Math 140
Introductory Statistics

Professor Silvia Fernández

Lecture 11
Based on the book *Statistics in Action*
by A. Watkins, R. Scheaffer, and G. Cobb.

6.1 Probability Distribution from Data

We have three ways of specifying a population:

- 1. List of all (individual) units
- 2. Frequency Table (p. 68)
- 3. Relative Frequency or
  Proportion Table (p. 359)

How can we calculate mean and SD on each?

List of all units

Number	Type	Value x	x - μ
1	Penny	1 ¢	-3
2	Penny	1 ¢	-3
3	Penny	1 ¢	-3
4	Penny	1 ¢	-3
5	Penny	1 ¢	-3
6	Nickel	5 ¢	1
7	Nickel	5 ¢	1
8	Nickel	5 ¢	1
9	Dime	10 ¢	6
10	Dime	10 ¢	6
	Total = 10 coins	Sum = 40 cents	

$$\mu = \text{population mean} = \frac{\sum x}{n}$$

$$\mu = \frac{1+1+1+1+1+5+5+5+10+10}{10} = 4$$

$$\sigma_n = \text{SD} = \sqrt{\frac{\sum (x-\mu)^2}{n}}$$

$$\sigma_n = \sqrt{\frac{9+9+9+9+9+1+1+1+36+36}{10}} =$$

$$\sigma_n = \sqrt{\frac{120}{10}} = \sqrt{12} \approx 3.4641$$

Frequency Tables

Type	Value x	Frequency f	$x \cdot f$
Penny	1 ¢	5	5
Nickel	5 ¢	3	15
Dime	10 ¢	2	20
Sum		10	40

$$n = \sum f = 10$$

$$\mu = \text{population mean} = \frac{\sum x \cdot f}{n}$$

$$\mu = \frac{5 + 15 + 20}{10} = 4$$

$$\sigma_n = \text{SD} = \sqrt{\frac{\sum (x - \mu)^2 \cdot f}{n}}$$

$$\sigma_n = \sqrt{\frac{9 \cdot 5 + 1 \cdot 3 + 36 \cdot 2}{10}} =$$

$$\sigma_n = \sqrt{\frac{120}{10}} = \sqrt{12} \approx 3.4641$$

Relative Frequency or Proportion Table

Type	Value x	Proportion of coins $P(x)$	$x \cdot P(x)$
Penny	1 ¢	0.5	0.5
Nickel	5 ¢	0.3	1.5
Dime	10 ¢	0.2	2.0
Sum		1	4.0

$$\mu = \text{population mean} = \sum x \cdot P(x)$$

$$\mu = 0.5 + 1.5 + 2.0 = 4$$

$$\sigma_n = \text{SD} = \sqrt{\sum (x-\mu)^2 \cdot P(x)}$$

$$\sigma_n = \sqrt{9 \cdot (0.5) + 1 \cdot (0.3) + 36 \cdot (0.2)} =$$

$$\sigma_n = \sqrt{4.5 + 0.3 + 7.2} = \sqrt{12} \approx 3.4641$$

Summary of Mean/SD

List of all units

$$\mu = \frac{\sum x}{n}$$

$$\sigma_n = \sqrt{\frac{\sum (x-\mu)^2}{n}}$$

Frequency Table

$$\mu = \frac{\sum x \cdot f}{n}$$

$$\sigma_n = \sqrt{\frac{\sum (x-\mu)^2 \cdot f}{n}}$$

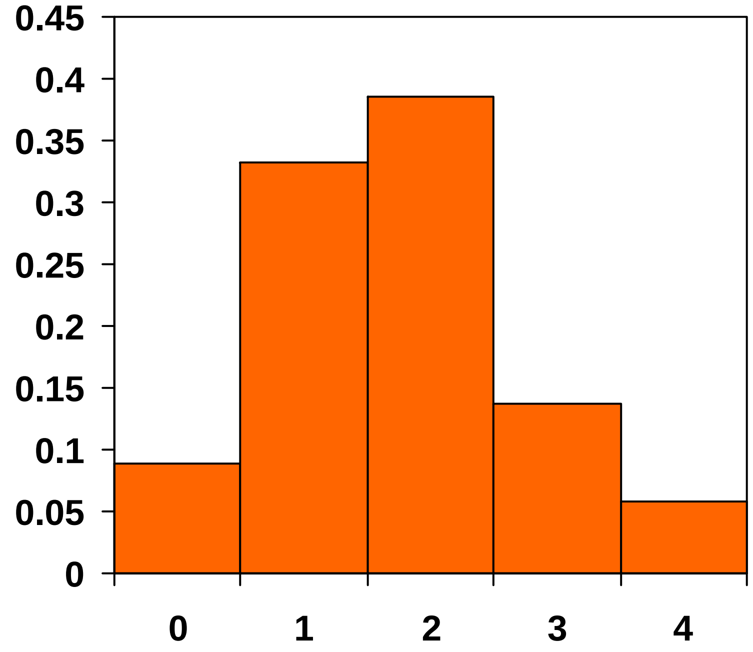
Relative Frequency (or Proportion) Table

$$\mu = \sum x \cdot P(x)$$

$$\sigma_n = \sqrt{\sum (x-\mu)^2 \cdot P(x)}$$

Example (page 359)

Number of Motor Vehicles (per household), x	Proportion of households, $P(x)$
0	0.088
1	0.332
2	0.385
3	0.137
4	0.058



The number of motor vehicles per household. The “4” represents “4 or more,” but the proportion of households with more than four vehicles is very small.

[Source: U.S. Census Bureau, American Community Survey, 2004, factfinder.census.gov.]

How to Sample

We have three ways of specifying a population:

- 1. List of all (individual) units
- 2. Frequency Table (p. 68)
- 3. Relative Frequency or
  Proportion Table (p. 359)

How can we get (or simulate) a sample from these distributions?

How to Sample

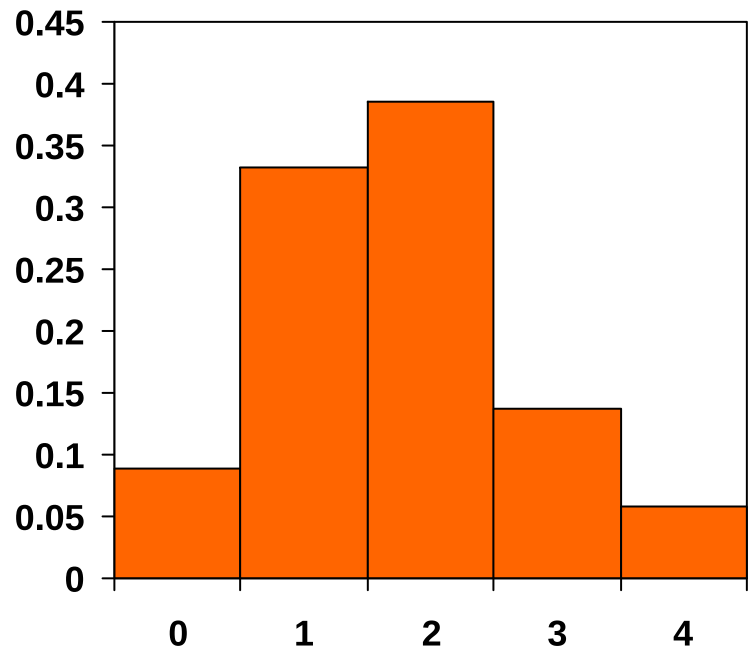
- List of all Units and Frequency Tables.
  - 1. Make a numbered list of all units. If using a frequency table remember to repeat each value x according to its **multiplicity**.
  - 2. Draw as many numbers at random as needed.
- Relative Frequency (or Proportion) Tables
  - 1. Assign a number to each proportion in the table, using as many digits as the decimal precision of the proportions.
  - 2. Select as many of these numbers at random as needed.
  - 3. Use the assignation to realize the characteristics of the values in the sample.

Ways to generate random numbers

- By using **`rand`** or **`randint`** in your calculator (Preferred way)
- By using a string or a table of random digits (page 828)
- By writing numbers in slips of paper, mixing, and drawing at random.

Example (page 359)

Number of Motor Vehicles (per household), x	Proportion of households, $P(x)$
0	0.088
1	0.332
2	0.385
3	0.137
4	0.058



The number of motor vehicles per household. The “4” represents “4 or more,” but the proportion of households with more than four vehicles is very small.

[Source: U.S. Census Bureau, American Community Survey, 2004, factfinder.census.gov.]

Example (part 1)

Number of Motor Vehicles (per household), x	Proportion of households, $P(x)$	Random Numbers Representing this Category
0	0.088	$001\text{-}088$
1	0.332	$089\text{-}420$
2	0.385	$421\text{-}805$
3	0.137	$806\text{-}942$
4	0.058	$943\text{-}999{,}000$

420 = 88 + 332

805 = 420 + 385

942 = 805 + 137

1000 = 942 + 58

Note: If you are using a random-digit table then you should use 000 as the equivalent of 1000

Example (part 2)

Number of Motor Vehicles (per household), x	Proportion of households, $P(x)$	Random Numbers Representing this Category
0	0.088	$001\text{-}088$
1	0.332	$089\text{-}420$
2	0.385	$421\text{-}805$
3	0.137	$806\text{-}942$
4	0.058	$943\text{-}999$ 000

- Suppose we have the following string of random digits (already separated by groups of 3)

391 | 545 | 177 | 981 | 016 | 845 | 248

- The corresponding number of vehicles per household in this sample are:

Random Number	Vehicles in household
391	1
545	2
177	1
981	4

Sampling with or without replacement

- Without replacement.
  - Usual choice in practice.
  - Do not return slips of paper to the box.
  - Disregard repetitions when using tables of random digits or random numbers from a calculator.

Note: The two possibilities yield almost the same results as long as the sample size is small compared to the population size.

- With replacement
  - Seldom used in practice.
  - However it is much easier to do calculations in this case. (see 5.3, 5.4)
  - Do return slips of paper to box.
  - Allow repetitions of numbers when using tables of random digits or random numbers from a calculator.

7.1 Generating Sampling Distributions

- Sampling Distributions. Distribution of summary statistics obtained from taking repeated random samples.
- Steps for generating a sampling distribution:
  - I. Take a random sample of a fixed size n from a population.
  - II. Compute a Summary Statistic for this sample.
  - III. Repeat steps I and II many times.
  - IV. Display the distribution of the Summary Statistic.

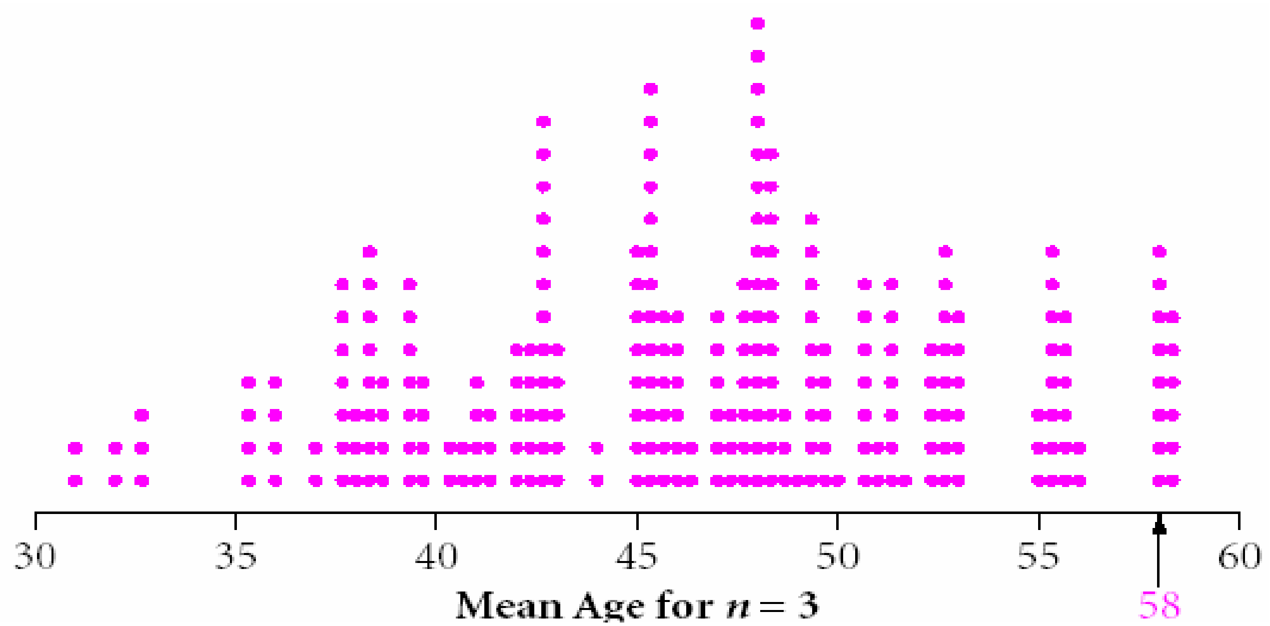
Note: A way to remember these steps is, Random Sample, Summary Statistic, Repetition, Distribution.

Example

- Westvaco case.

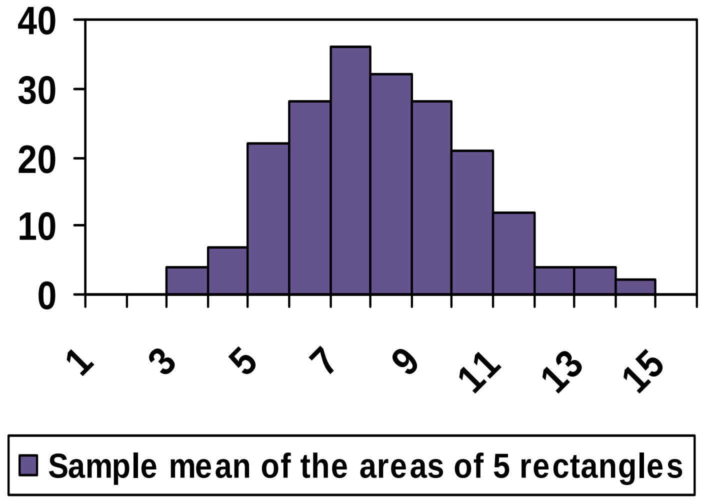
  Randomly select three workers from the group of 10 with ages above, and calculate the mean age of the three selected.

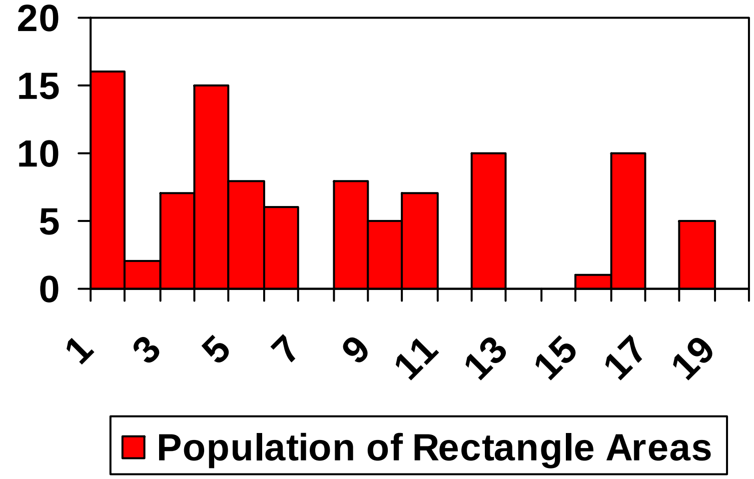
  25 33 35 38 48 55 55 55 56 64



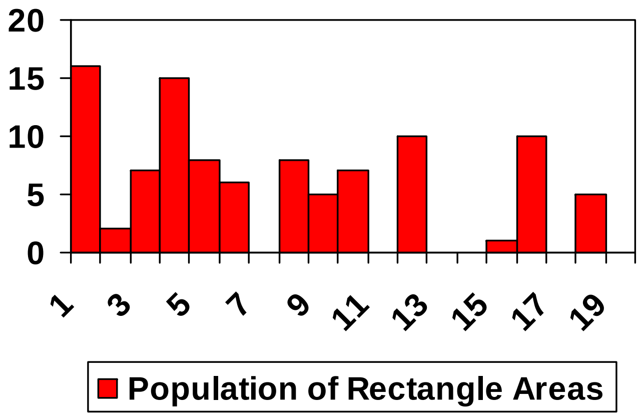
Shape, Center, and Spread

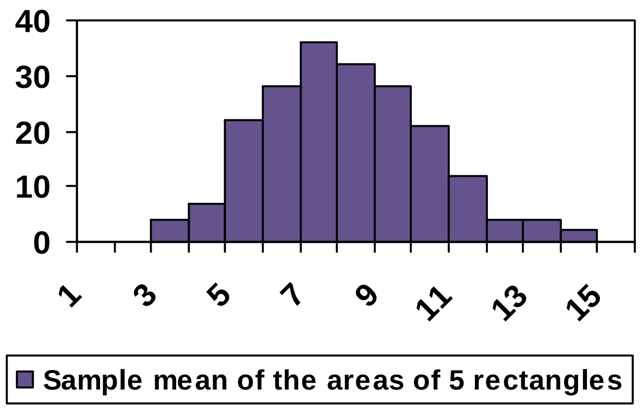
- A good description of a sampling distribution is the trio shape, center, and spread.
- Recall the rectangles activity 4.2. (See displays 5.7 and 5.8)





Shape, Center, and Spread

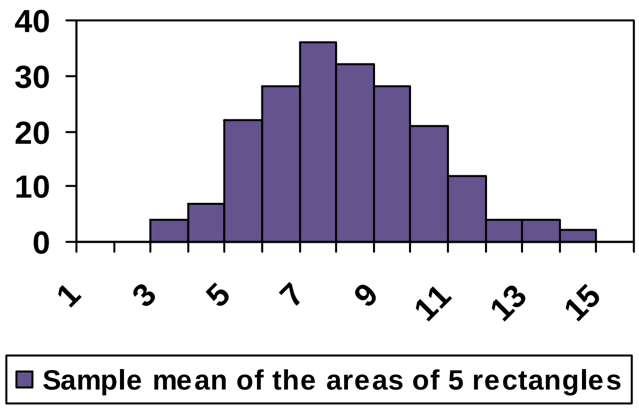




- Shape: Irregular
- Center= Mean = μ = 7.41
- Spread =
  Standard Deviation = σ = 5.23

- Shape: Normal with a hint of skew to the right.
- Center= Mean= $\overline{x}$ = 7.377
- Spread =
  Standard Deviation = SE = 2.23

Shape, Center, and Spread



- Shape: Normal with a hint of skew to the right.
- Center= Mean= $\overline{x}$ = 7.377
- Spread = Standard Deviation = SE = 2.23

Notes.

- The standard deviation of the sampling distribution is often called the Standard Error (SE)
- Most sample distributions are nearly normal, we'll see more about this later.
- Values that are in the middle 95% of a random distribution are called **Reasonably Likely.**
- Values that are in the outer 5% of a random distribution are called **Rare Events**.

Notation

	Population	Sample	Sampling Distribution
Mean	μ	$\overline{x}$	$\mu_{\overline{x}}$
Standard Deviation	σ	s	$\sigma_{\overline{x}}$
Size	N	n	

Properties of The Sampling Distribution of The Sample Mean

- The mean $\mu_{\overline{x}}$ of the sampling distribution of $\overline{x}$ equals the mean of the population μ:

$$\mu_{\overline{x}} = \mu$$

- The standard deviation $\sigma_{\overline{x}}$ of the sampling distribution of $\overline{x}$, also called the **standard error** of the mean, equals the standard deviation of the population σ divided by the square root of the sample size n:

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

- The Shape of the sampling distribution will be approximately normal if the population is approximately normal; for other populations, the sampling distribution becomes more normal as n increases. This property is called the **Central Limit Theorem**.

Example 1

- Problems usually involve a combination of the three properties of the Sampling Distribution of the Sample Mean, together with what we learned about the normal distribution.

- Example: Average Number of Children

  What is the probability that a random sample of 20 families in the United States will have an average of 1.5 children or fewer?

Example 1

- Example: Average Number of Children

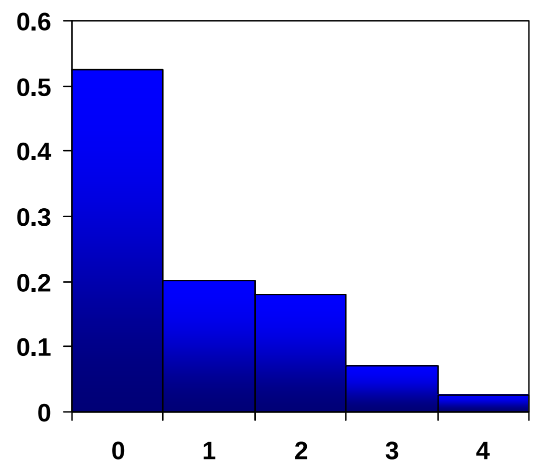
  What is the probability that a random sample of 20 families in the United States will have an average of 1.5 children or fewer?

Number of Children (per family), x	Proportion of families, $P(x)$
0	0.524
1	0.201
2	0.179
3	0.070
4 or more	0.026

- Mean (of population)

  μ = 0.873
- Standard Deviation

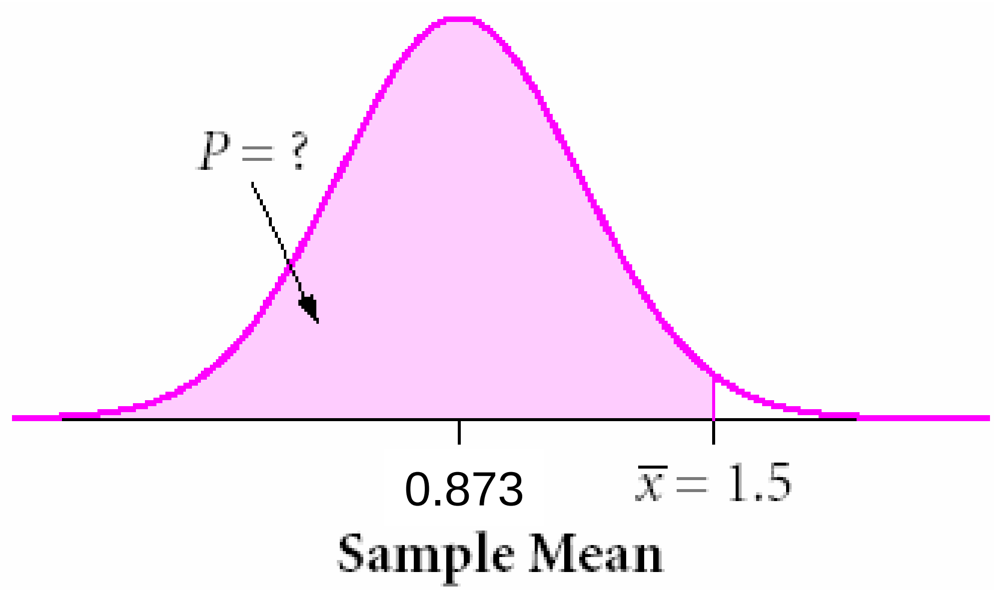
  σ =1.095



Example 1

$$\mu_{\overline{x}} = \mu = 0.873$$

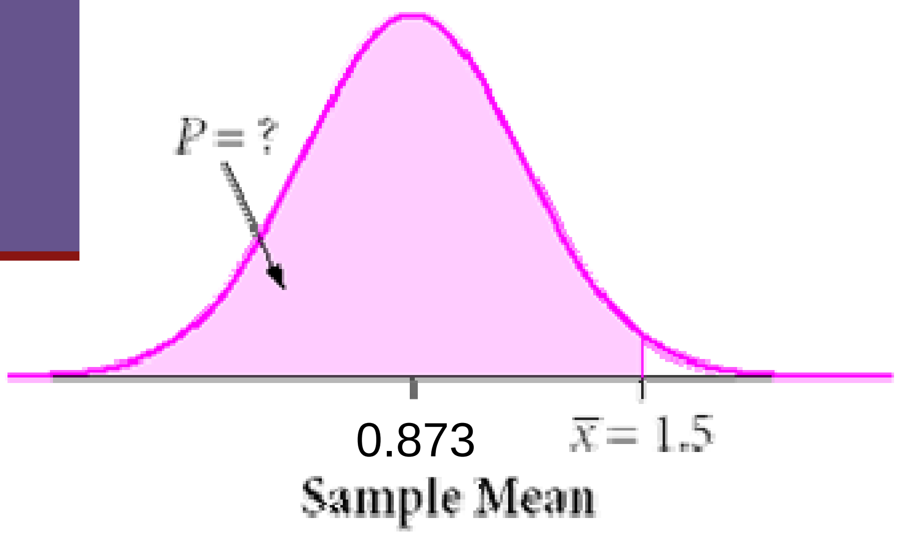
$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.095}{\sqrt{20}} = 0.2448$$



Example 1

$$\mu_{\overline{x}} = \mu = 0.873$$

$$\sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.095}{\sqrt{20}} = 0.2448$$



- Find z-score of the value 1.5

$$z = \frac{\overline{x} - mean}{SD} = \frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} = \frac{1.5 - 0.873}{0.2448} \approx 2.56$$

$$\texttt{normalcdf}(-99999, 2.56) \approx .9947$$

- So in a random sample of 20 families there is a 99.47% probability that the mean number of children per family will be less than 1.5

Example 2

- Example: Reasonably Likely Averages

  What average numbers of children are reasonably likely in a random sample of 20 families?

- Recall that the values that are in the middle 95% of a random distribution are called **Reasonably Likely.**

Example 2

- Example: Reasonably Likely Averages

  What average numbers of children are reasonably likely in a random sample of 20 families?

- Recall that the values that are in the middle 95% of a random distribution are called **Reasonably Likely.**

Note that by calculating the z-scores of 2.5% and 97.5% we find that the Reasonably Likely values are those values within 1.96 standard deviations from the mean.

That is, between $\mu - 1.96\,\sigma$ and $\mu + 1.96\,\sigma$

Finding Probabilities for Sample Totals

- Sometimes situations are stated in terms of the total number in the sample rather than the average number: "What is the probability that there are 30 or fewer children in a random sample of 20 families in the United States?" You have the choice of two equivalent ways to do this problem.
- **Method I:** Find the equivalent average number of children, $\bar{x}$, by dividing the total number of children, 30, by the sample size, 20:

$$\bar{x} = \frac{30}{20} = 1.5$$

Then you can use the same formulas and procedure as in the previous examples.

- **Method II:** Convert the formulas from the previous examples to equivalent formulas for the sum, then proceed as in the next example.

Sampling Distribution of the Sum of a Sample

- If a random sample of size n is selected with mean μ and standard deviation σ, then
  - the mean of the sampling distribution of the sum is

$$\mu_{sum} = n\mu$$

  - the standard error of the sampling distribution of the sum is

$$\sigma_{sum} = \sqrt{n} \cdot \sigma$$

  - the shape of the sampling distribution will be approximately normal if the population is approximately normal; for other populations, the sampling distribution becomes more normal as n increases.

Note: To get the “sum” formulas just multiply by n

Examples 3 and 4

- Ex3: The Probability of 25 or fewer Children

  What is the probability that a random sample of 20 families in the United States will have a total of 25 children or fewer?

- Ex4: Reasonably Likely Totals

  In a random sample of 20 families, what total numbers of children are reasonably likely?