3 Numerical Summaries of Data

3.1 Measures of Center

Measures of Center of Quantitative Data

Summarize the data by taking an average.

Example: Suppose that a student has the following 6 quiz grades: 95, 100, 85, 89, 10, 97.

In general, the *sample mean* of n observations on x is denoted

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum x}{n}.$$

The **sample median** is the middle value when the measurements are arranged from smallest to largest. For an even number of data, the median is the average of the two middle values.

When outliers exist or at least one tail of the distribution is heavy, then the **sample**

median typically is preferred over the **sample mean**, as a measurement of center. Otherwise, the **sample mean** typically has less variability and is preferred over the **sample median**.

For **symmetric** distributions (with a finite mean, μ), the population mean and population median are equal.

For **left** skewed distributions, the population mean is less than the population median.

For **right** skewed distributions, the population mean is greater than the population median.

Example: Explain how to estimate the population average income, μ, in Virginia. Also explain how to estimate the population median income in Virginia.

Estimating a Proportion

Example: Suppose the governor is interested in the proportion of Virginian adults who approve of his budget.

Let p be the population proportion of Virginian adults who approve of his budget.

A **sample proportion** is the **sample mean** of zeros and ones, where "one" indicates

success and "zero" indicates failure.

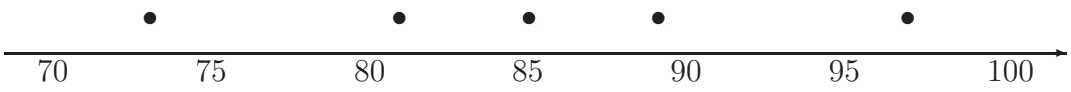
3.2 Measures of Spread

Measures of Spread of Quantitative Data

The **sample range** is the maximum value minus the minimum value; this measure is reasonable for small data sets, but not large ones.

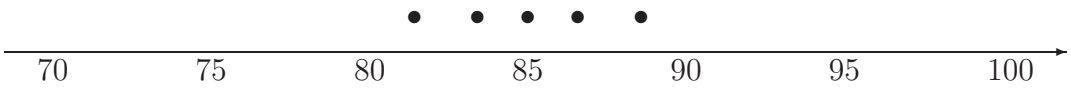
Example: Suppose a sample of 100 incomes among employed Virginians might have the smallest value of \$10,000 and the largest value of \$300,000.

Example: Grades by instructor **Jill** for 5 students on a chemistry exam are {81, 85, 97, 73, 89}.



Dot plot of grades for Jill's class

Example: Grades by instructor **Susan** for 5 students on a chemistry exam are {84, 88, 85, 86, 82}.



Dot plot of grades for Susan's class

Which instructor do you prefer?

$$\text{sample variance} = s^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum x^2 - \frac{1}{n}(\sum x)^2}{n-1}$$

$$\text{sample standard deviation} = s = \sqrt{s^2}$$

Example: Jill's class

Example: Susan's class

Remark: Adding a constant to the data does not affect s. For example, adding 10 points to everyone's grade increases $\bar{X}$ and the sample median by 10 points but does not affect s, the spread.

Remark: Multiplying data by a constant c changes s by a factor of $|c|$.

Example: At a company the salaries are \$50,000, \$55,000, and \$60,000.

Example: Consider the data {8, 8, 8, 8, 8, 8}.

Brief review of means and standard deviations

The **mean**, μ, of a random variable is the **average** of all outcomes in the population, and is the limiting value of $\bar{X}$ as n gets large.

The **standard deviation**, σ, of a random variable measures the **spread** of all outcomes in the population, and is the limiting value of s as n gets large.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

The **variance**, σ^2, of a random variable also measures the spread in the population, and is the limiting value of s^2 as n gets large.

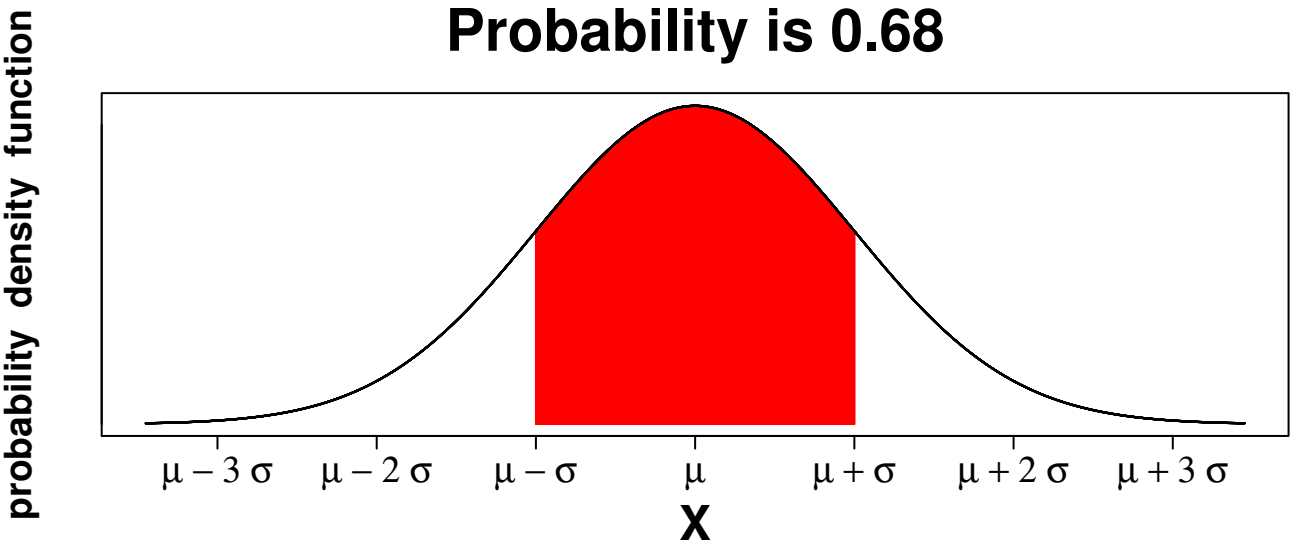
$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

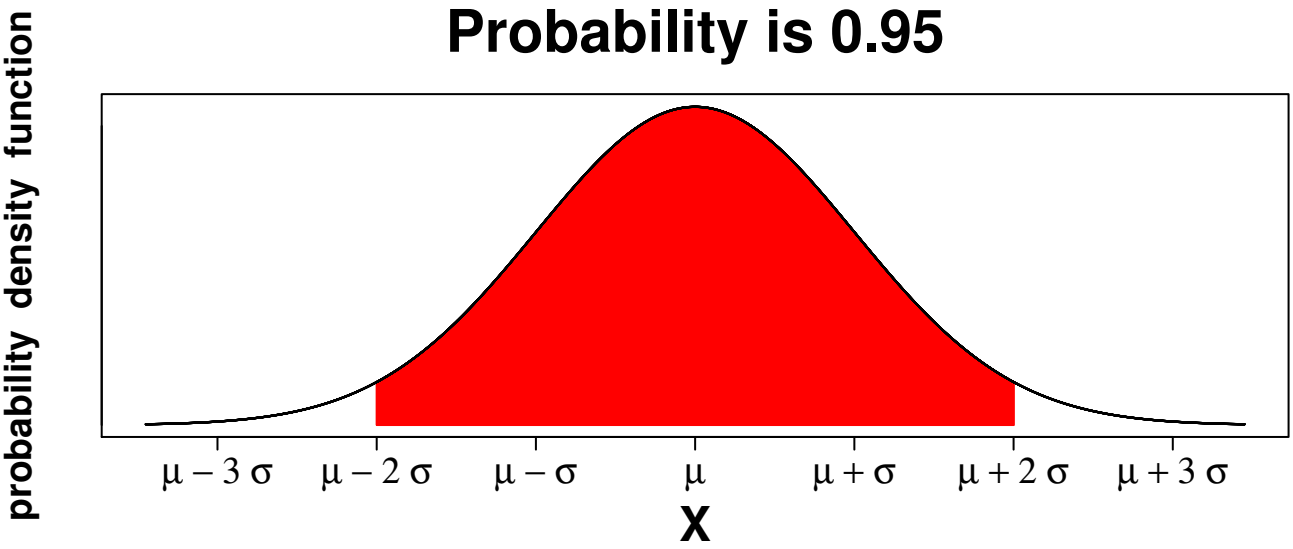
σ is more intuitive than σ^2, partly because σ has the same units as the original data.

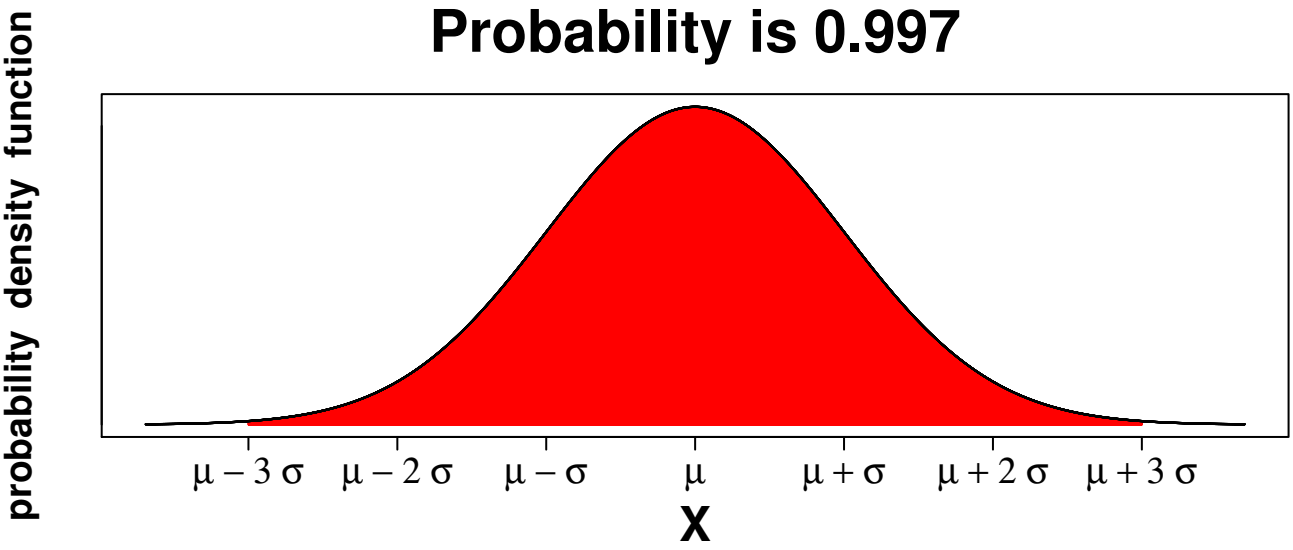
Empirical Rule

If a large number of observations are sampled from an approximately **normal** distribution, then (usually)

1. Approximately 68% of the observations fall within **one** standard deviation, σ, of the mean, μ.
2. Approximately 95% of the observations fall within **two** standard deviations, σ, of the mean, μ.
3. Approximately 99.7% of the observations fall within **three** standard deviations, σ, of the mean, μ.


Probability is 0.68
probability density function
μ − 3 σ
μ − 2 σ
μ − σ
μ
μ + σ
μ + 2 σ
μ + 3 σ
X


Probability is 0.95
probability density function
μ − 3 σ
μ − 2 σ
μ − σ
μ
μ + σ
μ + 2 σ
μ + 3 σ
X


Probability is 0.997
probability density function
μ − 3 σ
μ − 2 σ
μ − σ
μ
μ + σ
μ + 2 σ
μ + 3 σ
X

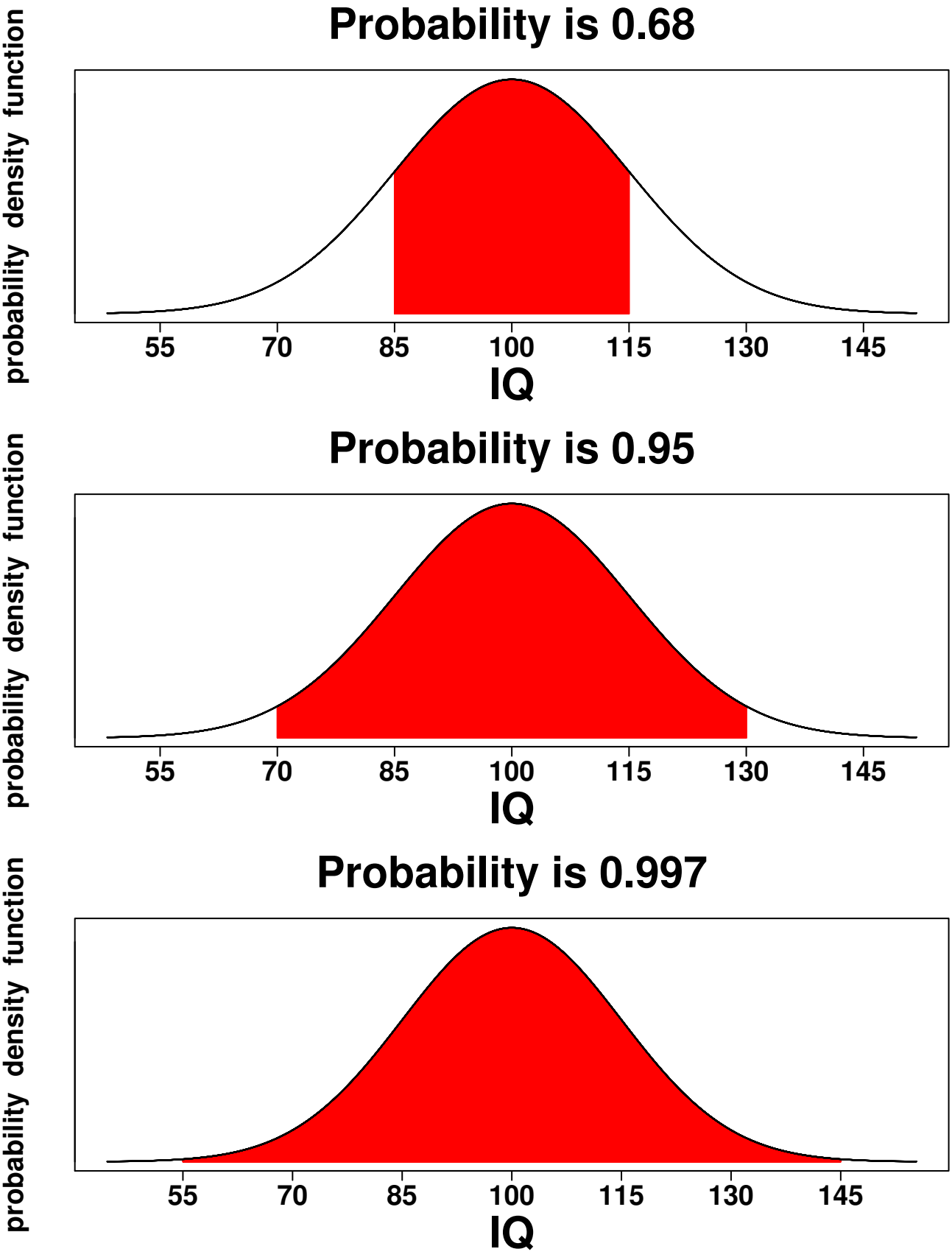
Example: IQ scores of normal adults on the Weschler test have a symmetric bell-shaped distribution with a mean of 100 and standard deviation of 15.

(a) If 1000 adults are sampled, approximately how many have IQs between 85 and 115?

(b) If 1000 adults are sampled, approximately how many have IQs between 70 and 130?

(c) If 1000 adults are sampled, approximately how many have IQs between 55 and 145?

(d) If 1000 adults are sampled, approximately how many have IQs greater than 130?


Probability is 0.68
probability density function
55
70
85
100
115
130
145
IQ
Probability is 0.95
probability density function
55
70
85
100
115
130
145
IQ
Probability is 0.997
probability density function
55
70
85
100
115
130
145
IQ

3.3 Measures of Position

The z-Score

Z represents the number of standard deviations, σ, away from the mean, μ.

Z is the "standardized" variable, known as the Z-score, and has **no units**.

$$z = \frac{x - \mu}{\sigma} \qquad\qquad x = \mu + \sigma\ z$$

Definition: The ***p*th percentile** separates the lowest $p\%$ of the data from the highest $(100 - p)\%$.

Finding the pth percentile

* Sort the n observations in increasing order.
* Calculate $L = (p/100)n$.
* If L is a whole number, then the pth percentile is the average of the Lth and $(L + 1)$th ordered observations.
* If L is NOT a whole number, then the pth percentile is $\lceil L \rceil$th ordered observation, where $\lceil L \rceil$ is the integer rounded UPWARD from L.

Definition:

* Q_1 is the **first quartile**, which is 25th percentile.
* Q_2 is the **second quartile**, which is 50th percentile, also the sample median.
* Q_3 is the **third quartile**, which is 75th percentile.

Interquartile Range

To determine the **interquartile range**, first determine the **lower quartile** and the **upper quartile**.

The **interquartile range** or **IQR** is **upper quartile** minus **lower quartile**.

The **5-number summary** is {minimum, lower quartile, median, upper quartile, maximum}.

The **5-number summary** divides the data into four (roughly) equal sections (fourths).

Example: Consider the following 14 observations:

{45, 48, 53, 103, 160, 10, 63, 68, 70, 71, 55, 58, 81, 82}.

How do the above results change if we replace 160 by 1,000,000?

□

Definition: An observation is an **outlier** if it is at least $1.5 \times$ IQR from its nearest quartile.

The Box Plot: Graphing a Five-Number Summary of Position

Procedure:

1. Draw rectangle with edges at lower and upper quartiles.
2. Draw a line through the box at the sample median.
3. Plot the outliers using asterisks.

4. Draw **whiskers**; i.e., lines from the edge of the box to the most extreme observations which are not outliers.

Previous example: Consider the following 14 observations:

{45, 48, 53, 103, 160, 10, 63, 68, 70, 71, 55, 58, 81, 82}.

Mathematically determine if there are any outliers on the left.

Mathematically determine if there are any outliers on the right.

Which measures of center and spread should one use?

Suppose a data set has outliers (or the distribution has at least one heavy tail).

Suppose a data set seems to be from a distribution which is **normal** (or approximately normal).