**Basic statistics in Excel**

This data is from an experiment where 16 *P. aeruginosa* samples were grown on 2 surfaces and the size of the biofilm was measured. The question is whether the bacteria grow better on one surface compared to the other.

**Load the data analysis ToolPak**

Follow the instruction here to get the add-on: <https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4#OfficeVersion=Windows>

**Replicate combining and outlier removal**

1. Download the dataset from NOW named ‘drug comparisons.xls’
2. First we need combine the 3 technical replicates to be a single observation
3. For each dataset (sheet) create a column called “Observation”
   1. i.e. place “Observation” in the column head row
4. Get the mean of the replicates
   1. For the first row (sample 1) in the observation column cell type =average(
   2. You can now select the cells that you wish to get the average (mean) of. Select the first replicate then holding shift select the last replicate.
   3. Once done, close the bracket on the formula
      1. It should look like =AVERAGE(B2:D2) in the formula box
   4. Press enter
   5. Place your cursor over the bottom right corner of the cell you just filled and you will see a black + appear
   6. Click and hold that cross and drag it down the other 15 rows so that all 16 samples have the mean observation

Next we will look for any outliers in our data. To do this we need to first see if the data is normally distributed

1. Create a column that has the header “Bins” and then in that column put the numbers 1-15 (1 number per row):

|  |
| --- |
| Bins |
| 1 |
| 2 |
| 3 |
| 4 |
| 5  Etc. |

1. Click the Data tab at the top of your workbook and then the data analysis which should be at the far right
2. Select Histogram
   1. Put the Observation column as the Input range
      1. Click the Input range box and then on your sheet click the letter at the top of the Observation column
   2. Put the Bins column in the Bin range
   3. Select “Output range” and put in $J:$K
      1. This will put the histogram data in these 2 columns (J and K)
   4. Tick the box for chart output
   5. Click OK
3. Does it seem like the top or bottom value is an outlier?
   1. If so we need to do a Grubbs test for outliers
4. The steps to do a Grubbs test in Excel are outlined here: <https://www.statology.org/grubbs-test-excel/>
5. Follow it, changing the values as needed in the final example
   1. Remember is it’s a minimum, then you need to do mean-value not value-mean for G
   2. Do we remove the bottom value?
      1. If so, do it now
6. Repeat all steps for the Surface 2 data

**Calculate descriptive statistics for each dataset**

For each surface, let’s get the descriptive statistics (mean, median etc)

1. Go to the Surface 1 sheet
2. In the Data tab, select Data analysis
3. Select Descriptive Statistics
4. For the input range select the Observations data
5. For the output range select ‘New workbook ply’ and give it the name ‘Statistical tests’
6. Click ‘Summary statistics’ and press ok
7. Go to the Statistical Tests worksheet and rename “Column 1” to ‘Surface 1’
8. Redo this for Surface 2 but instead of ‘New workbook ply’ as output range put in 'Statistical Tests'!$D:$D
   1. This will place it in column D of the new worksheet we created in step 5
9. By eye, do you think there will be a difference in the mean, median and range of biofilm sizes on these two surfaces?

**Look for statistical difference between conditions**

Now that we have our datasets, we want to compare them to look for a statistically significant difference in biofilm size

1. In the Data tab, select Data analysis
2. We have 2 sets of data to compare (2 conditions) so we will use a T-test
   1. Normally we would do a Shapiro-Wilk and Bartlett’s test to see if we can use a parametric test here but that is complicated in Excel
   2. If doing these in your research, best to use GraphPad, SPSS or R
3. Which T-test should we choose for our data?
4. Select the correct one and click ‘OK’
5. Put the Surface 1 observations column as variable 1 range and the Surface 2 observations column as the variable 2 range
   1. You can click on the sheet at the bottom of the workbook to switch between them
6. In the output range options type 'Statistical Tests'!$G:$G
7. Click OK
8. Look at the P-value
   1. Should we be using a 2-tailed or 1-tailed p-value?
   2. What is the p-value?
   3. Is there a statistically significant difference in the biofilm sizes on these two surfaces?