A derivation for Armitage's trend test for the 2x3 genotype table

statgen.org

January 2007

We have sampled R cases and S controls.

	Number of disease alleles			
	0	1	2	Totals
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Totals	n_0	n_1	n_2	N

We want to test H_0 against H_1.

H_0: All entries in the table are proportional, vs.

H_1: Within a column, the absolute value of the difference between the probability of an observation being classified as "Case" or "Control" increases monotonically across the table.

We'll work with the difference between the values in the column so we first standardize the rows to have the same sums:

	Number of disease alleles			
	0	1	2	Totals
Cases	Sr_0	Sr_1	Sr_2	RS
Controls	Rs_0	Rs_1	Rs_2	RS

We choose a set of scores x_1, x_2, and x_3 and form the test statistic

$$U = \sum_{i=0}^{2} x_i(Sr_i - Rs_i).$$

Under $\mathrm{H}_0, \mathrm{P}(\text{Case} \mid \text{Count} = i) = \mathrm{P}(\text{Control} \mid \text{Count} = i) = n_i/N$ and E(U)=0.

Letting $S = N - R$, $s_i = n_i - r_i$, and choosing scores $x_0 = 0, x_1 = 1$, and $x_2 = 2$, the test statistic becomes

$$U = N(r_1 + 2r_2) - R(n_1 + 2n_2).$$

Returning to generic scores x_i, we calculate the variance of U as:

$$
\begin{aligned}
\mathrm{var}(U) &= \mathrm{var}(\sum_{i=0}^{2} x_i(Sr_i - Rs_i)) \\
&= \mathrm{var}(S\sum_{i=0}^{2} x_i r_i - R\sum_{i=0}^{2} x_i s_i) \\
&= S^2\mathrm{var}(\sum_{i=0}^{2} x_i r_i) + R^2\mathrm{var}(\sum_{i=0}^{2} x_i s_i) \\
&= S^2\left[\sum_{i=0}^{2} x_i^2\mathrm{var}(r_i) + 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j \mathrm{cov}(r_i, r_j)\right] \\
&\quad + R^2\left[\sum_{i=0}^{2} x_i^2\mathrm{var}(s_i) + 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j \mathrm{cov}(s_i, s_j)\right]
\end{aligned}
$$

(Under H_0)

$$
\begin{aligned}
&= S^2\left[\sum_{i=0}^{2} x_i^2 R\left(\frac{n_i}{N}\right)\left(\frac{N-n_i}{N}\right) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j R\left(\frac{n_i}{N}\right)\left(\frac{n_j}{N}\right)\right] \\
&\quad + R^2\left[\sum_{i=0}^{2} x_i^2 S\left(\frac{n_i}{N}\right)\left(\frac{N-n_i}{N}\right) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j S\left(\frac{n_i}{N}\right)\left(\frac{n_j}{N}\right)\right] \\
&= \frac{S}{N^2}\left[\sum_{i=0}^{2} x_i^2 SRn_i(N-n_i) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j SRn_i n_j\right] \\
&\quad + \frac{R}{N^2}\left[\sum_{i=0}^{2} x_i^2 SRn_i(N-n_i) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j SRn_i n_j\right] \\
&= \frac{SR}{N}\left[\sum_{i=0}^{2} x_i^2 n_i(N-n_i) - 2\sum_{i=0}^{1}\sum_{j=i+1}^{2} x_i x_j n_i n_j\right]
\end{aligned}
$$

(Let $S = N - R$. Choose scores $x_0 = 0,\ x_1 = 1,$ and $x_2 = 2$.)

$$
= \frac{(N-R)R}{N}\left[N(n_1 + 4n_2) - (n_1 + 2n_2)^2\right]
$$

For N large, we then have:

$$\begin{aligned} \frac{U}{\mathrm{SD}(U)} &= \frac{N(r_1 + 2r_2) - R(n_1 + 2n_2)}{\sqrt{\frac{(N-R)R}{N}\left[N(n_1 + 4n_2) - (n_1 + 2n_2)^2\right]}} \\ &\sim \mathrm{N}(0, 1) \end{aligned}$$

so, using Sasieni's (1997) notation,

$$\chi_G^2 = \frac{N\left[N(r_1 + 2r_2) - R(n_1 + 2n_2)\right]^2}{(N-R)R\left[N(n_1 + 4n_2) - (n_1 + 2n_2)^2\right]} \sim \chi_1^2.$$