GRAND VALLEY STATE UNIVERSITY
DEPARTMENT OF STATISTICS

PAPER #801

Estimating Metropolitan Area Population

Mary Richardson, richamar@gvsu.edu
Grand Valley State University

Key Words: Active learning; Introductory statistics; Statistical simulation; Sampling distribution for a mean; Confidence interval for a mean; Hypothesis test on a mean

Abstract

This article describes an interactive activity that has students use simulation to illustrate the sampling distribution of a sample mean, properties of confidence intervals, and properties of hypothesis testing. The activity is completed in three parts. The three parts can be used in a sequence or they can be used individually.

1. Introduction

Metropolitan statistical areas are geographic entities defined by the U.S. Office of Management and Budget (OMB) for use by Federal statistical agencies in collecting, tabulating, and publishing Federal statistics. A metropolitan area contains a core urban area of 50,000 or more population. Each metropolitan area consists of one or more counties and includes the counties containing the core urban area, as well as any adjacent counties that have a high degree of social and economic integration (as measured by commuting to work) with the urban core.

The Population Sheet (Appendix D) displays a population of 178 ‘large’ metropolitan statistical areas (where for purposes of our discussion here, ‘large’ is defined as having 300,000 or more residents). For each of the metropolitan areas the number of residents in July 2005 is given (data source: *U.S. Census Bureau Public Information Office*).

Students draw simple random samples from this population of large metropolitan areas and use their sample mean number of residents to make inferences about the population mean number of residents.

2. Activity 1 - Sampling Distribution of the Sample Mean

Prior to completing Activity 1, students should be familiar with stem-and-leaf plots and calculating means and standard deviations. Each student needs the Activity 1 Worksheet (Appendix A), the Population Sheet, and a random number table or calculator capable of generating random numbers.

Ten groups of students are formed and each group is asked to do the following: (1) Select two different random samples of n = 5 metropolitan areas. (2) Select two different random samples of n = 15 metropolitan areas. (3) Select two different random samples of n = 25 metropolitan areas. For smaller class sizes, rather than forming groups, each individual student can be asked to complete the previous tasks.

For each of the samples, students calculate the mean number of residents, reinforcing the idea that the sample mean is a random variable with values that differ from sample to sample. To complete the Data Collection Sheet (on the Activity 1 Worksheet), group results are placed on the white board and combined to obtain 20 sample means for each of the sample sizes n = 5, 15, and 25. Example class results are shown in Table 1 (rounded to the nearest whole number). The true mean number of residents in the population is μ = 1,637,811.

Table 1. Example class sample means for samples of size 5, 15, and 25.

Sample Number	n = 5	n = 15	n = 25
1	1,448,755	1,881,244	1,803,163
2	2,971,892	2,775,578	2,029,830
3	794,053	1,708,526	1,861,056
4	1,832,733	1,682,090	1,597,200
5	1,509,579	1,991,842	1,843,339
6	808,659	2,084,950	1,478,230
7	1,137,194	1,279,757	1,564,694
8	985,521	1,498,579	1,357,386
9	1,986,916	1,794,888	1,831,800
10	1,325,020	1,700,124	1,725,960
11	525,531	1,991,859	2,034,606
12	609,340	806,696	964,183
13	1,291,928	1,336,883	1,219,984
14	3,836,178	2,987,863	2,193,030
15	4,152,034	2,217,635	2,448,689
16	846,443	1,405,726	1,432,462

17	3,445,581	2,127,085	1,957,900
18	1,499,253	1,023,487	1,459,937
19	2,286,597	2,982,072	2,299,444
20	1,288,551	1,336,679	1,985,075

Students answer a series of questions based on the sample means for the different sample sizes. Through answering these questions, students discover properties of the distribution of a sample mean; namely, (i) the distribution of sample mean values is centered at the population mean, (ii) the distribution of sample mean values approaches a normal distribution as the sample size increases, (iii) the distribution of sample mean values has less variability than the distribution of the original population, and (iv) the variability of sample mean values decreases as the sample size increases. An Answer Sheet for the Activity 1 questions is provided in Appendix E.

3. Activity 2 - Confidence Interval for the Population Mean

Prior to completing Activity 2, students should be familiar with how to construct confidence intervals. Each student needs the Activity 2 Worksheet (Appendix B), the Population Sheet, and a random number table or a calculator capable of generating random numbers. Each student selects a simple random sample of 25 metropolitan areas. The population distribution of the number of residents is not a normal distribution (the distribution is skewed to the right), but a confidence interval procedure based on the *t* distribution may still be used, since the sampling distribution of the sample mean is approximately normal for samples of size 25. A $(1-\alpha)100\%$ confidence interval for μ, the population mean, is given by:

$$\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right) \qquad (1)$$

where $t_{n-1,\frac{\alpha}{2}}$ is an appropriate percentile of the *t* distribution having $(n-1)$ degrees of freedom, $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size.

First, each student uses her sample to construct an 80% confidence interval for the population mean number of residents and writes her interval on the white board. The white board displays all of the confidence intervals constructed by the class and students see the results of drawing repeated samples from the same population and calculating 80% confidence intervals for the population mean. Example 80% class confidence intervals are displayed in Table 2.

Some of the confidence intervals will contain the population mean $(\mu = 1{,}637{,}811)$ and some will not. Students discover that if we claim that we are 80% confident that a mean lies within the endpoints of a confidence interval, we are saying that the endpoints of the confidence interval were calculated by a method that gives correct results in 80% of all possible random samples. There is not an 80% chance that an individual interval contains the population mean (an individual interval either contains the mean or it does not). Students are asked to write a statement explaining how to properly interpret an 80% level of confidence.

Next, students are asked to use their samples to construct a 99% confidence interval for the population mean number of residents. The class confidence intervals are displayed on the white board and the results are discussed. Example 99% class confidence intervals are displayed in Table 2. Students are asked to write a statement explaining how a 99% level of confidence should be interpreted. Students are asked to write a statement explaining how increasing the confidence level from 80% to 99% changed the confidence intervals. In addition, students are asked to give one advantage of using 99% confidence rather than 80% confidence and to give one disadvantage. An Answer Sheet for the Activity 2 questions is provided in Appendix E.

Table 2. Example class 80% and 99% Confidence Intervals.

Student	80% Confidence Intervals	99% Confidence Intervals
1	(663,600; 1,831,292)	(8,307; 2,486,585)
2	(1,137,220; 2,061,956)	(618,271; 2,580,905)
3	(1,212,453; 2,349,287)	(574,477; 2,987,263)
4	(980,731; 1,507,522)	(685,102; 1,803,150)
5	(1,431,279; 3,077,799)	(507,274; 4,001,804)
6	(986,739; 2,216,243)	(296,758; 2,906,223)
7	(1,470,217; 2,672,523)	(795,499; 3,347,242)
8	(1,251,883; 2,446,147)	(581,678; 3,116,351)
9	(1,240,407; 2,693,918)	(424,717; 3,509,609)
10	(758,981; 2,454,501)	(-192,522; 3,406,003)
11	(1,336,021; 2,460,389)	(705,041; 3,091,369)
12	(1,146,993; 2,421,230)	(431,909; 3,136,315)
13	(787,881; 1,445,051)	(419,086; 1,813,846)
14	(1,287,874; 2,871,155)	(399,358; 3,759,672)
15	(1,106,365; 2,564,313)	(288,184; 3,382,494)
16	(952,044; 2,843,083)	(-109,181; 3,904,309)
17	(906,825; 1,497,666)	(575,253; 1,829,239)
18	(1,172,205; 2,627,273)	(355,641; 3,443,838)
19	(1,184,796; 2,565,896)	(409,741; 3,340,951)
20	(1,365,667; 3,012,852)	(441,290; 3,937,230)
21	(1,128,383; 1,905,100)	(692,500; 2,340,984)
22	(1,215,115; 2,568,651)	(455,528; 3,328,238)
23	(1,043,166; 1,597,679)	(731,981; 1,908,864)
24	(988,394; 2,898,551)	(-83,561; 3,970,505)

25	(1,201,447; 2,170,322)	(657,728; 2,714,041)
26	(1,242,834; 2,348,573)	(622,308; 2,969,098)
27	(991,929; 1,449,904)	(734,919; 1,706,913)
28	(900,505; 1,968,574)	(301,119; 2,567,960)
29	(1,266,636; 3,295,531)	(128,047; 4,434,120)
30	(981,221; 2,351,098)	(212,464; 3,119,855)

NOTE: Shaded intervals do not cover μ = 1,637,811.

4. Activity 3 - Hypothesis Test on the Population Mean

Prior to completing Activity 3, students should be familiar with how to perform hypothesis tests, including the calculation of test statistics and p-values. Each student needs the Activity 3 Worksheet (Appendix C), the Population Sheet, and a random number table or a calculator capable of generating random numbers. Each student selects a simple random sample of 25 metropolitan areas or uses the random sample selected for Activity 2. Note that a *t*-test procedure may be applied in this case, regardless of whether the population distribution is normal, since the sampling distribution of the sample mean is approximately normal for samples of size 25.

Students use their sample data to perform a hypothesis test of $H_0: \mu = 1{,}637{,}811$ versus $H_A: \mu \neq 1{,}637{,}811$ using a 10% level of significance. Since the true mean number of in-migrations for the population of metropolitan areas is 1,637,811 -- performing this test will provide an opportunity to illustrate how to interpret a Type I error rate.

Each student writes her calculated p-value on the white board on a class stem-and-leaf plot. Example class p-values are shown in Figure 1.

Figure 1. Stem-and-leaf Plot of Class P-Values.

```
Leaf Unit = 0.010 [0|5 is read .05, 3|4 is read .34]
```

```
0 | 2466 →    Type I Errors
1 | 4
2 |
3 | 3588
4 | 16
5 | 45
6 | 2234578
7 | 12246
8 | 9
9 | 1356
```

The p-values are calculated under the assumption that the null hypothesis is true, so the p-values will tend to be large. However, a few students will not obtain large p-values. On the stem-and-leaf plot, a cut-off value is marked at $\alpha = .10$. Each p-value falling at or below this cut-off represents a rejection of H_0 (a Type I error). Each p-value falling above this cut-off represents a failure to reject H_0 (a correct decision). Since many samples are taken, and many tests are performed, students can see that some samples result in a correct decision and other samples result in an incorrect decision. Students are asked to calculate the fraction of rejections of H_0 out of the tests performed and thus obtain a simulated value for α, the Type I error probability. Students are asked to interpret a Type I error rate in terms of repeatedly sampling, then using the sample data to test a true null hypothesis about a population parameter. For the example class data shown in Figure 1, the simulated Type I error rate is $4/30 \approx 13\%$.

The instructor may wish to discuss with students that in practice the true value of the population mean would not be known. If the population mean were known, then there would be no point in utilizing sample data to draw an inference about this parameter. The population mean is assumed known so that some of the properties of confidence intervals and hypothesis testing can be explored. An Answer Sheet for the Activity 3 questions is provided in Appendix E.

5. Conclusions

Activity 1 can be used by the instructor to introduce the concept of the sampling distribution of a sample mean. Active data collection helps reinforce the idea of repeated sampling with different samples producing different results. Changing the sample size allows students to formulate ideas about the mean and the variability of the distribution of a sample mean and to determine the relationship between sample size and variability.

The instructor can refer to this activity when discussing the distribution of a sample mean from a theoretical perspective. Students have examined empirical properties of the sampling distribution and will be ready to advance to a discussion of the theoretical properties.

Through completing Activity 2, students will see a demonstration of the process of repeatedly selecting samples from a population and constructing confidence intervals for the population mean. Plotting class results on the white board allows for a discussion of the meaning of the level of confidence. Students can visualize that, in repeated sampling, the confidence level represents the percentage of the time that the process of constructing a confidence interval will result in an interval that successfully encloses the true value of the population parameter. Changing the confidence level allows students to see the relationship between level of confidence and confidence interval width.

Activity 3 can be used to demonstrate the interpretation of Type I error rates. Students can see that the Type I error rate represents the fraction of incorrect decisions when repeatedly sampling, then using the sample data to test a true null hypothesis about a population parameter.

Appendix A Activity 1 Worksheet

Background: The Population Sheet displays a population of 178 large metropolitan areas. For each of the metropolitan areas, the number of residents in July 2005 is given. The true mean number of residents in the population is $\mu =$ 1,637,811. The true standard deviation of the number of residents in the population is $\sigma = 2{,}343{,}899$. If we did not know μ and wished to estimate it, we could draw a simple random sample of metropolitan areas from the population and use the mean number of residents for the sampled areas to estimate μ. The value of the sample mean, $\bar{x}$, will vary from sample to sample. The distribution of all of the $\bar{x}$ values for many simple random samples of size *n* is called the **sampling distribution of $\bar{x}$**.

Instructions: Work in groups. Each group should have a random number table or a calculator capable of generating random numbers.

1. Select two different simple random samples of size 5 from the population (sample with replacement -- so that it is possible to select the same metropolitan area more than once). For each sample calculate the value of the sample mean number of residents. Write your two sample mean values on the whiteboard under the heading 'Sample Size $n = 5$'. Once every group has written their sample mean values on the white board, complete the $n = 5$ column on the data table on the Data Collection Sheet. Take calculations to the nearest whole number.

Random Sample 1 $\bar{x} =$ Random Sample 2 $\bar{x} =$

2. Select two different simple random samples of size 15 from the population (sample with replacement). For each sample calculate the value of the sample mean number of residents. Write your two sample mean values on the whiteboard under the heading 'Sample Size $n = 15$'. Once every group has written their sample mean values on the white board, complete the $n = 15$ column on the data table on the Data Collection Sheet. Take calculations to the nearest whole number.

Random Sample 1 $\bar{x} =$ Random Sample 2 $\bar{x} =$

3. Select two different simple random samples of size 25 from the population (sample with replacement). For each sample calculate the value of the sample mean number of residents. Write your two sample mean values on the whiteboard under the heading 'Sample Size $n = 25$'. Once the entire class has finished with random samples of size *n* = 25, complete the $n = 25$ column on the data table on the Data Collection Sheet. Take calculations to the nearest whole number.

Random Sample 1 $\bar{x} =$ Random Sample 2 $\bar{x} =$

Answer the following questions using the data table on the Data Collection Sheet.

4. For each sample size n = 5, 15, and 25 construct a stem-and-leaf plot of the sample mean values.

5. For each sample size, describe the shape of the distribution of the sample mean values.

6. Based on your stem-and-leaf plots, what do you think is the relationship between the sample size and the shape of the distribution of the sample mean?

7.
(a) For each sample size, calculate the standard deviation and the mean of the sample means. Take calculations to the nearest whole number.

(b) For which sample size is the standard deviation the largest and for which sample size is the standard deviation the smallest?

8. How does the standard deviation of the sample mean values compare to the standard deviation of the population? What does this tell you about the spread of the sample mean values compared to the spread of the population values?

9. Find an expression for the mean of the sample means, $\mu_{\bar{X}}$, as a function of the mean of the population, μ.

10. Try to develop a formula to relate the standard deviation of the sample means, $\sigma_{\bar{X}}$, to the population standard deviation, σ, and the sample size, n. (Hint: the formula involves $\sqrt{n}$.)

Data Collection Sheet

Data Table. Class Sample Means

Sample Number	$n = 5$	$n = 15$	$n = 25$
1			
2			
3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

Appendix B   Activity 2 Worksheet

Statistical Guide: A $(1-\alpha)100\%$ confidence interval for μ, the population mean, is given by:
$\bar{x} \pm t_{n-1,\frac{\alpha}{2}}\left(\frac{s}{\sqrt{n}}\right)$ where $t_{n-1,\frac{\alpha}{2}}$ is an appropriate percentile of the t distribution having $(n-1)$ degrees of freedom, $\bar{x}$ is the sample mean, s is the sample standard deviation, and n is the sample size.

This interval gives a range of values within which we expect the population mean to fall. The procedure assumes that the data are a random sample from a normal population with unknown standard deviation, σ. If the sample size is large, the assumption of normality is not crucial. However, outliers are always a concern.

The sample mean is a point estimate for the population mean, μ. A **confidence interval estimate** for the population mean is an interval of values, computed from the sample data, that we believe contains μ. The **confidence level** is the probability that the **estimation method** will give an interval that contains the parameter (μ, in this case). The confidence level is denoted by $1-\alpha$, where common values of α are 0.10, 0.05, and 0.01, corresponding to 90%, 95%, and 99% confidence.

Instructions: The Population Sheet displays a population of 178 large metropolitan areas. For each of the metropolitan areas, the number of residents is given for July 2005. The true mean number of residents in the population is $\mu = 1{,}637{,}811$.

1. Select a simple random sample of 25 metropolitan areas (sample with replacement -- so that it is possible to select the same area more than once).

2. Calculate the mean and the standard deviation of the number of residents for your sampled metropolitan areas. Take calculations to the nearest whole number.

3. Use your data to construct an 80% confidence interval for the mean number of residents in the population of metropolitan areas. Write your confidence interval on the white board. On the white board, you will see all of the confidence intervals constructed by the class.
(a) How many of the confidence intervals include the value of $\mu = 1{,}637{,}811$? What percent is this?
(b) What percent of the confidence intervals did we expect to include the value of $\mu = 1{,}637{,}811$?
(c) Explain how to interpret a level of confidence of 80%.

4. Use your data to construct a 99% confidence interval for the mean number of residents in the population of metropolitan areas. Write your confidence interval on the white board. On the white board, you will see all of the confidence intervals constructed by the class.
(a) How many of the confidence intervals include the value of $\mu = 1{,}637{,}811$? What percent is this?

(b) What percent of the confidence intervals did we expect to include the value of $\mu = 1{,}637{,}811$?

(c) Explain how to interpret a level of confidence of 99%.

(d) Explain how increasing the confidence level from 80% to 99% changed the confidence intervals.

(e) Give one advantage of using 99% confidence rather than 80% confidence. Give one disadvantage.

Appendix C Activity 3 Worksheet

Statistical Guide: We want to test a hypothesis about a population mean, μ. The null hypothesis is $H_0 : \mu = \mu_0$, where μ_0 is the hypothesized value for μ. The data are assumed to be a random sample of size *n* from a population that has a normal distribution with unknown standard deviation, σ. If the sample size is large, the assumption of normality is not crucial. However, outliers are always a concern.

From the sample data, we calculate the sample mean, $\bar{x}$, and the sample standard deviation, *s*. We base our decision about μ on the standardized sample mean, $t_{n-1} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$. This is the test statistic, and under H_0 it has a *t*-distribution with $n-1$ degrees of freedom.

To make our decision, we calculate the **p-value** for the test. The p-value depends on how the alternative hypothesis is expressed:
(1) If $H_A : \mu > \mu_0$, then the p-value is the area to the right of the observed test statistic, under the H_0 model. (2) If $H_A : \mu < \mu_0$, then the p-value is the area to the left of the observed test statistic, under the H_0 model. (3) If $H_A : \mu \neq \mu_0$, then the p-value is the sum of the area to the left of negative the absolute value of the observed test statistic and the area to the right of the absolute value of the observed test statistic, under the H_0 model.

The **p-value** is the probability, computed under the assumption that H_0 is true, of obtaining a test statistic value at least as favorable to H_A as the value that actually resulted from the data. If the p-value is small enough, H_0 is rejected.

Rejecting the null hypothesis, when in fact it is *true*, is called a **Type I error**. The significance level, α, is the chance of committing a Type I error. If the p-value $\leq \alpha$, H_0 is rejected. If the p-value $> \alpha$, H_0 is not rejected. *Failing to reject* the null hypothesis, when in fact it is *not true*, is called a **Type II error**.

Instructions: The Population Sheet displays a population of 178 large metropolitan areas. For each of the metropolitan areas, the number of residents in July 2005 is given. The true mean number of residents in the population is $\mu = 1{,}637{,}811$.

1. Select a simple random sample of 25 metropolitan areas (sample with replacement -- so that it is possible to select the same area more than once).

2. Calculate the mean and the standard deviation of the number of residents for your sampled metropolitan areas.

3. Test $H_0 : \mu = 1{,}637{,}811$ versus $H_A : \mu \neq 1{,}637{,}811$.
Use a p-value to perform your test. Take calculations to two significant digits. Since H_0 is true, a correct decision would be to fail to reject H_0. An incorrect decision would be to reject H_0. (This would be a Type I error.) Use a 10% level of significance ($\alpha = .10$).

(a)
calculated test statistic =
p-value =
decision =
(Write your p-value on the white board.)
expected number of rejections of H_0 for the class =
number of rejections of H_0 for the class =

(b) Explain how to interpret a Type I error rate in terms of repeatedly performing the procedure of selecting a sample and using the sample data to test a null hypothesis that should not be rejected.

Appendix D Population Sheet

Label	Metropolitan Statistical Area	Population July 1, 2005
001	Akron, OH	702,235
002	Albany-Schenectady-Troy, NY	848,879
003	Albuquerque, NM	797,940
004	Allentown-Bethlehem-Easton, PA-NJ	790,535
005	Anchorage, AK	351,049
006	Ann Arbor, MI	341,847
007	Asheville, NC	392,831
008	Atlanta-Sandy Springs-Marietta, GA	4,917,717
009	Augusta-Richmond County, GA-SC	520,332
010	Austin-Round Rock, TX	1,452,529
011	Bakersfield, CA	756,825
012	Baltimore-Towson, MD	2,655,675
013	Baton Rouge, LA	733,802
014	Beaumont-Port Arthur, TX	383,530
015	Birmingham-Hoover, AL	1,090,126
016	Boise City-Nampa, ID	544,201
017	Boston-Cambridge-Quincy, MA-NH	4,411,835
018	Boston-Quincy, MA	1,800,432
019	Cambridge-Newton-Framingham, MA	1,459,011
020	Essex County, MA	738,301
022	Rockingham County-Strafford County, NH	414,091
022	Bridgeport-Stamford-Norwalk, CT	902,775
023	Brownsville-Harlingen, TX	378,311
024	Buffalo-Niagara Falls, NY	1,147,711
025	Canton-Massillon, OH	409,996
026	Cape Coral-Fort Myers, FL	544,758
027	Charleston, WV	306,435
028	Charleston-North Charleston, SC	594,899
029	Charlotte-Gastonia-Concord, NC-SC	1,521,278
030	Chattanooga, TN-GA	492,126
031	Chicago-Naperville-Joliet, IL-IN-WI	9,443,356
032	Chicago-Naperville-Joliet, IL	7,882,729
033	Gary, IN	697,401
034	Lake County-Kenosha County, IL-WI	863,226
035	Cincinnati-Middletown, OH-KY-IN	2,070,441
036	Cleveland-Elyria-Mentor, OH	2,126,318
037	Colorado Springs, CO	587,500
038	Columbia, SC	689,878
039	Columbus, OH	1,708,625
040	Corpus Christi, TX	413,553
041	Dallas-Fort Worth-Arlington, TX	5,819,475
042	Dallas-Plano-Irving, TX	3,893,123
043	Fort Worth-Arlington, TX	1,926,352
044	Davenport-Moline-Rock Island, IA-IL	376,309
045	Dayton, OH	843,577

046	Deltona-Daytona Beach-Ormond Beach, FL	490,055
047	Denver-Aurora, CO	2,359,994
048	Des Moines-West Des Moines, IA	522,454
049	Detroit-Warren-Livonia, MI	4,488,335
050	Detroit-Livonia-Dearborn, MI	1,998,217
051	Warren-Troy-Farmington Hills, MI	2,490,118
052	Durham, NC	456,187
053	El Paso, TX	721,598
054	Eugene-Springfield, OR	335,180
055	Evansville, IN-KY	349,543
056	Fayetteville, NC	345,536
057	Fayetteville-Springdale-Rogers, AR-MO	405,101
058	Flint, MI	443,883
059	Fort Wayne, IN	404,414
060	Fresno, CA	877,584
061	Grand Rapids-Wyoming, MI	771,185
062	Greensboro-High Point, NC	674,500
063	Greenville, SC	591,251
064	Harrisburg-Carlisle, PA	521,812
065	Hartford-West Hartford-East Hartford, CT	1,188,241
066	Hickory-Lenoir-Morganton, NC	355,654
067	Honolulu, HI	905,266
068	Houston-Sugar Land-Baytown, TX	5,280,077
069	Huntsville, AL	368,661
070	Indianapolis-Carmel, IN	1,640,591
071	Jackson, MS	522,580
072	Jacksonville, FL	1,248,371
073	Kalamazoo-Portage, MI	319,348
074	Kansas City, MO-KS	1,947,694
075	Killeen-Temple-Fort Hood, TX	351,528
076	Kingsport-Bristol-Bristol, TN-VA	301,294
077	Knoxville, TN	655,400
078	Lakeland, FL	542,912
079	Lancaster, PA	490,562
080	Lansing-East Lansing, MI	455,315
081	Las Vegas-Paradise, NV	1,710,551
082	Lexington-Fayette, KY	429,889
083	Little Rock-North Little Rock, AR	643,272
084	Los Angeles-Long Beach-Santa Ana, CA	12,923,547
085	Los Angeles-Long Beach-Glendale, CA	9,935,475
086	Santa Ana-Anaheim-Irvine, CA	2,988,072
087	Louisville-Jefferson County, KY-IN	1,208,452
088	Madison, WI	537,039
089	Manchester-Nashua, NH	401,291
090	McAllen-Edinburg-Mission, TX	678,275
091	Memphis, TN-MS-AR	1,260,905
092	Miami-Fort Lauderdale-Miami Beach, FL	5,422,200
093	Fort Lauderdale-Pompano Beach-Deerfield Beach, FL	1,777,638
094	Miami-Miami Beach-Kendall, FL	2,376,014
095	West Palm Beach-Boca Raton-Boynton Beach, FL	1,268,548
096	Milwaukee-Waukesha-West Allis, WI	1,512,855

097	Minneapolis-St. Paul-Bloomington, MN-WI	3,142,779
098	Mobile, AL	401,427
099	Modesto, CA	505,505
100	Montgomery, AL	357,244
101	Naples-Marco Island, FL	307,242
102	Nashville-Davidson--Murfreesboro, TN	1,422,544
103	New Haven-Milford, CT	846,766
104	New Orleans-Metairie-Kenner, LA	1,319,367
105	New York-Northern New Jersey-Long Island, NY-NJ-PA	18,747,320
106	Edison, NJ	2,303,709
107	Nassau-Suffolk, NY	2,808,064
108	Newark-Union, NJ-PA	2,152,978
109	New York-White Plains-Wayne, NY-NJ	11,482,569
110	Ogden-Clearfield, UT	486,842
111	Oklahoma City, OK	1,156,812
112	Omaha-Council Bluffs, NE-IA	813,170
113	Orlando-Kissimmee, FL	1,933,255
114	Oxnard-Thousand Oaks-Ventura, CA	796,106
115	Palm Bay-Melbourne-Titusville, FL"	531,250
116	Pensacola-Ferry Pass-Brent, FL	439,877
117	Peoria, IL	369,161
118	Philadelphia-Camden-Wilmington, PA-NJ-DE-MD	5,823,233
119	Camden, NJ	1,245,902
120	Philadelphia, PA	3,890,181
121	Wilmington, DE-MD-NJ	687,150
122	Phoenix-Mesa-Scottsdale, AZ	3,865,077
123	Pittsburgh, PA	2,386,074
124	Portland-South Portland-Biddeford, ME	514,227
125	Portland-Vancouver-Beaverton, OR-WA	2,095,861
126	Port St. Lucie-Fort Pierce, FL	381,033
127	Poughkeepsie-Newburgh-Middletown, NY	667,742
128	Providence-New Bedford-Fall River, RI-MA	1,622,520
129	Provo-Orem, UT	452,851
130	Raleigh-Cary, NC	949,681
131	Reading, PA	396,314
132	Reno-Sparks, NV	393,946
133	Richmond, VA	1,175,654
134	Riverside-San Bernardino-Ontario, CA	3,909,954
135	Rochester, NY	1,039,028
136	Rockford, IL	339,178
137	Sacramento--Arden-Arcade--Roseville, CA	2,042,283
138	St. Louis, MO-IL	2,778,518
139	Salem, OR	375,560
140	Salinas, CA	412,104
141	Salt Lake City, UT	1,034,484
142	San Antonio, TX	1,889,797
143	San Diego-Carlsbad-San Marcos, CA	2,933,462
144	San Francisco-Oakland-Fremont, CA	4,152,688
145	Oakland-Fremont-Hayward, CA	2,466,692
146	San Francisco-San Mateo-Redwood City, CA	1,685,996
147	San Jose-Sunnyvale-Santa Clara, CA	1,754,988

148	Santa Barbara-Santa Maria, CA	400,762
149	Santa Rosa-Petaluma, CA	466,477
150	Sarasota-Bradenton-Venice, FL	673,035
151	Savannah, GA	313,883
152	Scranton--Wilkes-Barre, PA	550,546
153	Seattle-Tacoma-Bellevue, WA	3,203,314
154	Seattle-Bellevue-Everett, WA	2,449,527
155	Tacoma, WA	753,787
156	Shreveport-Bossier City, LA	383,233
157	South Bend-Mishawaka, IN-MI	318,156
158	Spokane, WA	440,706
159	Springfield, MA	687,264
160	Springfield, MO	398,124
161	Stockton, CA	664,116
162	Syracuse, NY	651,763
163	Tallahassee, FL	334,886
164	Tampa-St. Petersburg-Clearwater, FL	2,647,658
165	Toledo, OH	656,696
166	Trenton-Ewing, NJ	366,256
167	Tucson, AZ	924,786
168	Tulsa, OK	887,715
169	Vallejo-Fairfield, CA	411,593
170	Virginia Beach-Norfolk-Newport News, VA-NC	1,647,346
171	Visalia-Porterville, CA	410,874
172	Washington-Arlington-Alexandria, DC-VA-MD-WV	5,214,666
173	Bethesda-Gaithersburg-Frederick, MD	1,148,284
174	Washington-Arlington-Alexandria, DC-VA-MD-WV	4,066,382
175	Wichita, KS	587,055
176	Wilmington, NC	315,144
177	Winston-Salem, NC	448,629
178	Worcester, MA	783,262

# Appendix E	Answers to Activity Questions

Note: answers to activity questions are based on the example class results shown in Table 1, Table 2, and Figure 1.

Activity 1.

Activity Questions.

4.

Sample Size of 5: Stem-and-Leaf Plot

```
Frequency    Stem &  Leaf

     6.00        0 .  567889
     6.00        1 .  122344
     3.00        1 .  589
     1.00        2 .  2
     1.00        2 .  9
     1.00        3 .  4
     1.00        3 .  8
     1.00 Extremes    (>=4152034)

 Stem width:  1000000
 Each leaf:       1 case(s)
```

Sample Size of 15: Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     1.00        0 .  8
     6.00        1 .  023344
     7.00        1 .  6777899
     3.00        2 .  012
     3.00        2 .  799

 Stem width:  1000000
 Each leaf:       1 case(s)
```

Sample Size of 25: Stem-and-Leaf Plot

```
 Frequency    Stem &  Leaf

     1.00        0 .  9
     5.00        1 .  23444
     9.00        1 .  557888899
     5.00        2 .  00124

 Stem width:  1000000
 Each leaf:       1 case(s)
```

5. For samples of size 5, the shape is somewhat skewed (to the right). The distribution is roughly mound-shaped for samples of size 15 and 25.

6. As the sample size increases, the distribution of the sample mean becomes more normal.

7.
(a)
n=5: mean = 1,729,088, standard deviation = 1,076,633
n=15: mean = 1,830,678, standard deviation = 596,130
n=25: mean = 1,754,398, standard deviation = 373,036
(b) The standard deviation of the $\bar{x}$ values is smallest for samples of size 25 and largest for samples of size 5. Averages of 25 numbers will be much less variable than averages of only 5 numbers.

8. The standard deviation of the $\bar{x}$ values is smaller than the standard deviation of the population. Therefore, the spread of the $\bar{x}$ values is less than the spread of the population values.

9. $\mu_{\bar{x}} = \mu = 1{,}637{,}811$

10. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{2{,}343{,}899}{\sqrt{25}} = 468{,}780$ (for the samples of size 25)

Activity 2.

Activity Questions.

3. (a) 27/30, or approximately 90%
(b) 80%
(c) If we repeatedly select random samples from a population and construct a confidence interval for the mean of the population using each selected sample, 80% of the confidence intervals will successfully enclose the true mean of the population and 20% will not.

4. (a) 30/30, or 100%
(b) 99%
(c) If we repeatedly select random samples from a population and construct a confidence interval for the mean of the population using each selected sample, 99% of the confidence intervals will successfully enclose the true mean of the population and 1% will not.
(d) The 99% confidence intervals are wider than the 80% confidence intervals.
(e) A 99% confidence interval is a highly reliable interval estimate but it may be imprecise. An 80% confidence interval is more precise but has relatively low reliability. The advantage of the 99% confidence interval is that it is very likely to enclose μ. The disadvantage of the 99% confidence interval is that it may be too wide to give us a good estimate of the value of μ.

Activity 3.

Activity Questions.

3.

(a) expected number of rejections of H_0 for the class = 10% of 30 = 3

number of rejections of H_0 for the class = 4 (or approximately 13%)

(b) If we repeatedly perform the procedure of selecting a sample and using the sample data to test a hypothesis about a population parameter, the Type I error rate is the percentage of the samples that would lead us to reject a true null hypothesis.

Acknowledgments

The author gratefully acknowledges the helpful comments and suggestions of colleagues during the preparation of this manuscript.

References

Data Source: *U.S. Census Bureau Public Information Office*: http://www.census.gov/.