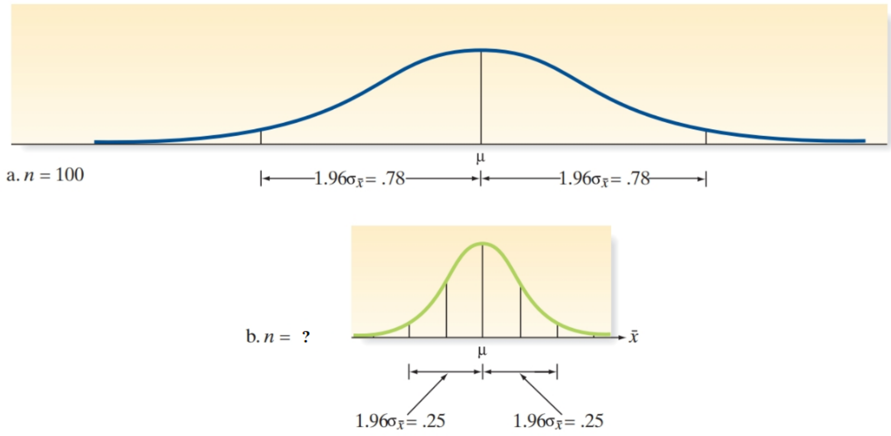
# Math 216-001 Statistical Thinkings
# Note 18: Determining the Sample Size - Estimating a Population Mean

**Instructor: Dr. Jiacheng Cai** **Date: 03/16/2022**

## Section 7.5: Determining the Sample Size - Estimating a Population Mean

Recall from Section 1.5 that one way to collect the relevant data for a study used to make inferences about a population is to implement a designed (planned) experiment. Perhaps the most important design decision faced by the analyst is to determine the size of the sample. We show in this section that the appropriate sample size for making an inference about a population mean depends on the desired reliability.

Consider Example 1 and 2 in the previous section, in which we estimated the mean length of stay for patients in a large hospital. A sample of 100 patients' records produced the 95% confidence interval $\bar{x} \pm 1.96\sigma_{\bar{x}} = 4.5 \pm .78$. Consequently, our estimate $\bar{x}$ was within .78 day of the true mean length of stay, $\mu$, for all the hospital's patients at the 95% confidence level. That is, the 95% confidence interval for $\mu$ was $2(.78) = 1.56$ days wide when 100 accounts were sampled.

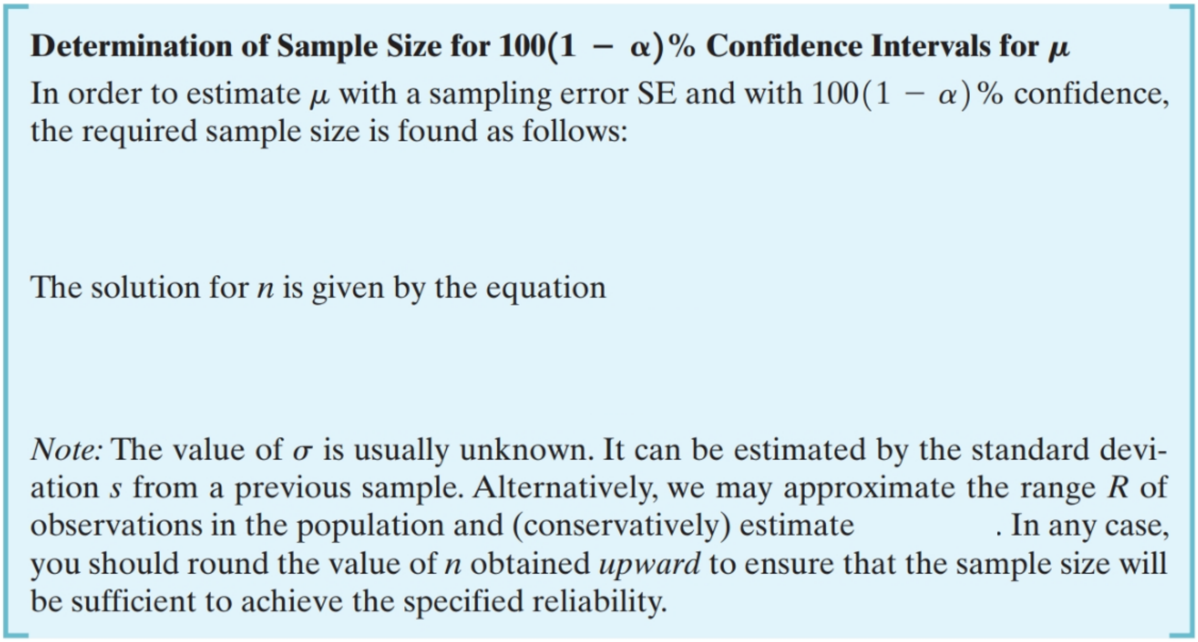

a. $n = 100$
$\mu$
$1.96\sigma_{\bar{x}} = .78$
$1.96\sigma_{\bar{x}} = .78$
b. $n =$ ?
$\bar{x}$
$\mu$
$1.96\sigma_{\bar{x}} = .25$
$1.96\sigma_{\bar{x}} = .25$

Now suppose we want to estimate $\mu$ to within .25 day with 95% confidence. That is, we want to narrow the width of the confidence interval from 1.56 days to .50 day, as the above figure (b). How much will the sample size have to be increased to accomplish this?

Consequently, over patients' records will have to be sampled to estimate the mean length of stay, $\mu$, to within .25 day with (approximately) 95% confidence. The confidence interval resulting from a sample of this size will be approximately .50 day wide.

In general, we express the reliability associated with a confidence interval for the population mean $\mu$ by specifying the **sampling error** within which we want to estimate $\mu$ with $100(1-a)\%$ confidence. The sampling error (denoted SE) is then equal to the half-width of the confidence interval

The procedure for finding the sample size necessary to estimate $\mu$ with a specific sampling error is given in the following box. Note that if $\sigma$ is unknown (as is usually the case, in practice), you will need to estimate the value of $\sigma$.



**Determination of Sample Size for $100(1-\alpha)\%$ Confidence Intervals for $\mu$**

In order to estimate $\mu$ with a sampling error SE and with $100(1-\alpha)\%$ confidence, the required sample size is found as follows:

The solution for $n$ is given by the equation

*Note:* The value of $\sigma$ is usually unknown. It can be estimated by the standard deviation $s$ from a previous sample. Alternatively, we may approximate the range $R$ of observations in the population and (conservatively) estimate . In any case, you should round the value of $n$ obtained *upward* to ensure that the sample size will be sufficient to achieve the specified reliability.

Sometimes the formula will lead to a solution that indicates a small sample size is sufficient to achieve the confidence interval goal. Unfortunately, the procedures and assumptions for small samples differ from those for large samples. Therefore, if the formulas yield a small sample size, one simple strategy is to select a sample size n = 30.

**Example 1** Suppose the manufacturer of official NFL footballs uses a machine to inflate the new balls to a pressure of 13.5 pounds. When the machine is properly calibrated, the mean inflation pressure is 13.5 pounds, but uncontrollable factors cause the pressures of individual footballs to vary randomly from about 13.3 to 13.7 pounds. For quality control purposes, the manufacturer wishes to estimate the mean inflation pressure to within .025 pound of its true value with a 99% confidence interval. What sample size should be specified for the experiment?

## Distinguish Between Confidence and Probability

Consider the following example, a 99% confidence interval for the population mean weight μ was computed to be $20.04 < \mu < 20.46$. It is tempting to say that the probability is 99% that μ is between 20.04 and 20.46. This, however, is not correct. The term probability refers to random events, which can come out differently when experiments are repeated. The numbers 20.04 and 20.46 are fixed, not random. The population mean is also fixed. The population mean weight is either between 20.04 and 20.46 or it is not. There is no randomness involved. Therefore, we say that we have 99% confidence (not probability) that the population mean is in this interval.

On the other hand, let's say that we are discussing a method used to construct a 99% confidence interval. The method will succeed in covering the population mean 99% of the time, and fail the other 1% of the time. In this case, whether the population mean is covered or not is a random event, because it can vary from experiment to experiment. Therefore it is correct to say that a method for constructing a 99% confidence interval has probability 99% of covering the population mean.

In summary, we have the following

1. C.I. is for the **population mean $\boldsymbol{\mu}$**, not for the **sample mean $\bar{\boldsymbol{x}}$**
2. We can say we are xx% confidence that $\mu$ falls in a specific C.I. We can say in repeated sampling, xx% of similarly constructed intervals contain the value of the population mean $\mu$.
3. However we can **NOT** say that, the probability that $\mu$ falls in a specific C.I. is xx%

**Example 2** A hospital administrator draw a simple random sample of 100 records of patients who were admitted for cardiac bypass surgery. She compute the sample mean number of days spent in the hospital, and construct a 95% confidence interval for the mean, which is (7.1, 7.5). Which of the following statement is correct?

a. 95% of the patients spent between 7.1 days and 7.5 days in the hospital

b. We are 95% confident that the mean number of days spent in hospital of all patients falls between 7.1 and 7.5

c. We are 95% confident that the mean number of days spent in hospital of the sampled patients falls between 7.1 and 7.5

d. The probability that a patient stay 7.1 to 7.5 days is .95

Reference: W. Navidi, B. Monk, Elementary Statistics, 3th Edition, McGraw-Hill Education 2019