Statistical Thinking for Forensic Practitioners

Excel Lab on Part 4: Types of Data

We will practice making different types of visualizations using Excel. Open the `Glassdata_part4_lab.xlsx` file found on the course website. This file contains a dataset of the refractive index and chemical composition of samples of different types of glass.

1 Visualizing data from individual columns

Let us first consider visualizing data in columns individually rather than jointly.

1. What type of variable is `Refractive Index`: Categorical Nominal, Categorical Ordinal, Continuous, or Discrete?

   Continuous

2. Create a histogram of the data in the column `Refractive Index` by following the section in this guide corresponding to your version of Excel.

3. Describe the shape of the histogram you created including the modality (uniform, unimodal, bimodal, multimodal) and the symmetry (symmetric, left-skewed, right-skewed).

   Unimodal and right-skewed

4. Format the histogram you created by right-clicking on it and selecting `Format` (refer to this guide if you experience problems). In particular, change the Chart Title to say "Refractive Index." You may format the other aesthetics to your liking.

5. Under the `Series "Refractive Index"` header, change the way that the number of bins are calculated from "Auto" to "Number of bins." Set the number of bins to 25. Is there any evidence of outlier values based on this visual? Explain.

   Some evidence of an outlier due to the large gap in the histogram on the right side.

6. Now consider the elemental concentration columns `Na` through `Ca`. What type of variable are these columns: Categorical Nominal, Categorical Ordinal, Continuous, or Discrete?

   Continuous

7. We may want to visualize these concentrations side-by-side rather than as individual plots. This can be accomplished using boxplots. This guide describes how to create them in Excel 2013 and this guide describes how to create them in Excel 2016 (and newer). You may need to do some searching on the internet if you have a different version of Excel. Use the appropriate guide to create 8 side-by-side boxplots, one for each column.

8. Change the chart title to say "Elemental Concentrations (ppm)" (note: stands for parts per million) similar to how you changed the histogram title. Again, you can format the aesthetics to your liking. Under the `Horizontal Axis` header, shift the `Gap Width` slider to 0. This will make it easier to see the spread of each boxplot. Excel does not add a legend to the plot by default, so we must add it ourselves. Click on the plot and a `Chart` header should appear at the end of the ribbon at the top of the spreadsheet (this may say `Design`

under `Chart Tools` depending on your Excel version). Click on `Chart` and you should see a `Legend` drop-down menu (you might have to look under `Add Chart Element` drop-down menu depending on your Excel version). Click the drop-down and select `Show Legend at Right`. A legend should appear on the plot associating each color to an element.

9. Compare the boxplots across each type of element and summarize your findings in a few sentences.

   It's clear that Silicon has the highest elemental concentration, although the distribution appears to be approximately symmetric. The distribution of Calcium seems to have a few more outlier points (which are the visible points on the boxplot) indicating a possible right-skewed distribution. The distribution of Barium and Iron are difficult to describe based on these plots since the scales of the elemental concentrations are extremely different compared to the other elements (this might be fixed by performing a log-transformation of the concentration values.)

10. Now consider the `Type` column. What type of variable is `Type`: Categorical Nominal, Categorical Ordinal, Continuous, or Discrete?

    Categorical Nominal

11. To create visualizations for data like those in the `Type` column, Excel requires us to create a count table summarizing the frequency of each category. In columns `L` and M you will see a table template has been created. To populate the rows of the table, use the same procedure as you did on question 6 of the Excel Lab on Part 3: Data Collection (i.e., using the `COUNTIF` function).

12. After filling in the `Count` column of the table, create a column chart from the data. Change the chart title to "Glass Type" and format the aesthetics of the chart to your liking. Add a label to the vertical axis by selecting the chart, clicking `Chart` in the ribbon at the top of the spreadsheet, clicking the arrow under `Axis Titles`, hovering over `Primary Vertical Axis Title`, and selecting `Rotated Title` (you might have to look under `Add Chart Element` drop-down menu depending on your Excel version). Make "Count" the vertical axis label.

13. A glass analyst hopes to use this data set to compare the average refractive index of Building float glass and Tableware glass. They assume that the precision with which they can estimate these averages will be the same for these two types of glass. Based on this column chart, what would you tell this glass analyst?

    Because Building float glass is much more prevalent in the data set than Tableware glass, it would be safe to assume that the average estimate obtained for the refractive index of Building float glass would be much more precise than that of Tableware glass. It would be wise to caution the glass analyst against making this assumption in their analysis.

2 Joint visualizations of quantitative data

Scatterplots are a very useful tool for visualizing the relationship between two quantitative variables. They can uncover, for example, whether the refractive index of a glass sample is heavily associated with the concentration of a particular element. Statisticians sometimes refer to 2D scatterplots as a visualization of the "joint distribution" of two variables.

14. Create a scatterplot between the `Ca` column (on the horizontal axis) and the `Refractive Index` column (on the vertical axis) using this guide. Remove the chart title as it will be redundant once we add axes labels. By default, the axes are not labeled and the legend provided is redundant. Remove the legend following a procedure similar to question 8 above. Label the vertical axis "Refractive Index" and the horizontal axis "Calcium Concentration (ppm)" following a similar procedure to question 11 above.

15. Describe the Form, Strength, Direction, and Outliers of the scatterplot. (Note: refer to slides 29-33 of the Part 4 lecture slides)

    Linear, Strong, Positive, with no apparent outliers

16. A glass analyst is interested in determining whether the Calcium concentration of a glass fragment could be used to estimate its refractive index before measuring. Based on this dot plot, what would you say to this glass analyst? For example, if a glass fragment had a Calcium concentration of 10 ppm, what might be a good estimate for its refractive index (based on this plot)?

    Because there is such a strong relationship between the Calcium concentration and refractive index in this data set, it would appear that Calcium concentration is a good predictor of refractive index. For a fragment with a Calcium concentration of 10 ppm, we might estimate its refractive index to be around 1.52 based on this plot.