Problem 1. Testing for completely random dropouts

Let P_{ij} denote the probability that the $i-$th unit drops out at time t_j, $j = 1, \ldots, m$.
Under the assumption of completely random dropouts, the probability P_{ij} may depend on time, treatment, or other explanatory variables, but cannot depend on the observed measurements $y_i = (y_{i1}, \ldots, y_{i\,m_i})$.

Testing Method:

(a) Choose the score function $h_k(y_1, \ldots, y_k)$ so that extreme values constitute evidence against completely random dropouts. A sensible choice is

$$h_k(y_1, \ldots, y_k) = \sum_{j=1}^{k} \omega_j y_j.$$

(b) For each of $k = 1, \ldots, (m-1)$, define

$$R_k = \{i : m_i \geq k\},$$
$$r_k = \{i : m_i = k\},$$

and compute the set of scores $h_{ik} = h_k(y_{i1\ldots,y_{ik}})$ for $i \in R_k$.

(c) If $1 \leq |r_k| \leq |R_k|$, test the hypothesis that the r_k's scores so defined are a random sample from the "populations" of R_k's scores.

—

Remark:

1. The implicit assumption that the separated $p-$values are mutually independent is valid precisely because once a unit drops out, it never returns.
2. A natural test statistics is $\overline{h}_k = \frac{1}{|r_k|}\sum_{\{j \in r_k\}} h_{jk}$. Under the assumption of completely random dropouts,

$$\overline{h}_k \sim N\left(\overline{H}_k, \frac{|R_k| - |r_k|}{|r_k|(|R_k| - 1)} \sum_{\{j \in R_k\}} (h_{jk} - \overline{H}_k)^2\right),$$

where

$$\overline{H}_k = \frac{1}{|R_k|} \sum_{\{j \in r_k\}} h_{jk}.$$

   - When $|R_k|$ or $|r_k|$ is small, evaluate the randomization distribution of $\overline{h}_k$ under the null hypothesis.
   - Alternative method ...

3. The Final stage consists of analyzing the resulting set of $p-$values via
   (a) Empirical distribution of the $p-$values
   (b) Kolmogorov-Smirnov statistic $D_+ = \sup |\hat{F}_n(p) - p|$

Suppose that σ^2 is the population variance. This implies: If the random variable X is the result of a single draw from the population, then $\mathrm{Var}(X) = \sigma^2$. Now consider drawing a sample of n items $X_1, \ldots, X_n$ without replacement from the population. Since every pair (X_i, X_j) for $i \neq j$ has the same joint distribution, the variance of the sum $S_n := X_1 + \ldots + X_n$ is

$$\mathrm{Var}(S_n) = n\mathrm{Var}(X_1) + \left(n^2 - n\right)\mathrm{Cov}(X_1, X_2) = n\sigma^2 + n(n-1)c \quad (1),$$

where we write c for the covariance between the results of two distinct draws. Formula (1) applies in the case $n = N$ as well, with the extra bonus that S_N is a constant (equal to the sum of all N values in the population). It follows that

$$0 = \mathrm{Var}(S_N) = N\sigma^2 + N(N-1)c \quad (2).$$

Solve equation (2) for $c = -\frac{\sigma^2}{N-1}$ (3) and plug back into (1) to obtain

$$\mathrm{Var}(S_n) = n\sigma^2\left(1 - \frac{n-1}{N-1}\right) = \frac{N-n}{N-1} \cdot n\sigma^2 \quad (4)$$

and

$$\mathrm{Var}(\bar{X}_n) = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \quad (5).$$

Notice the difference between formulas (4) and (5) and the corresponding formulas for sampling with replacement is a factor $\frac{N-n}{N-1}$, which is the famous correction factor for sampling without replacement.