Phys 2426, Summer I, 2015
Running Lab Instructions

You have measured many resistances of many resistors. Our goal now is to analyze the situation statistically.

In statistics, you are interested in **individuals**, which could be people, or widgits, or whatever. They are not necessarily people, in a statistical context. For the running lab, individuals are individual resistors.

The **population** is the set of all individuals in which you are interested. For the running lab, the population is all the resistors in the box.

A **parameter** is a number that describes the population. The parameter we are most interested in is the number of resistors that are out of spec. In other words, each resistor is supposed to be within 1% of the nominal value. If a resistor is within that tolerance, it will *not* show up in the number of resistors that are out of spec. If a resistor is out of spec, then it *will* show up in the number of resistors that are out of spec. We are also interested in whether the average resistor value is equal to the nominal resistor value or not.

However, it is usually impractical to conduct a census, where you measure every single individual in the population for the parameter of interest. Thus, you select a **sample**, which is a subset of the population. The idea here is that you would measure the parameter value for your sample (called a statistic - more on that below), and then you would hope that the sample value is close to the population value. For the running lab, our sample is all the resistors you actually measured - five from each of the nominal values you chose. You should have roughly 40 to 50 individuals in your sample. As an important side note, **how you choose your sample is unbelievably important** when it comes to the inferences you can make from your sample. Truly random samples are usually the most powerful, although knowledge of the underlying structure of a population could lead you to pick a different kind of sample appropriate to that population. Much of statistics is knowing how strong your inferences can be, given the sample size and the data you actually got. Usually, larger sample sizes give you stronger inferences about the population.

A **statistic** is a number that describes a sample. For every parameter in which we might be interested, typically there is a corresponding statistic. The statistics we are interested in are the number of resistors in the sample that are out of spec (more than 1% away from the nominal value), and the average resistor value.

Now, all the resistors you measured are **supposed** to be within 1% of the nominal value. Does that hold for every resistor you measured? The average resistor value is **supposed** to be equal to the nominal value. Does that hold for your sample?

1. Open a new Excel or LibreOffice spreadsheet. Construct several columns.

   a. The first column (A column) will be the nominal value for each resistor you measured. This should be clumped in groups of five, since you measured five resistors in each size.

   b. The second column (B column) will be your actual multimeter data, corresponding to the nominal values you listed in Column A.

   c. The third column (C column) will be a "normalized" resistor value. To compute this, simply type in the formula =B2/A2, and fill down. The result should be a lot of numbers close to 1, presumably some greater and some lesser.

   d. Good statistical practice at this point would have you do a dotplot or histogram of the data. If you know how to do this easily, do it. It's not native to Libreoffice Calc, but there might be a ToolPak method for doing it in Excel. You also might be able to find some website that'll do it.

   e. We need to find out if there are any **outliers** - data points that are far away from the usual pattern. There is a standard method for checking for outliers. To use this method, you need to compute the so-called **five number summary**, which is:

i. Minimum (you can calculate this as =MIN(C2:CN), where cell CN is your last normalized data point).

ii. Q1, called the 25th percentile; this is the number such that 25% of the data lies below this number. Calculate it using =PERCENTILE(C2:CN,0.25).

iii. Median, also known as the 50th percentile. This is the number such that half or 50% of the data is below it. Calculate using either =PERCENTILE(C2:CN,0.5) or =MEDIAN(C2:CN).

iv. Q3, called the 75th percentile; this is the number such that 75% of the data lies below this number. Calculate it using =PERCENTILE(C2:CN,0.75).

v. Maximum (you can calculate this as =MAX(C2:CN)).

f. Once you have the five number summary, compute the so-called **interquartile range**, or IQR, as Q3-Q1. This is a useful measure of the spread of data, particularly if you think the data is skewed right or left.

g. Once you have the IQR, check Q1-(1.5)IQR and Q3+(1.5)IQR. Any data points that lie below the first number are outliers, and any data points that lie above the second number are outliers.

h. I don't anticipate that you will have outliers. Hopefully, the manufacturer has done its job.

i. Also, at this point, you can simply check if any of your data lies above 1.01 or below 0.99. Those would be resistors that are out of specification (or spec, for short).

j. Assuming you don't have any outliers, we are finally going to compute a **confidence interval** for the mean normalized resistor value. We think the mean normalized resistor value is simply 1. So a confidence interval asks the question where is the population's mean? Without a census, recall, we cannot compute this directly. But with a sample, we can **estimate** where we think the mean actually is. As an aside: many folks do what is called **hypothesis testing**. However, the problem with hypothesis testing is that it doesn't give as much information as a confidence interval. Some scientific journals have actually banned the use of hypothesis testing, because of the various issues with it. In any case, confidence intervals give the same information as an hypothesis test, and more besides. That is, confidence intervals are superior in every way to hypothesis tests. So, suppose we have constructed a confidence interval (a, b) at the 95% level. We say that we are 95% **confident that the interval captures the parameter in question.** This means that if we take lots of samples, and compute the 95% confidence interval each time, that we would expect 95% of those intervals to include the true parameter value. It does NOT mean that we have a 95% probability that the parameter is in the interval.

k. So how do we compute the confidence interval? I would recommend a simple online calculator to do this. The details would require more background that we have need of. Go to

```
https://www.easycalculation.com/statistics/confidence-limits-mean.php
```

and enter your confidence level (95% is a standard confidence level), sample size, standard deviation, and mean. Hit the "Calculate" button, and the interval will appear just below the button.

l. We think the mean is 1; if 1 is in the interval you calculated, then we can have "confidence" (I'm using this in the vernacular meaning, not the statistical meaning) that the manufacturer's claims are true.

2. In your report, do NOT give me all your data. That's too much data. Do report your mean, sample size, standard deviation, and five-number summary. Report if you have any outliers. If you managed a histogram, include that. Report the results of your confidence interval. Finally, report how many, if any, normalized resistor values are greater than 1.01 or less than 0.99.

3. The usual lab report guidelines are also in effect, as much as they apply. Use good judgment here. Give me a measurable hypothesis, tell me how you got your data, including how you chose your sample. Report your data, analyze the data, and draw your conclusion.