Formal Description of Statistical Modelling

The standard rejection test of whether who statistical samples are drawn from the same unobserved distribution is the Kolmogorov-Smirnov test. It is not applicable here, however, since name occurrences in our sample do not form an ordered set. The cumulative distribution function (CDF) is therefore undefined. One can scramble the occurrences without changing the data and the CDF will not be preserved. This prohibits us from using other rejection tests based on the CDF, most importantly the Cramer von Mises test.

Another test to look at differences in frequencies is the Kullback-Liebler Divergence. Unfortunately, given the size of the data set, this test is inappropriate. Further, subtracting out the Gospels-Acts sample from the contemporary population sample (i.e., the sample of all other attested male name occurrences in Palestine in 4 BCE-73 CE) gives infinity for many of these tests, because there are some names only in Gospels-Acts, so then the expected population is zero. The low-N problem also applies to the likelihood ratio test, the chi-squared test and the multinomial test. In the latter case, uncertainties in the underlying name probabilities are not considered and this can lead to misleading answers.

Instead of relying on these tests, we model the Gospels-Acts sample of name occurrences as being produced by a stochastic process which uses the probabilities for drawing the various names estimated from the contemporary population sample. We first estimate the probability of drawing each name, denoted by θ_i where i is an index across names. These parameters are determined by the properties of the Multinomial-Dirichlet model for many categories, which is a generalization of the Binomial-Beta model used for two categories.

Prior to observing any data, we assume that all values of θ are equally likely, which is commonly handled by using a Dirichlet distribution,

$$\begin{aligned} p(\{\theta_1, \theta_2, \cdots, \theta_K\}|\{\alpha_1, \alpha_2, \cdots, \alpha_K\}) &= \frac{1}{\mathbf{B}(\alpha_1, \alpha_2, \ldots, \alpha_K)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_K^{\alpha_K - 1} \\ &\equiv \mathrm{Dir}(\{\theta_1, \theta_2, \cdots, \theta_K\}|\{\alpha_1, \alpha_2, \cdots, \alpha_K\}) \end{aligned},$$

where we use $\alpha_1 = \alpha_2 = \cdots = \alpha_K = 1$ for the uniform prior.

The likelihood of seeing a certain number of draws of each possibility, $n_1, n_2, \cdots, n_K$ for a total of N draws of names, given the probabilities for drawing each of K possible names, $\theta_1, \theta_2, \cdots, \theta_K$, is expressed by the Multinomial distribution,

$$p(\{n_1, n_2, \cdots, n_K\}|N, \{\theta_1, \theta_2, \cdots, \theta_K\}) = \binom{N!}{n_1!\, n_2! \cdots n_K!} \theta_1^{n_1} \theta_2^{n_2} \cdots \theta_K^{n_K}.$$

Using the multinomial distribution as our likelihood and the Dirichlet distribution as our prior, we apply Bayes theorem to get the distribution over $\theta_1, \theta_2, \cdots, \theta_K$. This will allow us to get the best estimates for the probabilities of generating the name Simon, Joseph, etc., as well as the uncertainty in those estimates based off the contemporary population sample. Conveniently, we get an updated Dirichlet distribution for our posterior,

$$p(\{\theta_1, \theta_2, \cdots, \theta_K\}|\{n_1, n_2, \cdots, n_K\}, N) = \text{Dir}(\{\theta_1, \theta_2, \cdots, \theta_K\}|\{x_1 + \alpha_1, x_2 + \alpha_2, \cdots, x_K + \alpha_K\}),$$

or, in the case all the $\alpha = 1$, we have

$$p(\{\theta_1, \theta_2, \cdots, \theta_K\}|\{n_1, n_2, \cdots, n_K\}, N) = \text{Dir}(\{\theta_1, \theta_2, \cdots, \theta_K\}|\{x_1 + 1, x_2 + 1, \cdots, x_K + 1\}).$$

The marginal distribution for one name given our full data, $p(\theta_i|\{n_1, n_2, \cdots, n_K\}, N)$ follows a Beta distribution, with properties

$$p(\theta_i|\{n_1, n_2, \cdots, n_K\}, N) = \text{Beta}(\alpha_i + n_i, N + \textstyle\sum_i \alpha_i - \alpha_i - n_i).$$

Although the Dirichlet is a multivariate distribution, this marginal distribution allows us to plot the single-name probability estimates and their uncertainties. Once we have the posterior probabilities, $p(\{\theta_1, \theta_2, \cdots, \theta_K\}|\{n_1, n_2, \cdots, n_K\}, N)$, for the name fractions in the contemporary population sample, we can simulate the name-counts for the Gospels-Acts sample with a specific total number of names by drawing randomly from the contemporary population data. In this way we can compare how similar the Gospels-Acts sample is to the contemporary population sample by comparing the simulated name counts with the counts we observe. The process consists of several steps: First, we draw a random set of name probabilities $\{\theta_i\}$ from the Dirichlet posterior estimated from the contemporary population sample. Second, from those $\{\theta_i\}$, we draw a sample of names of size N_j, where j is the index for the Gospels-Acts sample. We then repeat this sampling many times (specifically, 500,000 iterations were conducted). And finally, we compare the actual source data counts with the simulated counts, and flag anything outside of the 95% percentile range as being not consistent.

Using this method, it is easy to see both the document-to-document variation and the name-to-name variation. The variation for both is high given the low numbers of occurrences of any given name.