

# 5: Introduction to Estimation

## Review Questions and Exercises

### Review Questions

1. Define “statistical inference.”
2. Name the two primary forms of statistical inference.
3. Name the two forms of statistical estimation.
4. Although parameters and estimators are related, they are *not* the same. List ways they differ.
5. What does it mean when we say that the sample mean is an unbiased estimator of  $\mu$ ?
6. A particular random sample of  $n$  observations can be used to calculate a sample mean. We can determine the characteristics of the distribution of means derived by other samples of the same size taken from the same populations under identical conditions without actually taking additional samples. The theorem that postulates that the sampling distribution of the mean (SDM) will tend to be normal (Gaussian) is called the central limit theorem. This SDM will have a mean of  $\mu$  and standard deviation that is equal to \_\_\_\_\_ [formula].
7. The standard deviation of the SDM is called the \_\_\_\_\_.
8. When the SDM is normal (Gaussian), 95% of the sample means will fall within 1.96 \_\_\_\_\_ of  $\mu$ .
9. True or false? The square root law says that the *SEM* is inversely proportional to the square root of the of the sample size.
10. Match the definitions below with each of the following terms: alpha, statistical inference, parameter, confidence interval, sampling distribution of the mean, standard error of the mean
  - (a) A numeric constant that describes something about a pmf, pdf, or statistical population.
  - (b) The act of generalizing from the data to a natural phenomenon or population with calculated degree of certainty
  - (c) The hypothetical frequency distribution of all possible sample means based on samples of size  $n$  taking from the same population under identical conditions.
  - (d) The chance researcher is willing to take in not capturing the parameter with a confidence interval.
  - (e) The standard deviation of the sampling distribution of the mean.
  - (f) An interval that is calculated so that it has known likelihood of capturing the parameter.
11. What percentage of 95% confidence intervals for  $\mu$  will *fail* to capture  $\mu$ ?
12. What percentage of (1-alpha)100% confidence intervals for  $\mu$  will fail to capture  $\mu$ ?
13. The CI for the mean *seeks* to capture the \_\_\_\_\_ mean. [M/C: (a) sample (b) population].
14. A confidence interval is  $3 \pm 1.2$ . The 3 in this statement is the \_\_\_\_\_ and the 1.2 is the \_\_\_\_\_.
15.  $\bar{x}$  is the point estimator of \_\_\_\_\_.
16. \_\_\_\_\_ is the point estimator of  $p$ .
17.  $s$  is the point estimator of \_\_\_\_\_.
18. The formula for calculating confidence intervals for  $p$  presented in this should only be used in large samples as tested by this “rule.”
19. What does SEP stand for? What does it measure? How do we estimate it (i.e., what’s the formula)?
20. The sample size requirements for estimating a mean with confidence depends on these things.
21. The sample size requirements for estimating a proportion with confidence depend on these things.
22. Increasing the sample size will \_\_\_\_\_ the precision of the estimate by \_\_\_\_\_ing the confidence interval length.
23. Increasing the confidence level of a CI will \_\_\_\_\_ the margin of error.

# Exercises

**5.1 Parameter or estimate?** Say whether each of the boldface numbers is either a parameter or an estimate. [Hint: Before providing an opinion, clearly state the implied research question.]

(A) A study of survival in 1225 newly diagnosed breast cancer cases finds that survival varies greatly by stage of diagnosis. The average seven-year survival rates for Stage I breast cancer was **92%**; the Stage II survival rate was **71%**; the stage III survival rate was **39%**; and the stage IV survival rate was **11%**.

(B) A review of divorce records for a county in Connecticut indicates that the marriages that end in divorce last an average of **72 months**.

(C) For U.S. men, the average life expectancy is **76**. For women, it's **81**. [Life expectancy is based on data from the National Death Index; deaths are uniformly recorded in the U.S.]

**5.2 Sampling experiment.** Let us conduct a small, artificial sampling experiment that will demonstrate what is meant by the sampling distribution for a mean (SDM). A tiny finite population consists of the following values: {1, 3, 5, 7, 9}. This population has  $\mu = 5$  and  $\sigma = 2.828$ .

(A) Construct a stemplot of the population.

(B) List of all possible unique samples each of  $n = 2$  from this population. For example, the first sample can be {1, 3}. The second such sample can be {1, 5}. There 10 such unique samples.

(C) Calculate the mean of each of the 10 samples.

(D) Construct a stemplot of the 10 sample means. This is an experimental sampling distribution of the mean (SDM).

(E) Is the SDM more or less normal shaped than the population? This is a demonstration of the **central limit theorem**, which states that the SDM tend to be more normal than the population. This is particularly true with when the sample size is large. Although the sample size in this experiment is small, you should still a *trend* toward the normalization of the SDM.

(F) Calculate the mean of the SDM (i.e., the mean of the means). Is the mean of the SDM equal to, less than, or greater than the mean of the population? This is a demonstration of the unbiasedness of the  $\bar{x}$  as a reflection of  $\mu$ .

(G) Use your calculator to calculate the standard deviation of the SDM (i.e., the standard deviation of the means is analogous to the *SEM*). Report both the  $s$  and  $\sigma$  derived by your calculator. Are these standard deviations equal to, less than, or greater than  $\sigma$  in population (which is 2.828)? This demonstrates that the SEM is smaller than the  $\sigma$ .

**5.3 Health survey.** A survey takes a simple random sample of 500 people from a town of 55,000. On the average, it finds 2.30 health problems per person (standard deviation = 1.65). **Say whether each of the following statements is true or false. Explain your reasoning in each instance.**

(A) The standard deviation of the mean (i.e., the *SEM*) is 0.0738.

(B) The 95% confidence interval for the average number of health problems in the sample is (2.16, 2.44).

(C) The 95% confidence interval for the average number of health problems in the town is (2.16, 2.44).

(D) While the number of health problems in the population is not normally distributed, according to the central limit theorem (Exercise 5.2E) it reasonable to assume that the sampling distribution of the mean (SDM) will be normal.

**5.4 Lab instrument.** Measurements from a particular lab instrument has known standard deviation  $\sigma = 10$ . A lab assistant takes 4 measurements with this instrument and calculates the mean of the four samples.

(A) Explain the advantage using the average of the four measurements rather than each individual measurement as the true value for a sample.

(B) Calculate the standard error of the mean when  $n = 4$ .

(C) How many times must the assistant repeat the measurement to reduce the standard error of the mean to 2.5? [Hint: Rearrange the formula  $SEM = \sigma / \sqrt{n}$  to get  $n = (\sigma / SEM)^2$  to determine the required sample size].

**5.5 Pharmacy survey.** A survey of 30 pharmacies found that the average cost of a month's supply of a particular drug was \$33. The standard deviation of the cost of the prescription is assumed to be \$7.  
(A) Based on the data, determine the mean price  $\mu$  with 90% confidence for the Rx from all pharmacies.  
(B) Now determine  $\mu$  with 95% confidence.  
(C) What is the margin of error for 90% confidence? What is the margin of error for 95% confidence?  
(D) A pharmacist reads that a 95% confidence interval for the average price is \$30.50 to \$35.50. Asked to explain the meaning of this, the pharmacist states "95% of all pharmacies sell the drug for between \$30.50 and \$35.50." Is the pharmacist correct? Explain your response.

**5.6 Laboratory scale.** A manufacturer of a laboratory scale with a digital readout claims the scale is accurate to 0.0015 of a gram. You read the fine print in the documentation that accompanies the scale and find that the manufacturer means to say that the weights will have standard deviation  $\sigma = 0.0015$  grams. You are willing to assume that the measurement error varies according to a normal pdf with a mean equal to the true weight  $\mu$  of an object. Duplicate measurements on an object produce weights of 24.31 and 24.34 grams. Estimate the true weight of the objects with 95% confidence.

**5.7 Hemoglobin survey, sample size.** Hemoglobin levels in 11-year old boys have a normal distribution with unknown mean  $\mu$  and  $\sigma = 1.209$  g/dl. How large a sample is needed to estimate  $\mu$  with 95% confidence and a margin of error of 0.5?

**5.8 Sugar consumption survey, sample size.** A nutritionist is willing to assume that the standard deviation of the weekly sugar consumption in children is 100 grams. How large a sample is needed to calculate a 95% confidence interval for  $\mu$  so that its margin of error is no greater than 10 grams?

**5.9 Birth weights.** Assume birthweights in a population of full-term infants vary normally with standard deviation  $\sigma = 2$  pounds. Random samples of size  $n$  are selected from this population of birth weights of full term infants. Calculate 95% confidence intervals for  $\mu$  based on each of the following samples. State the margin of error in each instance.  
(A)  $n = 81$ , sample mean = 6.2 pounds  
(B)  $n = 36$ , sample mean = 7.0 pounds  
(C)  $n = 9$ , sample mean = 5.8 pounds

**5.10 AIDS risk factor study.** A national survey of AIDS risk factors based on a random sample of 2673 heterosexual adults found the 170 individuals reported two or more sexual partners in the past 12 months. (Assume the responses are truthful.)  
(A) Describe the population to which inferences will be made. Describe the parameter to be estimated.  
(B) Calculate the point estimate for the prevalence of this risk factor.  
(C) What is the parameter of interest? Can the large sample method be used to calculate confidence intervals for the prevalence of this risk factor?  
(D) Calculate a 95% CI for the prevalence of multiple sexual partners in this population.  
(E) Calculate a 90% CI for the prevalence.

**5.11 Patient preference.** A doctor finds that 7 out of 8 of her patients preferred to have a "finger stick" blood sample compared to a venous blood draw. Can you use the large sample method to determine the proportion of patients preferring the finger-stick method with 95% confidence using this data? Justify your response with the "npq rule."

**5.12. Sample size for a proportion.** You are planning a study that wants to estimate a population proportion with 95% confidence. How many individuals do you need to study to achieve a margin of error of no greater than .06? A reasonable estimate for the population proportion is not available before the study is begun so you assume  $p^* = 0.50$ . How large a sample is needed to cut the margin of error to .03?

**5.13 Graduate student age.** The age distribution of students in a graduate program is approximately normally distributed with unknown mean  $\mu$  and standard deviation  $\sigma = 5$ . You sample 24 individuals from this population and find a sample mean of 25.0. Calculate the 95% confidence interval for mean age in the study population  $\mu$  based upon mean in your sample.

**5.14 Muscle strength scores.** A physical therapist studying muscular strength assumed muscle strength scores are normally distributed with  $\sigma = 12$ . A sample of 15 individuals demonstrates a mean muscular strength score of 84.3. Calculate a 95% confidence interval for  $\mu$ . Interpret the meaning of this confidence interval. [Similar to Daniel, 1999, p. 157.]

**5.15 Antigen titer.** A vaccine manufacturer analyzes samples a production batch of vaccine to check up on the concentration of its titer. Immunologic analyses are not perfect, so she repeats measurements on the same batch getting slightly different results each time. The public health scientist assumes that repeated measurements will vary according to a normal distribution with mean  $\mu$  and  $\sigma = 0.070$ . (The standard deviation is a characteristic of the titering technique as reported by the manufacturer.) Three ( $n = 3$ ) measurements reveal the following titers: {17.40, 17.36, 17.45}. Calculate a 95% confidence interval for the true concentration  $\mu$ .

**5.16 SIDS.** A study of 49 sudden infant death syndrome (SIDS) cases derives a mean birth weight of 2998 grams. From a listing of all birth weight, it is known that the standard deviation  $\sigma$  of birth weight in this population is 800 grams. Calculate a 95% confidence interval for the mean  $\mu$  birth weight of SIDS cases in the population. Interpret your results.