Estimation of the number of stem cells repopulating the marrow

John O'Quigley

Lancaster University, U.K.
j.oquigley@lancaster.ac.uk

Dept. Medical Informatics, Ljubljana.
May 25th, 2006

Basic set-up

- We call p_i the proportion of marked chromosome among stem cell progeny for the i^{th} donor
- $\tilde{p}_i$ is same quantity for the patient.
- p_i is regarded as being fixed and known,
- $\tilde{p}_i$ varies about p_i on the basis of a binomial law with parameters (n, p_i)
- var $(\tilde{p}_i) = p_i(1 - p_i)/n$ where n is the total number of cells involved in repopulating the marrow.
- n will likely vary from patient to patient

Table 1.

i	p_{i1}	p_{i2}	y_{i1}	y_{i2}
1	0.76	0.79	1.058	1.095
2	0.19	0.17	0.451	0.425
3	0.57	0.51	0.856	0.795
4	0.67	0.74	0.958	1.036
5	0.54	0.51	0.825	0.795
6	0.59	0.62	0.876	0.907
7	0.39	0.41	0.674	0.695
8	0.27	0.28	0.546	0.557
9	0.66	0.67	0.948	0.959
10	0.63	0.60	0.917	0.886
11	0.47	0.49	0.755	0.775
12	0.32	0.26	0.601	0.535
13	0.58	0.61	0.866	0.896
14	0.49	0.47	0.775	0.755
15	0.97	0.97	1.394	1.405
16	0.17	0.13	0.425	0.369
17	0.57	0.46	0.855	0.745

Simple inference

- $\sigma^2 = E(\tilde{p}_i - p_i)^2$ is taken as being an underlying unknown population value
- σ^2 estimated by

$$s^2 = \{k\sum_{i=1}^{k}(\tilde{p}_i - p_i)^2 - (\sum_{i=1}^{k}(\tilde{p}_i - p_i))^2\}/k(k-1)$$

- k moment estimates of n are obtained as

$$\hat{n}_i = p_i(1-p_i)/s^2$$

- $\hat{n} = \sum_i \hat{n}_i/k$ estimates n.
- expression for $\hat{n}_i$ implies that not only does n vary between patients but that it varies in a way which depends directly on p_i
- Eg, Donor for whom $p_i = 0.5$, has about five times as many stem cells involved in repopulation than recipient j for whom $p_j = 0.05$

Weighting by variance

We could estimate n by $\hat{n}_w$ where $\hat{n}_w = \sum_i w_i \hat{n}_i, \sum_i w_i = 1$ and

$$w_i = \{p_i^2(1-p_i)^2\}^{-1} / \sum_j \{p_j^2(1-p_j)^2\}^{-1}$$

- Precision of estimator improved.
- Increased precision associated with greater instability.
- $\hat{n}_w = 36$ (after rounding). For patient 15 not only does the value of p_{i1} being close to one result in a small value for $\hat{n}_i$ ie. 15, but it attaches an exaggerated weight to this value based on w_i
- 75% of the total weighting ends up being attributed to a single observation.
- Removing this value from the analysis we obtain $\hat{n}_w = 95$, a very substantial difference. Approximate confidence intervals for these two estimates are far from overlapping and the data point 15 is highly influential.

Maximum likelihood

Normal approximation for $\hat{p}_i$ enables the log-likelihood to be written as

$$\log L(n) = \text{constant} + \frac{k}{2}\log n - \frac{1}{2}\sum_{i=1}^{k} n(\hat{p}_i - p_i)^2/(p_i q_i)$$

- Solving $\partial \log L(n)/\partial n = 0$ leads to,

$$\hat{n} = k/\left(\sum_{i=1}^{k}(\hat{p}_i - p_i)^2/(p_i q_i)\right).$$

- Also $\{\partial^2 \log L(n)/\partial n^2\}_{n=\hat{n}} = -k/(2\hat{n}^2)$ so that $\text{var}(\hat{n}) \approx 2n^2/k$.

Bias of m.l.e.

- First two terms of Taylor expansion for $\hat{n}$ lead to

$$E(\hat{n}) \approx \frac{nk}{E(\chi_k^2)} + \frac{1}{2}\mathrm{var}\,(\chi_k^2) \times \frac{2nk}{\{E(\chi_k^2)\}^3},$$

  where χ_k^2 is a chi-square variate on k degrees of freedom. Since $E(\chi_k^2) = k$ and $\mathrm{var}(\chi_k^2) = 2k$.

- First order bias of the mle is $2n/k$.

More accurate inference

Assume that $y = \sum_{i=1}^{k} n(\hat{p}_i - p_i)^2/(p_i q_i) \sim \chi^2_k$. Letting $u = \hat{n}/n = k/\chi^2_k$ and noting that $|dy/du| = y^2/k$ then, after regrouping terms, we find that the density of u is given by

$$f(u) = k^{k/2} u^{-(k+2)/2} \exp\{-k/(2u)\}/D(k/2)$$

where $D(x) = 2^x \Gamma(x)$ and $\Gamma(\cdot)$ is the gamma function. Figure 1 shows the shape of this density for $k = 5$. It is clear that for such small values of k, by no means untypical in studies of the type described in the introduction, a normal approximation will not be very accurate.

Moment estimators

Replace the s_i by the pooled estimator;

$$s^2 = \{k\sum_{i=1}^{k}(\hat{p}_i - p_i)^2 - (\sum_{i=1}^{k}(\hat{p}_i - p_i))^2\}/k^2$$

and, ignoring the correction term for the mean, which has zero expectation since $\hat{p}_i$ unbiasedly estimates p_i, we obtain the estimator

$$\bar{n} = \left(\sum_{i=1}^{k} p_i q_i\right) / \left(\sum_{i=1}^{k}(\hat{p}_i - p_i)^2\right).$$

The above equation for $\bar{n}$ should be contrasted with the form of the maximum likelihood estimator $\hat{n}$.

Moment and mle estimators

$$\hat{n} = \frac{k}{\sum_{i=1}^{k}(\hat{p}_i - p_i)^2/(p_i q_i).}$$

$$\bar{n} = \frac{\sum_{i=1}^{k} p_i q_i}{\sum_{i=1}^{k}(\hat{p}_i - p_i)^2.}$$

Transformation

- Variance stabilizing transformations; $y_i = \sin^{-1}\sqrt{p_i}$ and $\hat{y}_i = \sin^{-1}\sqrt{\hat{p}_i}$.
- For each i define $\sigma_Y^2 = E(\hat{y}_i - y_i)^2$.
- σ_Y^2 does not depend on i to a high level of approximation.
- Thus σ_Y^2 does not depend on the particular value of y_i (and in consequence p_i).
- $$s_Y^2 = \{k\sum_{i=1}^{k}(\hat{y}_i - y_i)^2 - (\sum_{i=1}^{k}(\hat{y}_i - y_i))^2\}/k^2$$
- Finally, note that $\sigma_Y^2 \approx 1/4n$ (Johnson and Kotz 1969, page 65).
- A natural estimator for n is then

$$\bar{n}_Y = 1/(4s_Y^2).$$

Inference for $\bar{n}_Y$

Let $w = \bar{n}_Y/n$ and since ks_Y^2/σ_Y^2 is well approximated by a chi-square variate on k degrees of freedom then,

$$f(w) = k^{k/2} w^{-(k+2)/2} \exp\{-k/(2w)\}/D(k/2)$$

where $D(x) = 2^x\Gamma(x)$ and $\Gamma(\cdot)$ is the gamma function. Taylor series approximations give $E(\bar{n}_Y) \approx n(k+1)/(k-1)$ and that $\mathrm{var}(\bar{n}_Y) \approx 2n^2/(k-1)$. A simple bias correction factor, then, is given by $(k-1)/(k+1)$.

Confidence intervals

- Approximate $100(1-\alpha)\%$ C.I. for n is obtained by adding and subtracting, $z_{1-\alpha/2}$ times square root of variance.
- Alternative approximation given by (L, U) where

$$G(U(k-1)/\bar{n}_Y) - G(L(k-1)/\bar{n}_Y) = 1-\alpha$$

  and $G(u)$ is cumulative distribution function for a chi-squared variate on $k-1$ degrees of freedom. As k increases, the shape of a chi-squared variate approaches that of a normal and the two intervals converge.
- Intermediate solution obtains via a Cornish-Fisher expansion for the quantiles. Taking the first three terms of a normal based expansion, i.e. inverse function corresponding to a Gram-Charlier Type A series, amounts to making a skewness correction to the symmetric interval. Denoting $\bar{n}_Y(k+1)/(k-1)$ by m and $2\bar{n}_Y^2/(k-1)$ by s_m^2, this corrected interval can be written as (L_c, U_c) where

$$L_c = m - a_0 s_m \ ; \quad U_c = m + a_1 s_m$$

  and

$$a_i = z_{1-\alpha/2} - 0.471(-1)^i(z_{1-\alpha/2}^2 - 1)/\sqrt{k-1}.$$

Example

Nash et al (1988) studied 17 donor-patient pairs.

- $s_Y^2 = \{17 \sum (y_{i1} - y_{i2})^2 - (\sum (y_{i1} - y_{i2}))^2\}/(16 \times 17)$.
- We find that $s_Y^2 = 0.0021$ and $\hat{n}_Y = 119$.
- Unbiased estimate of n obtains by multiplying $\hat{n}_Y$ by 0.89 and equals, after rounding, 106.
- A 95% normal based confidence interval is (33,179).
- Second approximation denoted (L, U) yields (46,191)
- First three terms of a Cornish-Fisher expansion gives (61,192).
- All intervals are quite wide.
- Although the suggestion is that around 100 cells are involved in repopulation, the data are quite compatible with a figure as low as say 30 or possibly as high as 200.

Hypothesis tests

Suppose we wish to test the null hypothesis that few stem cells are involved in repopulation, i.e. that n is very small.

- Specifically suppose that $n = 5$, then $E(\hat{n}_Y) = 5.625$ and $\text{var}(\hat{n}_Y) = 3.125$ so that a simple hypothesis test of $n = 5$ versus $n > 5$ leads to a rejection of $n = 5$ with a t-statistic on 16 degrees of freedom equal to $(119 - 5.625)/1.77 = 64.05$ ($p < 0.0001$).
- For any value smaller than 5 the p-value would be even smaller.
- Evidence is then overwhelmingly against monoclonal or oligoclonal reconstitution of marrow grafts after allogeneic marrow transplantation.

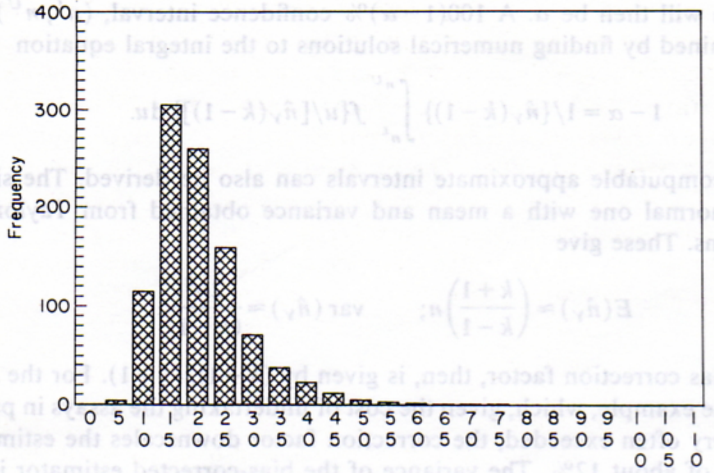

FIG. 1. The distribution of $\hat{n}_Y$ on the basis of 1000 simulations for $n = 20$.

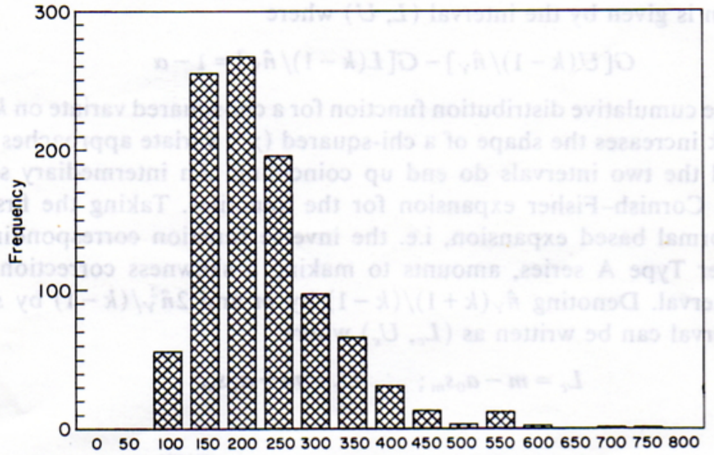

FIG. 2. The distribution of $\hat{n}_Y$ on the basis of 1000 simulations for $n = 200$.

If we wish to test simple hypotheses such as H_0: $n = n_0$ vs. H_1: $n > n_0$ then a critical region of size α is given by (n_0^c, ∞) where

$$\alpha = \frac{1}{n_0(k-1)}\int_{n_0^c}^{\infty} f\left(\frac{u}{n_0(k-1)}\right) \mathrm{d}u.$$

Likely to be of more interest is a test of the hypothesis H_0: $n = n^* < n_0$ vs. H_1: $n > n_0$. In this case n_0^c can be calculated in the same way and the maximum size of the