Week 7 Excel Project
Using the Sampling Distribution to Assess the Likelihood of Fraud

Due: Saturday, May 15 at 11:59pm
Submissions accepted until Sunday, May 16 at 9am

In this project you'll apply ideas that we studied in Chapters 4 and 5 to help assess the likelihood that a wholesale furniture retailer committed insurance fraud. Before going further, read pages 304—305 of the textbook to get the background information for this case. The data obtained by the CPA firm is in the Excel file FIRE.xlsx.

What you'll turn in: The Excel file you create while working on the project, and a neatly written report, prepared in Word, that includes all the statistics (means, medians, standard deviations, etc.), charts, probabilities, and other things asked for in the directions below. Charts and tables can be copied and pasted from Excel into Word.

In the context of this investigation, the 3005 invoices obtained by the CPA firm are our population of interest, and the variable of interest is the Gross Profit Factor (GPF) of each item. The GPF for each item is listed in the column labeled "Margin" (cells E2 through E3006).

1. Using Excel and the MIN, MAX, AVERAGE, and MEDIAN functions, calculate the minimum, maximum, population mean μ (Greek letter mu), and population median η (Greek letter eta). Also calculate the population standard deviation σ (Greek letter sigma).

   Note: When calculating the standard deviation for a population rather than for a sample in Excel, you must use the function =**STDEV.P( )**

2. Select Column E and then insert a Histogram chart, to visually illustrate the GPF data. You may need to enlarge the chart somewhat to see things clearly. Add informative titles for the horizontal axis, the vertical axis, and the chart overall (e.g., Gross Profit Factor, Frequency, and Gross Profit Factors of Items Sold in the Previous Year).

3. Based on the histogram you created, and the population mean and median you calculated in #1, discuss whether the GPF data appears to be symmetric about the mean, skewed to the left, or skewed to the right.

Because the retailer's records were not computerized, the retailer could not easily calculate the population mean μ, like you did in #1. Instead, they selected samples of sizes n=134 and n=119, calculated the sample mean $\bar{x}$ (x-bar) for each sample, and then averaged the two results.

First, we will consider the retailer's sample of n=134 items.

4. Assuming that we are randomly selecting n=134 items from our population, and using the information you calculated in #1, explain how to calculate the mean $\mu_{\bar{x}}$ and the standard deviation $\sigma_{\bar{x}}$ of the sampling distribution of $\bar{x}$. Then calculate those values.

5. Discuss what the Central Limit Theorem says the sampling distribution of $\bar{x}$ will be, at least approximately, when n is sufficiently large (which it is in our situation).

6. Now assume that $\bar{x}$ is normally distributed, with the mean and standard deviation you calculated in #4. What is the probability that the normal random variable $\bar{x}$ is smaller than the retailer's value of 50.6? In other words, what is $P(\bar{x} < 50.6)$? (The random variable $\bar{x}$ is not a standard normal random variable, so you'll have to do a z-score conversion before you can calculate the probability in Excel using the NORM.S.DIST function. In your report, explain how you do this conversation.)

   What then is the probability that the retailer would get a value for $\bar{x}$ that is *at least* 50.6, i.e., what is $P(\bar{x} \geq 50.6)$?

Next, we will consider the retailer's second sample of n=119 items.

7. Repeat #4, now for n=119.

8. Repeat #6, now using the mean and standard deviation you calculated in #7, and using the sample mean GPF of 51 that was given by the retailer.

Finally, we will consider the retailer's overall sample of n=253 items.

9. Repeat #4, now for n=253. (Technically, since 253 is more than 5% of 3005, the footnote at the bottom of page 284 of the textbook states that we must multiply by a finite population correction factor when calculating $\sigma_{\bar{x}}$, but for simplicity we will ignore it.)

10. Repeat #6, using the mean and standard deviation you calculated in #9, and using the sample GPF of 50.8 that was given by the retailer.

Now we will analyze the results from #6, #8, and #10.

11. The retailer told the insurance company that their loss estimates were based on selecting two independent, random samples. If the two samples are in fact selected independently, then the following two events are independent:
    o A random sample of 134 items has a sample mean greater than or equal to 50.6.
    o A random sample of 119 items has a sample mean greater than or equal to 51.
    Assuming that the retailer is telling the truth, and that these are independent events, use your answers from #6 and #8 to calculate the probability that the retailer would get a sample mean of 50.6 or larger from their first sample of n=134 items AND get a sample mean of 51 or larger from their second sample of n=119 items.

12. Based on all the information you have calculated, what is your opinion regarding whether the furniture retailer committed insurance fraud? Discuss how your answers to #3, #6, #8, #10, and #11 influenced your opinion.

In the last part of the project, you will use Excel to select your own random samples of 134 and 119 items from the population. This part of the project requires the Data Analysis ToolPak. See the instructions from the Week 3 Excel Project for information about loading the toolpak.

13. Open the Data Analysis tool (under the Data tab on the Ribbon) and choose Sampling from the Data Analysis window. Then enter
    o   Input Range: E2:E3006
    o   Labels: Leave the box UN-checked.
    o   Sampling Method: Random
    o   Number of Samples: 134
    o   Output: New Worksheet Ply (with a name like "random sample of n=134").
    You should now have a worksheet with the GPFs of 134 items that were selected randomly from among the 3005 items in the population, located in cells A1 through A134. Calculate the mean GPF for the items in your sample.

14. Repeat #13, this time for n=119. Don't forget to change the Number of Samples and the name of your output worksheet.

15. Calculate the average of the two sample means you found in #13 and #14.

16. How do your results from #13, #14, and #15 compare to the figures that were submitted by the furniture retailer? If you submitted these figures to the insurance company, do you think the insurance company would accuse YOU of committing insurance fraud? Why or why not? Back up your answer by calculating probabilities like those in #6, #8, and #11, but now using your sample GPFs instead of the retailer's.