

Due: Thursday, 27 March

You maintain a machine for your company that produces USB ports for computers. A year ago, when the machine was new and working properly, the dimensions (in inches) of 5000 ports were recorded to be used as a baseline for future comparison. To check the operation of the machine, a random sample of 10 parts is chosen every other day, and the average dimension (in inches) is recorded for each sample. Here are the observations over the past 48 days.¹

Day	2	4	6	8	10	12	14	16	18	20	22	24
mean dimension of 10 parts, \bar{x}	.5002	.4984	.5019	.4993	.5036	.4984	.5010	.4976	.5045	.5010	.5002	.4967
Day	26	28	30	32	34	36	38	40	42	44	46	48
mean dimension of 10 parts, \bar{x}	.5019	.4958	.4958	.4967	.5002	.4976	.5019	.4993	.4993	.4976	.5054	.5028

You will need the following information to make your report.

- Compute the mean and standard deviation of the original 5000 measurements. We will use this baseline mean and standard deviation in place of the (true) mean and standard deviation of the population of USB ports manufactured by a properly working machine. Create a histogram of the original data.
- *If the machine is working correctly*, what is the expected value and standard error of the mean of 10 parts. (be careful here!)
- Create a scatterplot showing the day number across the horizontal axis and sampled average port size on the vertical axis. Add horizontal lines at the mean, ± 1 SE, ± 2 SE, and ± 3 SE port sizes.²

Precision Since the variation in these values are extremely small, we may need to increase the precision of the reported values in the spreadsheet. You should keep at least 6 digits after the decimal. To adjust the displayed precision select one or more cells, right click, and select “Format Cells. . .” → “Number” → “Decimal Places”

Write-up Write a report to your boss (who is not a statistician so needs all statistical terms to be explained to her) with a recommendation for whether to shut down the machine for maintenance or not. Do not include any tables of data or spreadsheet code in your report. Be sure to include (at least) the following in your report.

- Your values for the standard error ranges and an explanation of their significance.
- Include your graphs in the relevant section within your report.
- Comment on the patterns of the original data and of the samples taken over the last 48 days.
- If the machine is working properly, the central limit theorem says that the sample means should follow the 68/95/99.7 rule. Count how many of the sample means are within the corresponding limits and compute the percentage of means within the limits. Do these percentages seem to indicate that the machine is working properly according to the central limit theorem? Explain.
- The central limit theorem only works when the individual samples are large enough. How can we tell here that our samples of size 10 are large enough.
- Do you recommend that the machine be shut down or not? Explain.

¹A file containing all data (original 5000 measurements and the 24 measurements from the last 48 days) is available on the course webpage, <http://dean.serenevy.net/>.

²An easy way to do this is to right click on the vertical axis labels, select “Format Axis...” → “Axis Options”, then set the “Minimum” to $\mu - 3SE$ and the “Major Unit” to the SE.

Extra Credit 1 (Due 3 April)

You are hired by a veterinarian to help improve patient compliance (all shots are gotten on time, heartworm, senior bloodwork, etc.). Below is a table of compliance data for the 12 months prior to your hiring (followed by the compliance data for the month after your hiring). Each number represents the number of animals who have properly complied with the corresponding recommendations of the American Animal Hospital Association.³

Month	patients	C	D	E	F	G	H
Apr 2006	621	92	22	279	209	25	123
May 2006	657	87	28	289	197	21	107
Jun 2006	796	100	46	340	262	38	151
Jul 2006	864	99	58	424	259	29	232
Aug 2006	812	135	40	326	259	37	214
Sep 2006	860	169	34	394	249	29	146
Oct 2006	829	114	27	291	254	22	174
Nov 2006	882	127	46	363	276	33	131
Dec 2006	869	125	55	302	262	38	201
Jan 2007	644	90	30	214	171	26	124
Feb 2007	695	124	39	262	225	23	206
Mar 2007	809	78	34	335	254	39	159
May 2007	<i>933</i>	<i>184</i>	<i>58</i>	<i>419</i>	<i>282</i>	<i>48</i>	<i>292</i>

You were hired in April 2007, so the May 2007 data in the last row represent the first month in which the results of your efforts should be visible. Your position is only temporary unless you can show that you are making a real difference in customer compliance. Write a report to the review committee which explains how each of the compliance variables has changed and which variables you claim have been improved due to your efforts.

Hints/Suggestions:

- Since the number of patients changes each month, each compliance variable should be converted to a percentage before the data are analyzed. For example, create a new sheet where each cell contains percentages rather than raw numbers (I.E. =C2/B2, ...).
- Use a 97.5% *confidence level* to decide which variables have varied sufficiently to be unlikely caused by random fluctuations.
- If you wish to compute exact probabilities, you may use the NORMDIST function. For example, if you have a mean of $\mu = 0.6$ and standard error $SE = 0.05$, then you can compute the probability of drawing a sample whose mean is 0.5 or less by computing, =NORMDIST(0.5, 0.6, 0.05, 1). (The last “1” is a bit magical, you will need it, it will always be “1”, but I won’t go into why you need it.) The result is 0.023 telling us that there is a 2.3% chance of drawing a sample with mean less than 0.5 (and 97.7% chance of drawing a sample with mean larger than 0.5).

Extra Credit 2 (Due 3 April)

Find data on per-child spending for K–12 education by state, county, or school district from before the No Child Left Behind Act of 2001. Then find average SAT scores by state for that same year. Create a scatterplot and compute the correlation coefficient and formula for the the least-squares regression line.

Do the same for some recent data from the past year or so. Compare these correlations and scatterplots. Use your regression equations to predict Indiana’s average SAT score using Indiana’s per-child spending on K–12 education. Conversely, predict spending from the average SAT score in Indiana.

Write a report describing your findings.

If there is some other topic you are personally interested in feel free to use variables relating to that topic (or even different variables relating to this topic). I merely provide these variables as an example.

³A file containing the data from before your hire is available on the course webpage, <http://dean.serenevy.net/>.