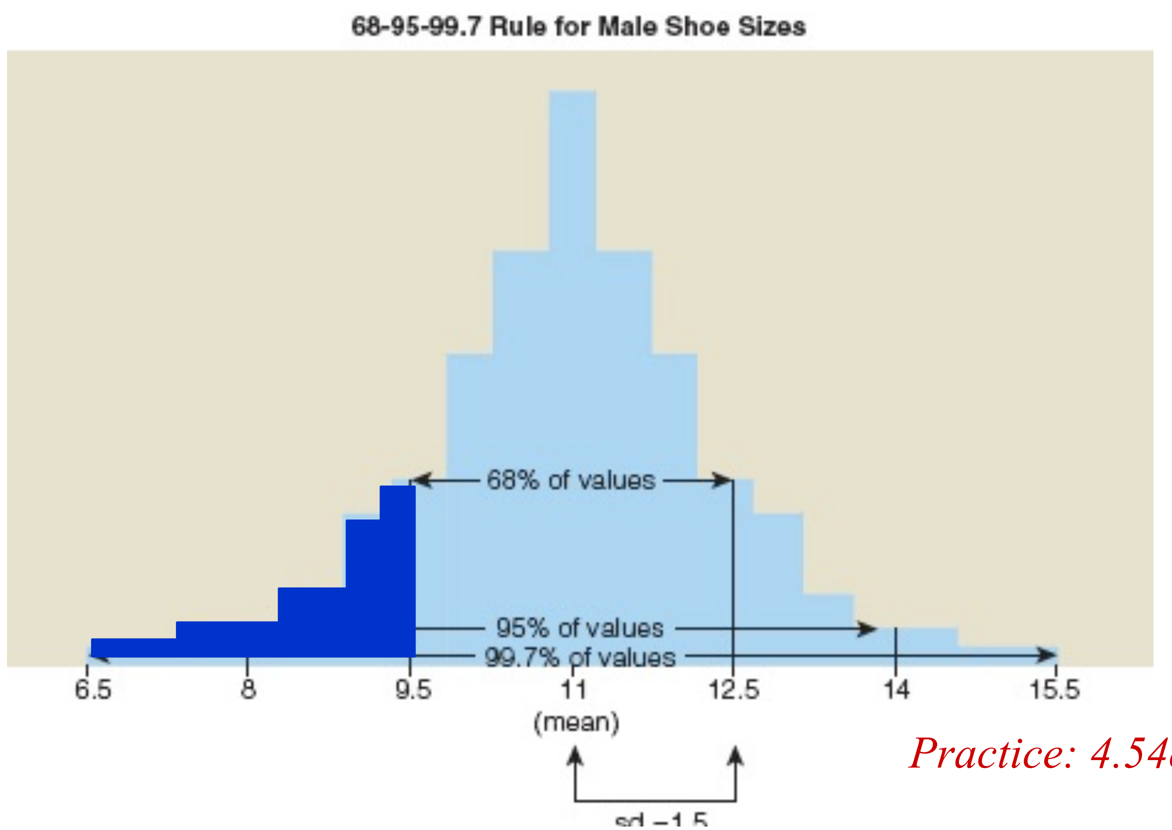
Practice: 4.54a p.121

Example: *Using Rule to Summarize*

- **Background**: Shoe sizes for 163 adult males normal with mean 11, standard deviation 1.5.
- **Question:** What does the 68-95-99.5 Rule tell us about those shoe sizes?
- **Response:**
  - 68% in ____________________________
  - 95% in ____________________________
  - 99.7% in ___________________________

Example: *Using Rule for Tail Percentages*

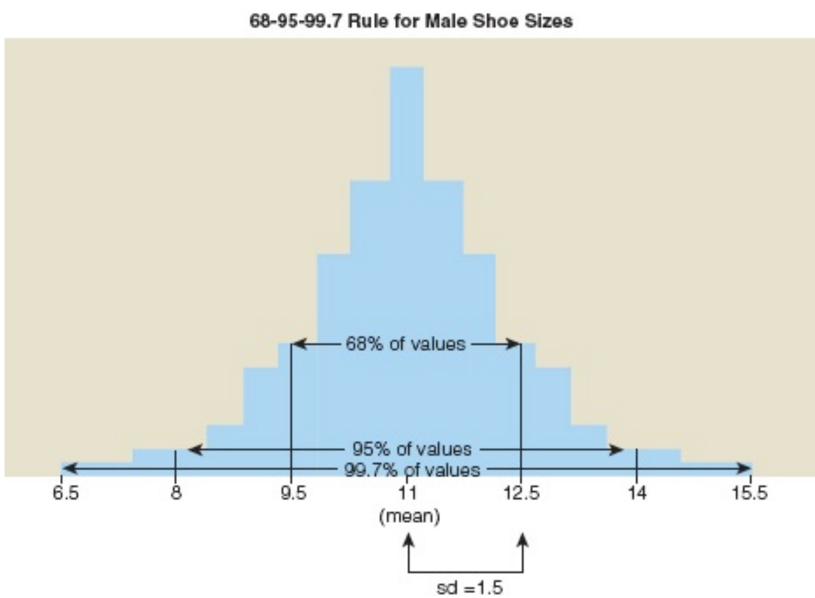
- **Background**: Shoe sizes for 163 adult males normal with mean 11, standard deviation 1.5.
- **Question:** What percentage are less than 9.5?
- **Response:**



Practice: 4.54c p.121

Example: *Using Rule for Tail Percentages*

- **Background**: Shoe sizes for 163 adult males normal with mean 11, standard deviation 1.5.
- **Question:** The bottom 2.5% are below what size?
- **Response:**



Practice: 4.54d p.121

From Histogram to Smooth Curve

- Start: quantitative variable with infinite possible values over continuous range.

(Such as foot lengths, not shoe sizes.)

- Imagine infinitely large data set.

(Infinitely many college males, not just a sample.)

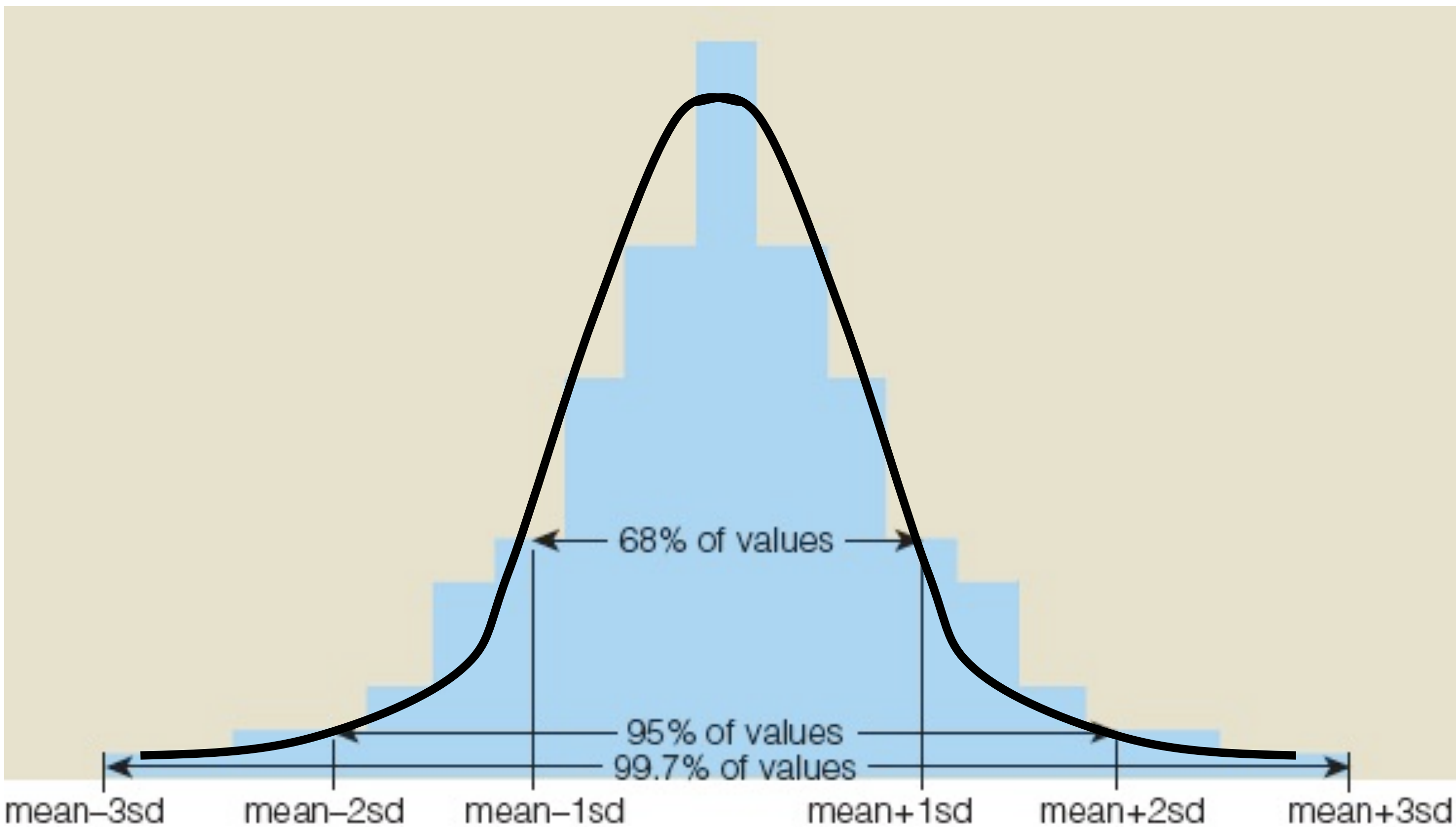
- Imagine values measured to utmost accuracy.

(Record lengths like 9.7333…, not just to nearest inch.)

- Result: histogram turns into smooth curve.
- If shape is normal, result is normal curve.

From Histogram to Smooth Curve

- If shape is normal, result is normal curve.

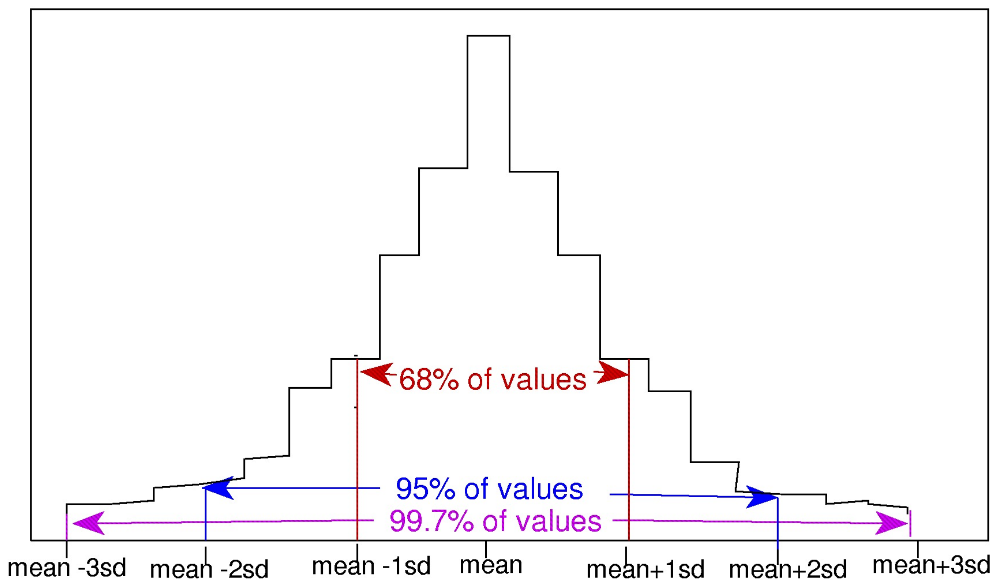


68-95-99.7 Rule *(Review)*

If we know the shape is normal, then values have

- 68% within 1 standard deviation of mean
- 95% within 2 standard deviations of mean
- 99.7% within 3 standard deviations of mean

68-95-99.7 Rule for Normal Distributions



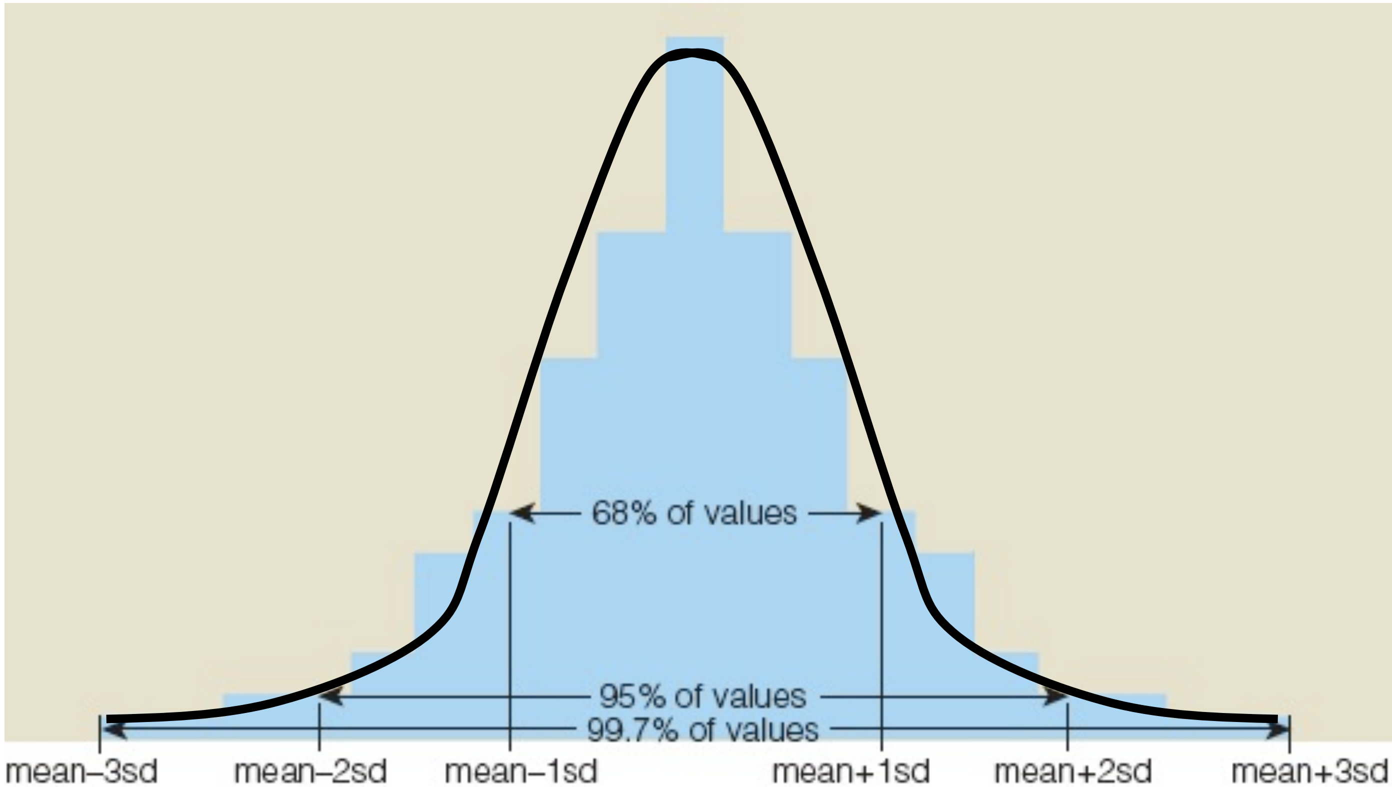
A Closer Look: *around 2 sds above or below the mean may be considered unusual.*

Quantitative Samples vs. Populations

- Summaries for sample of values
  - Mean $\bar{x}$
  - Standard deviation s
- Summaries for population of values
  - Mean μ (called "mu")
  - Standard deviation σ (called "sigma")

Notation: Mean and Standard Deviation

- Distinguish between sample and population



μ

σ

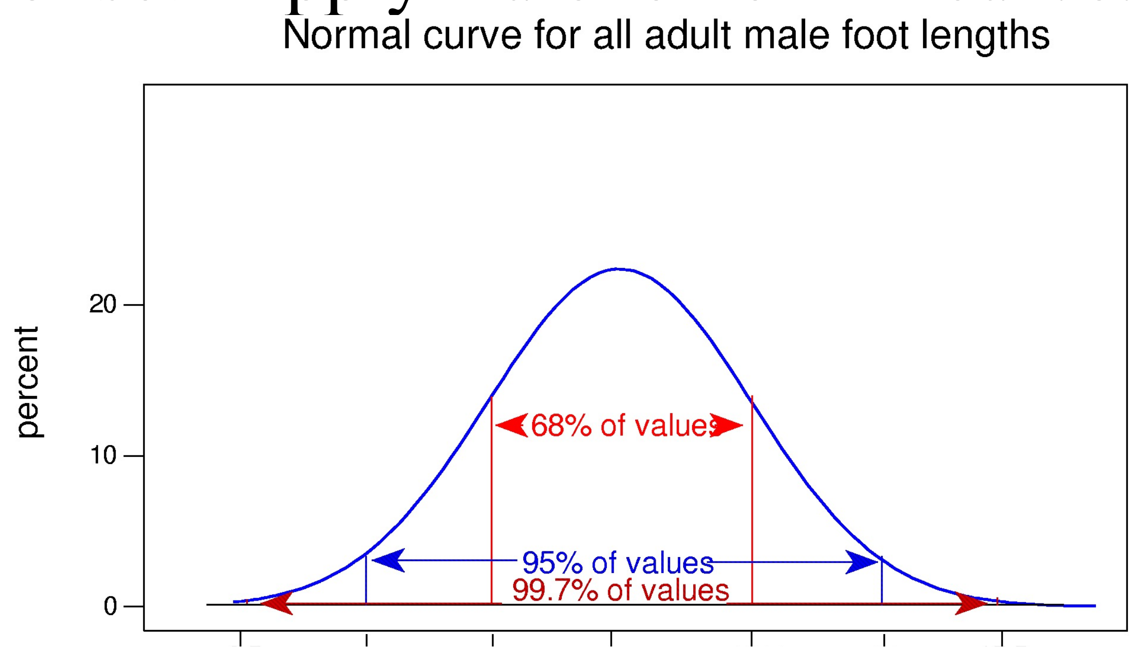
Example: *Notation for Sample or Population*

- **Background:** Adult male foot lengths are normal with mean 11, standard deviation 1.5. A sample of 9 male foot lengths had mean 11.2, standard deviation 1.7.
- **Questions:**
  - What notation applies to sample?
  - What notation applies to population?
- **Responses:**
  - If summarizing sample:
  - If summarizing population:

Practice: 4.55a-b p.121

Example: *Picturing a Normal Curve*

- **Background:** Adult male foot length normal with mean 11, standard deviation 1.5 (inches)
- **Question:** How can we display all such foot lengths?
- **Response:** Apply Rule to normal curve:



Example: *When Rule Does Not Apply*

- **Background:** Ages of all undergrads at a university have mean 20.5, standard deviation 2.9 (years).
- **Question:** How could we display the ages?
- **Response:**

Practice: 4.62 p.122

Standardizing Normal Values

We count distance from the mean, in standard deviations, through a process called **standardizing**.

68-95-99.7 Rule *(Review)*

If we know the shape is normal, then values have

- 68% within 1 standard deviation of mean
- 95% within 2 standard deviations of mean
- 99.7% within 3 standard deviations of mean

Note: around 2 sds above or below mean may be considered "unusual".

Example: *Standardizing a Normal Value*

- **Background:** Ages of mothers when giving birth is approximately normal with mean 27, standard deviation 6 (years).
- **Question:** Are these mothers unusually old to be giving birth? (a) Age 35 (b) Age 43
- **Response:**

(a) Age 35 is ____________________ sds above mean:
Unusually old? ____

(b) Age 43 is ____________________ sds above mean:
Unusually old? ____

Practice: 4.48a p.119

Definition

- ***z*-score**, or **standardized value**, tells how many standard deviations below or above the mean the original value ***x*** is:

$$z = \frac{\text{value} - \text{mean}}{\text{standard deviation}}$$

- Notation:
  - Sample: $z = \frac{x - \bar{x}}{s}$
  - Population: $z = \frac{x - \mu}{\sigma}$
- **Unstandardizing** *z*-scores:

Original value *x* can be computed from *z*-score.

Take the mean and add *z* standard deviations:

$$x = \mu + z\sigma$$

Lecture Summary
(Quantitative Summaries, Begin Normal)

- **Mean:** typical value (average)
- **Mean vs. Median:** affected by shape
- **Standard Deviation:** typical distance from mean
- **Mean and Standard Deviation:** affected by outliers, skewness
- **Normal Distribution:** symmetric, bell-shape
- **68-95-99.7 Rule:** key values of normal dist.
- **Sketching Normal Histogram & Curve**
- **Notation:** sample vs. population
- **Standardizing:** z=(value-mean)/sd