Sampling Distributions

The sampling distribution of a sample function, say $\bar{x}$, calculated from a random sample of size $n$ is simply the probability distribution of $\bar{x}$ (obtained from all possible samples of $n$ observations from a population with mean $\mu$ and variance $\sigma^2$).

Central Limit Theorem.

If $n$ is large, the sampling distribution of $\bar{x}$ will be approximately a normal distribution with mean $\mu_{\bar{x}} = \mu$ and standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$. The standard deviation $\sigma_{\bar{x}}$ is often referred to as the standard error of the sample mean $\bar{x}$. 
Example.

According to AAA, the average daily meal and lodging costs for a family of four is $213. Assume that the standard deviation of such cost is $15. Consider a random sample of 36 families of four and their travel expenses. Find the probability that sample mean exceeds $200.

Example.

The weight of corn chips dispensed into a 10-ounce bag by the dispensing machine has been identified as possessing a normal distribution with a mean of 10.5 ounces and a standard deviation of .2 ounces. Suppose 100 bags of chips were randomly selected from this dispensing machine. Find the probability that the sample mean weight of these 100 bags exceeded 10.45 ounces.

Example.

The amount of time it takes a student to walk from her home to class has a skewed right distribution with a mean of 16 minutes and a standard deviation of 1.6 minutes. If data were collected from 36 randomly selected walks, describe the sampling distribution of \(\bar{x}\), the sample mean time.
True or False: The Central Limit Theorem guarantees that the population is normal whenever \( n \) is sufficiently large.

True or False: The Central Limit Theorem guarantees an approximately normal sampling distribution for the sample mean for large sample sizes, so no knowledge about the distribution of the population is necessary for large-sample confidence intervals to be valid.
A 100(1 - \alpha) % confidence interval for \mu is a formula that gives an interval for each sample of \( n \) observations with the property that, in repeated sampling of \( n \) observations, 100(1-\alpha) % of all intervals enclose \( \mu \). In other words, the formula is 100(1-\alpha)% accurate in the sense that 100(1-\alpha)% of the time, in repeated sampling, the formula gives intervals that enclose \( \mu \).

The fraction \( (1- \alpha) \) is called confidence coefficient and 100(1-\alpha)% is often referred to as confidence level.

1. Large Sample 100(1-\alpha)% Confidence Interval for \mu:

\[
L = \bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \quad U = \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}
\]

Here \( Z_{\alpha/2} \) is obtained from standard normal distribution table so that the area to the left of \( Z_{\alpha/2} \) is equal to 1 - \( \alpha/2 \). If the population standard deviation \( \sigma \) is unknown, it is replaced by \( s \), the sample standard deviation to obtain an approximate confidence interval.
Example. How much money does the average professional football fan spend on food at a single football game? The question was posed to 40 randomly selected football fans. The sampled results show that the sample mean and standard deviation were $\bar{x} = $52 and $s =$17.50, respectively.

(a) Find a 95% confidence interval for the true mean $\mu$.

(b) Which of the following interpretations is correct for your interval in (a)?

(i) 95% of the population values will fall in the interval.

(ii) 95% of the similarly constructed intervals would contain the value of the sample mean.

(iii) The probability that the population mean falls in any confidence interval constructed is 0.95.

(iv) In repeated sampling, 95% of the intervals constructed would contain $\mu$.

Example. Each in a sample of 65 low-income children was administered the Communicative Development Inventory (CDI) exam. The sentence complexity scores had a mean of 7.62 and a standard deviation of 8.91. Construct a 90% confidence interval for the mean sentence complexity score of all low-income children.
2. Small Sample 100(1-\(\alpha\))% Confidence Interval for \(\mu\) (assuming that the random sample is selected from normal distribution):

\[
L = \bar{x} - (t_{\alpha/2,n-1}) \frac{s}{\sqrt{n}}, \quad U = \bar{x} + (t_{\alpha/2,n-1}) \frac{s}{\sqrt{n}}
\]

Here \(t_{\alpha/2,n-1}\) is obtained from the t-distribution table so that the area to the right hand side of \(t_{\alpha/2,n-1}\) is \(\alpha/2\). The formula requires that the sampled population be normal.

Example. A meteorologist wishes to estimate the mean amount of snowfall per year in Spokane, Washington. A random sample of the recorded snowfall for 20 years produces a sample mean equal to 54 inches and standard deviation of 9 inches. Estimate the true mean amount of snowfall in Spokane using a 90% confidence interval.

Example. Pulse rate is an important measure of the fitness of a person’s cardiovascular system. A random sample of 5 U.S adult males who jog at least 15 miles per week had had the following pulse rates per minutes.

\[
54 \quad 50.5 \quad 50.8 \quad 53 \quad 52.5
\]

\(\bar{x} = 52.16, s^2 = 2.203\)

Find a 95% confidence interval for the mean pulse rate of all U.S. adult males who jog at least 15 miles per week.
3. Large Sample 100(1-α)% Confidence Interval for the binomial proportion \( p \).

\[
L = \hat{p} - Z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \quad U = \hat{p} + Z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

Here \( Z_{\alpha/2} \) is obtained from standard normal distribution table so that the area to the left of \( Z_{\alpha/2} \) is equal to \( 1 - \alpha/2 \).

Note. The sample size \( n \) is considered large if \( \hat{p} \pm 3 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \) falls between 0 and 1.

Example. A university dean is interested in determining the proportion of students who receive some sort of financial aid. Rather than examine the records for all students, the dean randomly selects 200 students and finds that 118 of them are receiving financial aid. Use a 90% confidence interval to estimate the true proportion of students on financial aid.

Example. Suppose that in a random sample of 200 Americans, 85 were victims of a crime. Estimate the true proportion of Americans who were victims of a crime using a 95% confidence confidence interval.
Sample Size Calculations

4. Sample Size required to estimate $\mu$ with sampling error $e$ (i.e., bound on the error of estimation) and $100(1 - \alpha)\%$ confidence interval:

$$n = \frac{(Z_{\alpha/2})^2\sigma^2}{e^2}.$$  

Here $\sigma$ is often estimated by Range/4 if no prior estimate of $\sigma$ is available.

Example. Suppose you wish to estimate the mean $\mu$ of a population correct to within a bound $B = 0.2$ with probability equal to 0.95. The variance of the population is approximately equal to 16.0. Find the approximate sample size that will produce the desired accuracy of the estimate.

Example. As an aid in the establishment of personnel requirements, the director of a hospital wishes to estimate the mean number of people who are admitted to the emergency room during a 24-hour period. If the director wishes to estimate the mean number of admissions per 24-hour period to within 1 admission with 95% confidence, what size sample should she choose? It is known that the variance of number of people who are admitted to the emergency room during each 24-hour period is approximately 25.
5. Sample Size required to estimate the binomial proportion \( p \) with sampling error \( e \) (i.e., bound on the error of estimation) and \( 100(1 - \alpha)\% \) confidence interval:

\[
n = \frac{(Z_{\alpha/2})^2 p(1-p)}{e^2}.
\]

Use 0.5 for \( p \) if no prior estimate of \( p \) is available.

Example. A university dean is interested in determining the proportion of students who receive some sort of financial aid. If the dean wanted to estimate the proportion of all students receiving financial aid to within 3\% with 95\% confidence, how many students would need to be sampled?
Review Exercises

1. A local men’s clothing store is being sold. The buyers are trying to estimate the percentage of items that are outdated. They will randomly sample among its 100,000 items in order to determine the proportion of merchandise that is outdated. The current owners have never determined their outdated percentage and can not help the buyers. Approximately how large a sample do the buyers need in order to insure that they are 99% confident that the margin of error is within 4%?

2. A university is considering a change in the way students pay for their education. Presently, the students pay $16 per credit hour. The university is contemplating charging each student a set fee of $240 per quarter, regardless of how many credit hours each takes. To see if this proposal would be economically feasible, the university would like to know how many credit hours, on the average, each student takes per quarter. A random sample of 250 students yields a mean of 14.1 credit hours per quarter and a standard deviation of 2.8 credit hours per quarter. Suppose the administration wanted to estimate the mean to within 0.3 credit hours at 95% reliability. How large a sample would they need to take?
3. An article in a Florida newspaper reports on the topics that teenagers most want to discuss with their parents. The findings, the results of a poll, showed that 30% would like to talk about religion. This percentage was based on a national sampling of 505 teenagers. Estimate the proportion of all teenagers who want more family discussions about religion. Use a 90% confidence level.

4. The increasing cost of health care is an important issue today. Suppose that a random sample of 25 small companies that offer paid health insurance as a benefit was selected. The mean health insurance cost per worker per month was $124, and the standard deviation was $30. Construct a 95% confidence interval for the average health cost per worker per month for all small companies.
For the question below, answer True or False

1) As the sample size taken gets larger, the standard error of the mean gets larger as well.

2) The standard error of the mean is equal to sigma, the standard deviation of the population.

3) One way of reducing the width of a confidence interval is to reduce the confidence level.

4) One way of reducing the width of a confidence interval is to reduce the size of the sample taken.

5) If no estimate of $p$ exists when determining the sample size, we can use $.5$ in the formula to get a value for $n$. 
Elements of a Test of Hypothesis

1. Null Hypothesis ($H_0$) - A statement about the values of population parameters which we accept until proven false.

2. Alternative or Research Hypothesis ($H_a$)- A statement that contradicts the null hypothesis. It represents researcher’s claim about the population parameters. This will be accepted only when data provides sufficient evidence to establish its truth.

3. Test Statistic - A sample statistic (often a formula) that is used to decide whether to reject $H_0$.

4. Rejection Region- It consists of all values of the test statistic for which $H_0$ is rejected. This rejection region is selected in such a way that the probability of rejecting true $H_0$ is equal to $\alpha$ (a small number usually 0.05). The value of $\alpha$ is referred to as the level of significance of the test.

5. Assumptions - Statements about the population(s) being sampled.

6. Calculation of the test statistic and conclusion- Reject $H_0$ if the calculated value of the test statistic falls in the rejection region. Otherwise, do not reject $H_0$.

7. P-value or significance probability is defined as proportion of samples that would be unfavourable to $H_0$ (assuming $H_0$ is true) if the observed sample is considered unfavourable to $H_0$. If the p-value is smaller than $\alpha$, then reject $H_0$. 
Remark:

1. If you fix $\alpha = 0.05$ for your test, then you are allowed to reject true null hypothesis 5% of the time in repeated application of your test rule.

2. If the p-value of a test is 0.20 (say) and you reject $H_0$ then, under your test rule, 20% of the time you would reject true null hypothesis.
1. Large sample \((n > 30)\) test for \(H_0 : \mu = \mu_0\) (known).

\[
Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}
\]

Example. A study reported in the *Journal of Occupational and Organizational Psychology* investigated the relationship of employment status to mental health. Each of a sample of 49 unemployed men was given a mental health examination using the General Health Questionnaire (GHQ). The GHQ is widely recognized measure of present mental health, with lower values indicating better mental health. The mean and standard deviation of the GHQ scores were \(\bar{x} = 10.94\) and \(s = 5.10\), respectively.

(a). Specify the appropriate null and alternative hypothesis if we wish to test the research hypothesis that the mean GHQ score for all unemployed men exceeds 10.

Is the test one-tailed or two-tailed?

(b). If we specify \(\alpha = 0.05\), what is the appropriate rejection region for this test?

(c). Conduct the test, and state your conclusion clearly in the language of this exercise.

Find the p-value of the test.
Example. A consumer protection group is concerned that a ketchup manufacturer is filling its 20-ounce family-size containers with less than 20 ounces of ketchup. The group purchases 49 family-size bottles of this ketchup, weigh the contents of each, and finds that the mean weight is 19.86 ounces, and the standard deviation is equal to 0.22 ounces.

(a). Do the data provide sufficient evidence for the consumer group to conclude that the mean fill per family-size bottle is less than 20 ounces? Test using $\alpha = 0.05$.

(b). Find the $p$-value of the test in part (a).
Example. State University uses thousands of fluorescent light bulbs each year. The brand of bulb it currently uses has a mean life of 900 hours. A manufacturer claims that its new brands of bulbs, which cost the same as the brand the university currently uses, has a mean life of more than 900 hours. The university has decided to purchase the new brand if, when tested, the test evidence supports the manufacturer’s claim at the .10 significance level. Suppose 99 bulbs were tested with the following results: \( \bar{x} = 919 \) hours, \( s = 86 \) hours. Find the rejection region for the test of interest to the State University.
2. Small sample \((n \leq 30)\) test for \(H_0 : \mu = \mu_0\) (known).

\[
t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}
\]

This test requires that the sampled population is normal.

Example. A random sample of \(n\) observations is selected from a normal population to test the null hypothesis that \(\mu = 10\). Specify the rejection region for each of the following combinations of \(H_a\), \(\alpha\), and \(n\).

(a). \(H_a : \mu \neq 10\), \(\alpha = 0.01\), \(n = 14\).

(b). \(H_a : \mu < 10\), \(\alpha = 0.025\), \(n = 26\).
Example. According to advertisements, a strain of soybeans planted on soil prepared with a specified fertilizer treatment has a mean yield of 475 bushels per acre. Twenty farmers who belong to a cooperative plant the soybeans. Each uses a 40-acre plot and records the mean yield per acre. The mean and variance for the sample of 20 farms are $\bar{x} = 462$ and $s^2 = 9070$. Specify the null and alternative hypothesis used to determine if the mean yield for the soybeans is different than advertised.
Example. A psychologist was interested in knowing whether male heroin addicts’ assessments of self-worth differ from those of the general male population. On a test designed to measure assessment of self-worth, the mean score for males from the general population was found to be equal to 48.6. A random sample of 25 scores achieved by heroin addicts yielded a mean of 44.1 and a standard deviation of 6.2. Do the data indicate a difference in assessment of self-worth between male heroin addicts and general male population? Test using $\alpha = 0.01$. 


3. Large sample test for $H_0 : p = p_0$ (known).

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

For this test, sample size is considered large if $p_0 \pm 3\sqrt{\frac{p_0(1-p_0)}{n}}$ falls between 0 and 1.

Example. The National Science Foundation, in a survey of 2,237 engineering graduate students who earned their Ph.D. degrees, found that 607 were U.S. citizens; the majority (1,630) of the Ph.D degrees were awarded to foreign nationals. Conduct a test to determine whether the true percentage of engineering Ph.D. degrees awarded to foreign nationals exceeds 50%. Use $\alpha = 0.01$. 


Example. The business college computing center wants to determine the proportion of business students who have personal computers (PC’s) at home. If the proportion exceeds 30 percent, then the lab will scale back a proposed enlargement of its facilities. Suppose 250 business students were randomly sampled and 85 have personal computers at home. Conduct a test to see if the scale back of the proposed enlargement of its facilities is needed. Use $\alpha = 0.05$. 
Example. A method currently used by doctors to screen women for possible breast cancer fails to detect cancer in 15% of the women who actually have the disease. A new method has been developed that researchers hope will be able to detect cancer more accurately. A random sample of 70 women known to have breast cancer were screened using the new method. Of these, the new method failed to detect cancer in six. Specify the null and alternative hypothesis that the researchers wish to test. Calculate the test statistic, determine the rejection region if $\alpha = 0.05$, find the $p$-value, and state the conclusion clearly in the language of this exercise.
Example. The Midwest Organization of Retired Oncologists and Neurologists (M.O.R.O.N.) has recently taken flack from some of its members regarding the poor choice of the organization’s name. The association bylaws require that more than 60% of the organization must approve a name change. Rather than convene a meeting, it is first desired to use a sample to determine if a meeting is necessary. A random sample of 60 of M.O.R.O.N.’s members were asked if they want M.O.R.O.N. to change its name. Forty-five of the respondent’s said ”yes.” Find the p-value for the desired test of hypothesis.
Example. Increasing numbers of businesses are offering child-care benefits for their workers. However, one union claims that more than 80% of firms in the manufacturing sector still do not offer any child-care benefits to their workers. A random sample of 480 manufacturing firms is selected, and only 27 of them offer child-care benefits. Specify the rejection region that the union will use when testing at alpha = .05. Suppose the p-value for this test was reported to be p = .1113. State the conclusion of interest to the union. Use alpha = .10.