Branching processes and Apert syndrome

Apert syndrome is a birth defect caused by mutation of either of two specific base pairs. The syndrome occurs at a rate of about 1 in 200,000 live births, much higher than the estimated 1 in 100,000,000 expected from the average estimated human mutation rate. Mutation rates do vary significantly across the genome (e.g. "hot spots"); but such a large variation is still very unexpected. It has also been observed that the father's age has an effect on the likelihood of occurrence – older fathers are much more likely to have offspring with the syndrome.

There is an alternate explanation, but first, some background. Here is a cartoon of how sperm is produced in the testes. There is a large population of stem cells which usually divide *asymmetrically*, producing one sperm and one stem cell, but may occasionally divide *symmetrically*, producing two stem cells. They may also, eventually, die. Suppose, then, that the mutation causing Apert syndrome also causes the stem cells to divide symmetrically more often. In an individual with the mutation, then, this would eventually lead to a disproportionately large number of mutated sperm, and thus a much larger chance of offspring with the syndrome than we'd expect from the mutation rate.

We will make assumptions that allow us to treat the number of stem cells present as a branching process (ignoring the sperm), and will denote the probabilities of symmetric division and death, respectively, by b and d. Since the number of stem cells is more-or-less constant, we assume that $b = d$ for nonmutated cells.

This then suggests two models:

(I) Mutations occur with (relatively) high probability $1000m_0$ per cell generation, but that $b = d$ for the mutated cells.

(II) Mutations occur with usual probability m_0 per cell generation, and that $b = d + s$ for some $s > 0$ for the mutated cells.

For data, suppose we have the ages and numbers of mutated sperm cells in a number of fathers of affected children. The goal then is to distinguish Model I from Model II. We make the following assumptions:

1. The number of mutated sperm cells at a given age is proportional to the number of mutated stem cells, so that we can pretend we have numbers of stem cells.

2. The number of nonmutated stem cells stays approximately constant at N, which is large.

3. Splitting and mutation (and everything else that is convenient) is independent between cells.

We make one final assumption. Since our sample is of fathers *of affected children*, and (half) the child's genotype represents a random sample from the father's sperm, there is some sort of size-biasing going on: fathers with more mutant

cells are more likely to be represented in our sample. However, the probability of sampling a father depends on the *proportion* of his sperm that are mutated; but we'd rather sweep the denominator here under the rug, and assume that if M_n is the (random) number of mutated stem cells in a random individual of age n (in cell generations) in the population, and S_n is the number in an individual of age n in our sample, that

4. $\mathbb{P}\{S_n = k\} = \frac{k\mathbb{P}\{M_n=k\}}{\mathbb{E}[M_n]}$, e.g. that S_n is a size-biased sample from M_n.

(for more on size-biasing, see e.g. http://www.stat.berkeley.edu/∼aldous/157/Writeups/class10.27.pdf)

We are therefore interested in the size-biased mean and variance of M_n, as a function of n, which we define to be

$$f(n) := \frac{\mathbb{E}[M_n^2]}{\mathbb{E}[M_n]}$$
$$s(n)^2 := \frac{\mathbb{E}[M_n^3]}{\mathbb{E}[M_n]} - \frac{\mathbb{E}[M_n^2]^2}{\mathbb{E}[M_n]^2}.$$

Model I

Here we imagine a relatively large number of mutations, each giving rise to a critical branching process, e.g. one with mean offspring number $\mu = 1$. Each such branching process dies out quickly with large probability, and does not grow too quickly (at least not geometrically) if it does, so we might imagine the population of mutant cells being composed of a largeish number of branching processes, each having arisen at different times in the past. A better model would be to include back mutation, and to treat this as a two- or three-type branching process, but for pedagogical purposes, we won't.

Thanks to our cartoon model, we can decompose M_n, the number of mutant cells at age n, as

$$M_n = \sum_{k=1}^{n} \sum_{\ell=1}^{N} X_{k,\ell} Z_{n-k}^{(k,\ell)},$$

where $X_{k,l}$ are i.i.d. Bernoulli(m) random variables, and $(Z_m^{(k,\ell)})_{m\geq 0}$ are i.i.d. branching processes with $Z_0 = 1$, independent of the Xs, and with offspring distribution

$$Y = \begin{cases} 1 & \text{with probability } 1-2b \\ 2 & \text{with probability } b \\ 0 & \text{with probability } b. \end{cases}$$

Note that $\sigma^2 = \mathrm{Var}[Y] = 2b$.

First let's compute the first two moments of M_n. Since $\mathbb{E}[Z_k] = 1$,

$$\begin{aligned}\mathbb{E}[M_n] &= \sum_{k=1}^{n}\sum_{\ell=1}^{N}\mathbb{E}[X_{k,\ell}Z_{n-k}^{(k,\ell)}\\ &= \sum_{k=1}^{n} N\mathbb{E}[XZ_{n-k}]\\ &= Nmn.\end{aligned}$$

Also, since $\mathrm{Var}[Z_k] = \sigma^2 n$, and (for any random variables U and V) $\mathrm{Var}[UV] = \mathrm{Var}[U]\mathbb{E}[V^2] + \mathbb{E}[U^2]\,\mathrm{Var}[V]$, and the terms are independent by assumption,

$$\begin{aligned}\mathrm{Var}[M_n] &= \sum_{k=1}^{n}\sum_{\ell=1}^{N}\mathrm{Var}[X_{k,\ell}Z_{n-k}^{(k,\ell)}]\\ &= \sum_{k=1}^{n} N\,\mathrm{Var}[XZ_{n-k}]\\ &= N\sum_{k=1}^{n}\left(m\,\mathrm{Var}[Z_{n-k}] + m(1-m)\mathbb{E}[Z_{n-k}]^2\right)\\ &= Nm\left(\frac{n(n-1)\sigma^2}{2} - n(1-m)\right)\\ &\approx \frac{1}{2}Nmn^2\sigma^2.\end{aligned}$$

Let U_k have the distribution of $X_{k,1}Z_{n-k}^{k,1}$. Then Yaglom's law tells us that for n large, $\mathbb{P}\{U_k > 0\} \approx m\frac{2}{n\sigma^2}$, and that conditioned on $\{U_k > 0\}$, that U_k is approximately exponentially distributed with mean $n\frac{\sigma^2}{2}$. Therefore, the number of contributions to M_n is approximately $\sum_{k=1}^{n} Nm\frac{2}{k\sigma^2} \approx (2Nm/\sigma^2)\log n$, and the ratio of the largest contribution to the total size is approximately $(\sigma^2 n/2)/(Nmn) = \sigma^2/2Nm$. So, M_n is the sum over a random collection of independent exponentials with varying means. We could proceed in different directions at this point (see Model II), but the easiest way to computing the third moment of M_n at this point seems to be to assume that "large" means that $m >> 1/N$, and that we can therefore use the central limit theorem.

If the CLT applies, then we have that M_n is approximately Gaussian, with mean equal to Nmn and variance approximately $Nmn\frac{n\sigma^2}{2}$. Then it is straightforward to compute

$$\begin{aligned}f(n) &= n(\sigma^2/2 + Nm)\\ s(n)^2 &= n^2\frac{\sigma^2}{2}(Nm - \sigma^2/2).\end{aligned}$$

Model II

Now the mutation probability m is smaller, but the mutations spread faster, and the $\mu = 1 + s > 1$. If we assume that there is only a *single* origin with high probability, then we might be able to make some progress determining the generating function of M_n, but we opt instead to just compute $f(n)$ and $s(n)^2$, using the asymptotic theory for a supercritical ($\mu > 1$) branching process. Recall that since $\mu > 1$, that for large k, $Z_k \approx \mu^n W$, where W is random variable with an atom at zero and continuous distribution otherwise on $(0, \infty)$.

Therefore, since $\mathbb{E}[W] = 1$,

$$\begin{aligned}
\mathbb{E}[M_n] &= \sum_{k=1}^{n} \sum_{\ell=1}^{N} \mathbb{E}[X_{k,\ell} Z_{n-k}^{(k,\ell)}] \\
&= \sum_{k=1}^{n} N\mu^{n-k} \mathbb{E}[XW] \\
&= Nm \frac{\mu^n - 1}{\mu - 1} \\
&\approx \frac{Nm}{s} \mu^n.
\end{aligned}$$

Also, letting $\xi^2 = \mathrm{Var}[W]$,

$$\begin{aligned}
\mathrm{Var}[M_n] &= \sum_{k=1}^{n} \sum_{\ell=1}^{N} \mathrm{Var}[X_{k,\ell} Z_{n-k}^{(k,\ell)}] \\
&= \sum_{k=1}^{n} N\mu^{2(n-k)} \mathrm{Var}[XW] \\
&= N \left(m\xi^2 + m(1-m) \right) \frac{\mu^2 n - 1}{\mu^2 - 1} \\
&\approx \frac{Nm(\xi^2 + 1)}{2s} \mu^2 n.
\end{aligned}$$

Finally, denote the third cumulant $\kappa_3[V] = \mathbb{E}[V^3] - 3\mathbb{E}[V^2]\mathbb{E}[V] + 2\mathbb{E}[V]^3$, and let $\kappa = \kappa_3[W]$. Then, using that to first order in m, that $\kappa_3[XW] \approx m\kappa$, and that the cumulant of the sum of independent random variable is the sum of their cumulants,

$$\begin{aligned}
\kappa_3[M_n] &= \sum_{k=1}^{n} \sum_{\ell=1}^{N} \kappa_3[X_{k,\ell} Z_{n-k}^{(k,\ell)}] \\
&= \sum_{k=1}^{n} N\mu^{3(n-k)} \mathrm{Var}[XW] \\
&\approx \frac{Nm\kappa}{3s} \mu^3 n.
\end{aligned}$$

Therefore, we get that

$$\begin{aligned} f(n) &= \frac{\mathrm{Var}[M_n] + \mathbb{E}[M_n]^2}{\mathbb{E}[M_n]} \\ &\approx \mu^n((\xi^2+1)/2 + Nm/s) \end{aligned}$$

and that

$$\begin{aligned} s(n)^2 &= \frac{\kappa_3[M_n] + 3\mathbb{E}[M_n]\,\mathrm{Var}[M_n] - \mathbb{E}[M_n]^3}{\mathbb{E}[M_n]} - f(n)^2 \\ &\approx \mu^{2n}\left(\frac{\kappa}{3} - \frac{\xi^2+1}{2} + \frac{Nm}{s}\left[\frac{3(\xi^2+1)}{2} - 1 - \frac{Nm}{s}\right]\right). \end{aligned}$$

Conclusion

In conclusion, the age-conditioned means in each case are quite different – in the first case, they grow linearly with time, and in the second, geometrically. Given some reasonable values for N and the ages of the sampled men in cell generations since puberty, as well as the total incidence in the population, we could use data to fit the parameters m and s for each case, see if the fitted parameters are at all reasonable, and then compare the fits.

The conclusion in the literature for Apert syndrome is that Model II is clearly correct, and it is suspected that many other mutations showing a male bias have a similar background story.

References

[1] Morrison, SJ, Kimble J. *Asymmetric and symmetric stem-cell divisions in development and cancer.* http://www.ncbi.nlm.nih.gov/pubmed/16810241

[2] Choi SK, Yoon SR, Calabrese P, Arnheim N. *A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations.* http://www.ncbi.nlm.nih.gov/pubmed/18632557