Investigating Errors and Power in Significance Tests for Means (when σ is known)

A packing machine is set to fill packages so that package weights are normally distributed with mean 21 ounces and standard deviation 0.5 ounces. To monitor the accuracy of the machine, four packages are periodically selected at random, weighed together, and the mean weight is determined. It has been decided that the machine settings should be readjusted if the sample mean is significantly greater than 21 ounces at the 10% level (α=.10). To determine whether or not readjustments are needed, we would perform a significance test with hypotheses H_0: $\mu = 21$ and H_a: $\mu > 21$ and reject H_0 if the *P*-value associated with the test is less than or equal to .10.

Rather than conducting a test each time a sample of four packages is taken, we could determine in advance how large the sample mean would need to be to reject H_0 (and consequently readjust the machine). To determine what values of sample means would lead to rejection, you need to know the sampling distribution of $\bar{x}$. Since n=4 and the underlying population is N(21, 0.5), the sampling distribution of $\bar{x}$ is $N(21,\ 0.5/\sqrt{4})$.

1. Determine how large the sample mean must be to cause a readjustment when testing at the α=.10 level.

Tests of significance sometimes result in incorrect conclusions. One way that a wrong conclusion could be made would be to reject the null hypothesis when in fact it is true. This is called a **Type I** error. Clearly there could be samples drawn from the population specified by the null hypothesis that have mean values in the rejection region. If we were unlucky and selected a sample for which this happened, we would decide to reject H_0, an incorrect conclusion. For a test performed at significance level α, the probability that this will occur is α.

We can use simulations to verify that the probability of making a Type I error is .10 when a significance test of $H_0\!:\mu = 21$ versus $H_a : \mu > 21$ is performed at the α=.10 level.

Go to the following web site: http://wise.cgu.edu/power/power_applet.asp. Change values of μ_0 and μ_1 to 21 so that both are consistent with the null hypothesis. To make these changes you must highlight the initial value, type the value 21, and then press Enter. Using a similar procedure, set the value of $\sigma = .5$ (the known standard deviation), set the value of $\alpha = .10$, and set the value of $n = 4$. At this point you should see two normal curves. The top one represents the population, and the lower one shows the sampling distribution for sample means when $n = 4$. Notice the shading associated with the sampling distribution. The blue region indicates $\bar{x}$-values for which we would reject $H_0\!:\mu = 21$ in favor of $H_a : \mu > 21$.

Click on the bar labeled Sample and note that the weights of the 4 packages in this sample are displayed in the upper graph, the value of the sample mean is displayed in the lower graph, and the box below the bar labeled Sample displays $\bar{x}$-value, *P*-value, and the decision whether or not to reject the null hypothesis for this sample. Note that all samples are taken from a N(21, .5) population, so whenever the value of $\bar{x}$ falls in the rejection region, we will make a Type I error.

2. To verify that the probability of making a Type I error in this case is .10, you need to look at many sample means and determine how many fall in the rejection region. Use the applet to generate 50 samples of size 4. (You will need to press the Sample bar 50 times and keep track of the test results.) How many sample means lead to rejection of the null hypothesis? ______ Based on your 50 simulations, what is your estimate of the probability of Type I error? _____ How does this estimate compare to $\alpha = .10$?

A second type of error would occur if the decision were made to not reject H_0 when in fact H_0 is false. We fail to reject the null hypothesis whenever the $\bar{x}$ value falls outside the rejection region, but we have no guarantee that the null hypothesis is true just because the sample mean falls in this region. Many populations with means different than the one specified by the null hypothesis could result in sample means in this interval. A **Type II** error occurs whenever the decision is made to not reject the null hypothesis when in fact it is false. The probability of committing a Type II error is commonly denoted by β. You can use the applet to investigate the probability of making a Type II error, that is, the value of β, when the value of the population mean μ changes.

Suppose the packing machine target setting has shifted to 21.5 ounces, so that package weights are distributed according to N(21.5, .5). You can simulate sampling from this distribution by changing the value of μ_1 to 21.5. (Be sure to press Enter when you change the value.) Note that there are now two normal distributions displayed at the top of the screen. One is centered at $\mu = 21$, the other is centered at $\mu = 21.5$, and both have $\sigma = 0.5$. The lower graph displays the sampling distribution for sample means from both populations. Press the bar labeled Sample to see the weights of the 4 packages randomly selected from the "Alternative Population" in the top graph, the sample mean in the lower graph, and test results in the box.

3. Use simulations to approximate the probability of making a Type II error when testing $H_0: \mu = 21$ versus $H_a: \mu > 21$ at the α=.10 level if $\mu = 21.5$. Use the applet to generate 50 samples of size 4; keep track of how many samples result in failing to reject H_0. Based on your 50 simulations, what is your estimate of the probability of Type II error? _____ The exact value of β is given next to the graphs. How close is your approximation to the exact value? How does the shading in the lower graph relate to the value of β?

4. The **power** of a test is the probability of rejecting the null hypothesis. Confirm that the power given for this test is $1-\beta$. How does the shading in the lower graph relate to the power of the test?

Statisticians obviously prefer tests that have low probability of Type I and Type II errors and high power. The power of a test is a function of three quantities: the value of α, the sample size *n*, and the difference between the hypothesized mean and the true mean. You can use the applet to investigate how the power of this test changes with changes in each quantity.

5. To investigate the effect of the α-value on power, change the value of α to .05, leave all other settings the same, and note changes to the graph of the sampling distribution and to the value of the power. Then change the value of α to .01 and note changes to the graph of the sampling distributions and to the value of the power. Describe how the power changes as the alpha level decreases.

6. To investigate the effect of the sample size on power, restore the value of α to .10, leave all other settings the same, and use the slider to increase sample size. Note changes to the graphs of the sampling distributions and to the value of power. Describe how the power changes as sample size increases.

7. To investigate the effect of the difference between the hypothesized mean and the true mean on power, restore the value of *n* to 4 and leave all other settings the same. Now click on the graph associated with the alternative population and drag it to the right so that the mean increases. Note the changes to the graphs of the sampling distributions and to the value of power. Describe how the power changes as the difference between the hypothesized mean and the true mean increases.