Statistical Thinking for Forensic Practitioners

Excel Lab on Part 5: Probability Models and Uncertainty

1 Probability models in Excel

We will practice visualizing the probability density or mass functions of various probability models. Open the `Glassdata_part5_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. Additionally, the `Scratches` column contains the number of scratches counted by a forensic practitioner on the pane from which the glass sample was obtained.

1. The variables `Na` through `Scratches` (columns `A` through `J`) are all quantitative. However, for probability modeling purposes, there is an important distinction between the variables `Na` through `Refractive Index` and the variable `Scratches`. Identify this distinction.

   The variables `Na` through `Refractive Index` are continuous variables while the variable `Scratches` is discrete. Whether a variable is continuous or discrete informs whether we use continuous or discrete probability distribution models to describe the variable.

2. A glass evidence researcher would like to use a probability model to describe the variable `Scratches`. What types of values can the variable `Scratches` take on?

   Non-negative whole numbers (0 scratches, 1 scratch, 2 scratches, etc.)

3. The researcher is unsure whether they should use a binomial or Poisson distribution to model the variable `Scratches`. These two discrete probability models capture different data structures/behavior. Which model seems more appropriate? Explain.

   A Poisson distribution seems more appropriate. Recall that a Poisson distribution describes the number of events occurring in an interval of space (or time, but that isn't applicable here). The natural analogue to this situation would be to define an "event" as the occurrence of a scratch on an "interval of space" that is a pane of glass. Also recall that a binomial distribution describes the number of successes out of n binary, independent, random trials (e.g., number of heads out of n coin flips). In counting the number of scratches per pane of glass, it is not clear what a single "trial" would be analogous to in this situation or, even if we could define such a trial, whether they would be independent.

A simple yet often effective tool to determine how well a probability model "fits" a variable is a visualization. We assume that the variable of interest behaves in a manner dictated by the selected probability model. In our example, we assume the values that our variable `Scratches` take on should "look like" a typical, random sample from a Poisson distribution. For discrete variables, we can validate this assumption by creating an empirical frequency distribution (i.e., a bar chart of the data) and visually compare it to the theoretical frequency distribution dictated by the probability model (i.e., the probability mass function scaled by the sample size). This is a very simple (non-rigorous) example of what is called a goodness-of-fit test. Before we can visualize the theoretical frequency distribution, we need to obtain an estimate of the distribution's unknown parameters.

4. For the Poisson distribution, it turns out that an all-around good estimator of the parameter λ is the sample mean. That is, for a set of observations $\{x_1, ..., x_n\}$ (where $n = 215$ in this problem), the estimator $\hat{\lambda}$ (read "lambda hat") is

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

   Calculate $\hat{\lambda}$ for the observed values of the `Scratches` variable (recall the AVERAGE function in Excel).

   $\hat{\lambda} = 5.888$ **rounded to 3 decimal places**

5. With this estimate for λ, we can obtain an estimate of the theoretical frequency distribution under the Poisson model assumption. In columns `N` through `Q` you will see the beginning of a second table that we will fill-out to visualize the theoretical and observed frequency distributions of the `Scratches` variable. The `Scratches Values` column contains the range of observed values in the `Scratches` column, 0-13. Using the POISSON.DIST function, calculate the (estimated) theoretical probabilities entering the $\hat{\lambda}$ value calculated above as the `Mean` and setting `Cumulative` to `FALSE`.

6. Compute the estimated theoretical frequencies by multiplying the values calculated in the `Estim. Theoretical Prob.` column by 215.

7. Use the COUNTIF function in Excel to compute the observed frequency of each value.

8. Create a column chart from the `Estimated Theoretical Freq.` and `Observed Freq.` columns by highlighting both and selecting `Clustered Column` from the `2-D Column` sub-menu under the `Insert` tab. Comment on the similarities and differences between these two distributions. (Note: the horizontal axis label for the plot you created will range from 1 to 13. However the actual data ranges from 0 to 12. Keep this in mind when drawing conclusions.)

   The two distributions look similar in that they both reach a maximum at 5 scratches. The observed frequency distribution certainly looks more "ragged" than the theoretical frequency distribution, although this is to be expected. Oddly enough, there are a relatively high number of panes with 12 scratches in the observed data.

9. If we were to continue randomly sampling more panes and counting the number of scratches on each, would you expect the observed frequency distribution to look more or less similar than the theoretical frequency distribution?

   The observed frequency distribution should look more similar to the theoretical frequency distribution as the sample size increases.

We now turn our attention to continuous random variables. Many well-used discrete probability distributions describe random, physical phenomena (e.g., binomial distributions and number of successes in n binary trials). Continuous distributions, on the other hand, do not often have such natural physical analogues. Instead, a statistician might visualize a continuous variable first and choose a probability distribution with an agreeable shape.

1. Create a histogram of the data in the `Al` (aluminium) column. Comment on the skewness and modality of the histogram.

   Right-skewed and unimodal

2. The two continuous probability models discussed in the Part 5 lecture slides were the Normal and Log-Normal distributions. Comment on why the Log-Normal might be a more appropriate probability model for these data.

   The log-normal seems more appropriate because its probability density function is right-skewed. Also, elemental concentrations can not be less than 0 (parts per million, that is), so the normal distribution, which places non-zero probability on negative values, doesn't seem as appropriate.

We will visualize the estimated probability density function (PDF) of the Log-Normal distribution (similar to how we visualized the theoretical frequency distribution of the Poisson) as an extremely rough "goodness-of-fit" test for these data. Unfortunately, Excel does not allow for the creation of line + histogram combo plots, so we will have to create a separate visualization of the Log-Normal PDF. Similar to the Poisson problem, we will need to estimate the parameters of the Log-Normal distribution.

3. The Log-Normal distribution is parameterized by μ and σ^2 (or, equivalently, μ and σ, which we will use here). All-around good estimators for μ and σ are

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \ln(x_i)$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (\ln(x_i) - \hat{\mu})^2}$$

   where $\ln(x_i)$ is the natural logarithm of observation x_i. You will recognize $\hat{\mu}$ as the sample mean of the natural logarithm of the observations. The form of $\hat{\sigma}$ is commonly referred to as the *sample* standard deviation, in this case of the natural logarithm of the observations. Note that $\hat{\mu}$ is used in the definition of $\hat{\sigma}$.

   You will find the natural logarithms of the `Al` elemental concentrations in the `ln(Al)` column (column `T`). Using this column of values, the `AVERAGE` function, and the `STDEV.S` function, calculate $\hat{\mu}$ and $\hat{\sigma}$.

   $\hat{\mu} = .306$ **and** $\hat{\sigma} = .368$ **rounding to 3 decimal places.**

4. The `Al Range` column contains a sequence of evenly-spaced numbers from 0 to the maximum value observed in the `Al` column (3.48 approximately). These are the values at which we evaluate the PDF of the Log-Normal distribution. Using $\hat{\mu}$ and $\hat{\sigma}$ calculated in the previous problem and the `LOGNORM.DIST` function, determine the value of the Log-Normal PDF with mean $\hat{\mu}$ and standard deviation $\hat{\sigma}$ at each value in the `Al Range` column.

5. Create a line chart using the values you computed in the `Log-Normal PDF` column. By default, the horizontal axis of the plot corresponds to the row number of each value. We instead want the horizontal axis to correspond to the values in `Al Range` column. To change this, right-click on the plot and click `Select Data`. Click `Edit` under the `Horizontal (Category) Axis Labels` window on the right. Click on the column `U` header to highlight its contents. Select `OK`. Uncheck the `Al Range` box so that this isn't included on the horizontal/vertical axes. Select `OK`. The horizontal axis should now range from approximately 0 to approximately 3.4 (if any of these instructions are confusing, refer to this guide).

6. Comment on the similarities and differences between the histogram of observed values and the Log-Normal PDF line chart.

   Both plots show right-skewed behavior, which was the main reason for selecting the Log-Normal probability model in the first place. It also appears that both plots reach a maximum around 1.2 on the horizontal axis. The Log-Normal PDF appears to decay faster as it approaches 0 on the horizontal axis than the histogram of observations.