# Appendix: Mathematical Methods and Code

P. A. Robinson

*November 23, 2020*

## Preamble

This section briefly outlines the mathematical methods that enable us to (i) determine the probability of particular appointment outcome, assuming an unbiased process; and (ii) estimate bias in the process, if any. The details in this section are intended for readers with at least some undergraduate statistics, but are not needed to follow the discussion in the main document. A computer code that computes the measures discussed here is intended to be made available for public use. Users need only enter three numbers to use it.

## Probability of Observed Outcome

Suppose that a fraction $f$ of people in a pool of qualified applicants for a position are from a group G, where G is a characteristic that is irrelevant for the purposes of appointment. If a number $n$ of appointments is made, an unbiased process will be blind to characteristic G. Then the probability $P(r)$ of making $r$ appointments from among the $n$ is given by the Binomial Theorem:

$$P(r) = {}^nC_r f^r (1-f)^{n-r}, \tag{1}$$

$${}^nC_r = \frac{n!}{r!(n-r)!}, \tag{2}$$

where ${}^nC_r$ is the combinatorial coefficient and ! denotes the factorial.

The average expected number of appointments from group G is

$$\overline{r} = nf. \tag{3}$$

## Probability of Observed Number of Appointments or Fewer

If an actual number of appointments $r_a$ is made from group G, the probability of exactly $r_a$ appointments is not the most useful quantity because, if $n$ is large, nearby values can be almost equal in likelihood. Rather, we need to evaluate the cumulative probability of $r_a$ or fewer appointments from group G on the assumption of an unbiased process. This is denoted by $\mathrm{cdf}(r_a)$, where $\mathrm{cdf}(r)$ is given by

$$\mathrm{cdf}(r) = \sum_{j=0}^{r} P(j). \tag{4}$$

**95% Confidence Interval**

A common measure of the likely range over which outcomes can vary without being considered abnormal is the **95% confidence interval**, or 95% CI, of $r$. We expect 95% of outcomes to fall in this range for an unbiased process. This can be approximated for large $n$, $\overline{r}$, and $n - \overline{r}$ by the range $\overline{r} - 2\sigma(r)$ to $\overline{r} + 2\sigma(r)$, with

$$\sigma(r) \approx \sqrt{nf(1-f)}. \tag{5}$$

More generally, the 95% CI can be computed by defining it to be the interval between the 2.5% and 97.5% levels of $\mathrm{cdf}(r)$.

**Preference Ratio**

We expect an average of $\overline{r} = nf$ appointments from group G and $(1-f)n$ from amongst other candidates. Hence, if the actual numbers are $r_a$ and $n - r_a$, the preference ratio $B$ of the Other group relative to group G is defined

$$B = \frac{(n - r_a)/[(1-f)n]}{r_a/(fn)}, \tag{6}$$

$$= \frac{f(n - r_a)}{r_a(1-f)}. \tag{7}$$

This ratio tells us how much more likely selectors are to appoint a qualified person from the Other group than from group G when compared one-on-one.

**Selection Bias**

The fraction of the total appointments that were from group G was $f_a = r_a/n$ and this is the best available estimate of the fraction the selectors actually think are appointable from G. If $f_a > f$ group G was favored *prima facie*, whereas $f_a < f$ implies it was disfavored, but there will be fluctuations even in an unbiased process, so we need to do more detailed analysis.

We can determine the likely variation in the number of people $R$ from group G that the selectors would appoint if they had further sets of $n$ appointments to make, given that they actually appointed a number $r_a$. Bayesian statistics implies that the best estimate of the selectors' probability distribution $P_S(R)$ of outcomes is

$$P_S(R) = {}^nC_R f_a^r (1 - f_a)^{n-R}, \tag{8}$$

where $R$ is the number of appointments in a rerun of the process with an independent pool of fresh candidates. The part of this distribution that lies within the 95% CI of an unbiased process is an approximate measure of the probability $U$ that the selection process is unbiased; i.e.,

$$U \approx \sum_{R=\overline{r}-2\sigma(r)}^{\overline{r}+2\sigma(r)} P_S(R). \tag{9}$$

Note that if $f_a = f$ the distribution (8) is identical to the expected distribution (1) and $U = 95\%$. If $U \ll 1$ the process is likely to be biased.

**Computer Program**

A computer code that implements the above methods is intended to be made available for public use. All it requires to estimate the distributions, preference ratio, and selection bias are the total number of appointments $n$, the actual number $r_a$ from group G, and the fraction of qualified applicants $f$ who are expected to be from group G.