Notes to accompany Continuatio argumenti de mensura sortis ad fortuitam successionem rerum naturaliter contingentium applicata

Richard J. Pulskamp
Department of Mathematics and Computer Science
Xavier University, Cincinnati, Ohio

22 March 2006

This paper is, in some ways, the more interesting of the pair. Its topic is the comparison of two hypotheses concerning the ratio between male and female births — whether data supports a $1 : 1$ ratio or some other. However, in order to make this comparison, Bernoulli must develop methods of computing binomial probabilities corresponding to very large sample sizes.

However, Bernoulli, as he tests his theories upon tables of real data, is led to questions concerning correction of values and the stability of statistical ratios.

Throughout his paper the number of births is denoted by $2N$ and the ratio of male to female births as $a : b$. The probability that a birth is male is then estimated as $p = \frac{a}{a+b}$.

Let the random variable X have a binomial distribution with parameters $2N$ and p and we will identify two particular functions that will be found useful in describing the method of Bernoulli. Define the function q as

$$q(m) = \Pr(X = m|p = 1/2) = \binom{2N}{m}\left(\frac{1}{2}\right)^{2N}$$

and note that the maximum value of $q(m)$ occurs where $m = N$. This function q, which corresponds to the first hypothesis, will be seen to take on a kind of fiduciary role in his analysis. In the first paper, Bernoulli obtained an estimate of $q(N)$. We note that

$$q(N \pm \mu) \approx q(N)e^{-\mu^2/N},$$

a result which will ultimately be verified by Bernoulli.

Similarly, we define r as

$$r(m) = \Pr(X = m|p) = \binom{2N}{m}\left(\frac{a}{a+b}\right)^{m}\left(\frac{b}{a+b}\right)^{2N-m}$$

where $p = \frac{a}{a+b} \neq 1/2$. This function corresponds to his second or alternative hypothesis.

It is easy to see that

$$\frac{r(m)}{q(m)} = \left(\frac{a}{b}\right)^m \left(\frac{2b}{a+b}\right)^{2N} \tag{1}$$

and, in particular, that

$$r(N) = q(N) \times \left(\frac{a}{b}\right)^N \left(\frac{2b}{a+b}\right)^{2N}$$

It follows that it is easy to express the probability of any event under the second hypothesis in terms of the function q. Note that the computation of the second factor is trivial by use of logarithms regardless of the size of N. But equation (1) has another use. If a, b and m are given from data, then the two hypotheses may be compared as estimated and we may ask: Which hypothesis is more probable given the data?

We introduce, for fixed a and b, the function ϕ defined by

$$\phi(m) = \left(\frac{a}{b}\right)^m \left(\frac{2b}{a+b}\right)^{2N}$$

We have

$$r(N \pm \mu) \approx q(N) \times e^{-\mu^2/N} \times \phi(N \pm \mu)$$

It is quite clear how q plays a fiduciary role here in the computation of probabilities. However, given a, b and ϕ, the value of m may be determined and, consequently, the value of μ. Thus, Bernoulli is able to investigate deviations from N.

Example 1 (§ 3): In the previous paper, Bernoulli showed that when $N = 10000$, $q(N) = 0.005642 \approx \frac{1}{177}$. If now $\frac{a}{b} = 1.055$ (the observed ratio of male to female births), we find $\phi(N) = 0.000773 \approx \frac{1}{1294}$ and so $r(N) \approx \frac{1}{229401}$. Bernoulli, however, obtains the slightly larger figure of $\frac{1}{229392}$ since he finds $\phi(N) = \frac{1}{1296}$.

Finally, Bernoulli asserts

$$\Pr(X \leq 10,000 | a:b = 1055:1000) = \frac{1}{11469} \approx \frac{20}{229392}.$$

The probability $\frac{1}{11469}$ appears to be obtained by multiplying $r(10000)$ by 20. Implicitly, Bernoulli is saying that the tail probability is no more than 20 times the boundary value. This is not a bad estimate. We have from the normal approximation with continuity correction

$$\Pr(X \leq 10,000 | a:b = 1055:1000) = 0.00007864 \approx \frac{1}{12716}. \qquad \Box$$

Now let m denote the value for which $\phi(m) = 1, 2, 3, \ldots$. Routine algebra shows that we may write

$$m = A + \frac{\log \phi}{\log(a/b)}$$

where A corresponds to the case $\phi = 1$.

Example 2 (§ **6):** If $N = 10000$ and $a : b = 1.055$, then $A = 10134$ and we may write

$$m = 10134 + 43 \log_{10} \phi$$

It follows that if it were observed that the number of little boys born were 10134, then the evidence is equal for each of the two hypotheses. The special case, $\phi = 10^{10}$ yields $m = 10134 + 430 = 10564$.

What is $r(10564)$? Bernoulli gives $\frac{1}{1,416,000}$ which seems to have been applied here in error for it corresponds to using $\mu = 566$. We have

$$\begin{aligned} r(10564) =& \frac{1}{1,168,377} \\ &\approx q(10000) \times e^{-564^2/10000} \times 10^{10} \\ &\approx \frac{1}{1,156,949} \end{aligned}$$

Bernoulli then asserts, without proof, that the probability the number of boys will exceed 10564 is $\frac{10}{1416000}$, whereas $\frac{20}{1416000}$ would have been more accurate. In fact, we have, with the usual continuity correction for the normal distribution,

$$\Pr(X > 10564 | a : b = 1.055) \approx \frac{1}{74844}. \qquad \square$$

Example 3 (§ **8):** Bernoulli observes that we may just as easily put $\frac{r(m)}{q(m)} = \frac{1}{\phi}$ for which we will have

$$m = A - \frac{\log \phi}{\log(a/b)}$$

If all things are as before, we find $m = 10134 - 43 \log_{10} \phi = 10134 - 430 = 9704$. Now

$$q(9704) \approx q(10000) \times e^{-296^2/10000} \approx \frac{1}{177} \times \frac{1}{6384} = \frac{1}{1,129,968}$$

Bernoulli has used $\mu = 300$ so that $e^{-300^2/10000} \approx \frac{1}{8000}$ instead which yields $\frac{1}{141600}$ for the probability. Of course, $r(9704) = q(9704) \times 10^{-10}$.

By the normal approximation with continuity correction,

$$\Pr(X < 9704 | a : b = 1) \approx \frac{1}{72967}.$$

This then indicates that $\Pr(X < 9704 | a : b = 1) \approx 20 \times \Pr(X = 9704 | a : b = 1)$.

If we now combine these results we find that

$$\Pr(9704 \leq X \leq 10564 | a : b = 1.055) \approx 1 - \frac{1}{74844}. \qquad \square$$

Our next task is to find the maximum probability of r given a and b. Bernoulli observes that we should have at the maximum $r(m) = r(m + 1)$. This leads to the solution

$$M = \frac{2Na - b}{a + b} \quad \text{or, rather,} \quad M = \frac{2Na}{a + b}$$

by observing $2Na \gg b$ and therefore the term b in the numerator may be dropped. Of course, this is the standard result.

Example 4 (§ 12–16): If $N = 10000$ and $a/b = 1.055$, we have $M = \frac{2Na}{a+b} = 10268$. □

Bernoulli now seeks a formula for the probability of observing any number of boys out of $2N$ births. To this end he makes use of infinitesimal techniques. Let us put $M = \frac{2Na}{a+b}$, $m = M + \mu$ and $\pi = r(m)$. We seek a formula for π when a and b are nearly equal.

Now

$$r(m+1) = \frac{2N - m}{m+1} \times \frac{a}{b} \times \pi = \frac{2N - M - \mu}{M + \mu + 1} \times \frac{a}{b} \times \pi$$

Therefore

$$-\Delta\pi = \pi - \frac{2N - M - \mu}{M + \mu + 1} \times \frac{a}{b} \times \pi$$

corresponding to $\Delta\mu = 1$. Passing to differentials, we have

$$-\frac{d\pi}{\pi} = \left[1 - \frac{2N - M - \mu}{M + \mu + 1} \times \frac{a}{b}\right] d\mu$$

This differential equation, subject to the initial condition $Q = r(M)$ or rather, $\pi = Q$ when $\mu = 0$, has approximate solution

$$\pi = Qe^{-\mu^2/N}$$

as will now be demonstrated.

The solution to the differential equation is

$$-\ln\pi = \left(\frac{a+b}{b}\right)\mu - \left[\left(\frac{a+b}{b}\right)M + \frac{a}{b}\right]\ln(M + \mu + 1) + C$$

subject to the initial $\pi(0) = Q$. This gives

$$C = \left[\left(\frac{a+b}{b}\right)M + \frac{a}{b}\right]\ln(M+1) - \ln Q$$

Consequently, we may write

$$\ln\frac{Q}{\pi} = \frac{a+b}{b}\mu - \frac{a+b}{b}(M+1)\ln\frac{M+\mu+1}{M+1} + \ln\frac{M+\mu+1}{M+1}.$$

as Bernoulli finds.

To eliminate the factor of $M + 1$ in the second term, we make use of the fact that

$$\ln(1+x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 - \cdots$$

so that

$$\ln\frac{M+\mu+1}{M+1} = \ln\left(1 + \frac{\mu}{M+1}\right) = \frac{\mu}{M+1} - \frac{1}{2}\left(\frac{\mu}{M+1}\right)^2 + \frac{1}{3}\left(\frac{\mu}{M+1}\right)^3$$

truncated to 3 terms. Therefore

$$\ln\frac{Q}{\pi}=\frac{a+b}{b}\left(\frac{1}{2}\frac{\mu^2}{M+1}-\frac{1}{3}\frac{\mu^3}{(M+1)^2}\right)+\ln\frac{M+\mu+1}{M+1}$$

Further simplifications may be made. Since M is very large, we may replace $M+1$ by M. If further we assume $\mu \ll M$, then $\frac{\mu^3}{(M+1)^2}$ may be ignored as well as $\frac{M+\mu+1}{M+1}$. This then gives

$$\ln\frac{Q}{\pi}=\frac{a+b}{2b}\frac{\mu^2}{M}$$

Even more, if a is assumed nearly equal to b, we have

$$\ln\frac{Q}{\pi}=\frac{\mu^2}{M}$$

or

$$\pi=Qe^{-\mu^2/M}$$

Another approximation is obtained as follows:
Since $M=\frac{2Na}{a+b}$,

$$\frac{a+b}{2b}\times\frac{\mu^2}{M}=\frac{a+b}{2b}\times\frac{\mu^2(a+b)}{2Na}=\frac{(a+b)^2}{4ab}\times\frac{\mu^2}{N}$$

but the factor $\frac{(a+b)^2}{4ab}$ is very nearly 1 if a is little different from b. This then gives

$$\pi=Qe^{-\mu^2/N}$$

We return to the estimation of $r(M)$ where M is the maximum probability given by a and b. Since $M=\frac{2Na}{a+b}$, $\mu=|M-N|=|\frac{a-b}{a+b}|N$

$$q(M)=q(N)e^{-\mu^2/N}$$

and

$$r(M)=q(M)\times\left(\frac{a}{b}\right)^M\times\left(\frac{2b}{a+b}\right)^{2N}$$

Bernoulli wishes to argue that $r(M)\approx q(N)$ or equivalently, that

$$e^{\left(\frac{a-b}{a+b}\right)^2N}=\left(\frac{a}{b}\right)^{\frac{2Na}{a+b}}\times\left(\frac{2b}{a+b}\right)^{2N}$$

To this end, put $b=1$ and $a=1+\alpha$. We have

$$e^{\left(\frac{\alpha}{2+\alpha}\right)^2N}=(1+\alpha)^{\frac{2N(1+\alpha)}{2+\alpha}}\times\left(\frac{2}{2+\alpha}\right)^{2N}$$

which, by taking logarithms, is

$$\left(\frac{\alpha}{2+\alpha}\right)^2=\frac{2\alpha+2}{2+\alpha}\ln{(1+\alpha)}+2\ln\left(\frac{2}{2+\alpha}\right)$$

The Maclaurin series for each side agree up through order 3 and consequently Bernoulli concludes as long as α is very small, $r(M) \approx q(N)$.

By the previous paper,

$$q(N) = \frac{0.56413}{\sqrt{N}}$$

As an aside we note that the numerator is in fact $\frac{1}{\sqrt{\pi}}$ $(= 0.56413)$. We have

$$q(N \pm \mu) = \frac{0.56413}{\sqrt{N}} e^{-\mu^2/N}$$

Example 5 (§ **17–18):** Bernoulli first cites an example where of 5122 births, only 2020 were boys. Under the assumption that $a : b = 1.055$, the probability that this occur is 4.7×10^{-66}, somewhat larger than his estimate. In order to obtain his result, we find that the number of boys deviates from the hypothetical mean $M = 2651$ by -631. Consequently, he estimates he estimates this probability as

$$\frac{.56413}{\sqrt{N}} e^{-\mu^2/N} = \frac{.56413}{\sqrt{2561}} e^{-631^2/2561} = 3.3 \times 10^{-70}$$

In the same manner, if only 20 boys and 37 girls are born, under the assumption that $a : b = 1.055$, then the expected number of boys is 29.26. The deviation of the number of boys from this is -9.26. Therefore he estimates the probability as

$$\frac{.56413}{\sqrt{N}} e^{-\mu^2/N} = \frac{.56413}{\sqrt{28.5}} e^{-9.26^2/28.5} = 0.0052 \approx \frac{1}{192}. \qquad \square$$

In the previous paper, Bernoulli identified the quartiles of the binomial distribution with parameters $N = 10000$ and $p = \frac{1}{2}$ as $10000 \pm \mu$ where $\mu = 47\frac{1}{4}$. Thus

$$\Pr\left(9952\frac{3}{4} \leq X \leq 10047\frac{1}{4}\middle|p = \frac{1}{2}\right) \approx \frac{1}{2}$$

Actually, the probability is 0.4960.

What of $a : b = 1.055$? As long as μ is small relative to N and $a/b \approx 1$ the one distribution may be substituted for the other changing only the location of the center. Thus, since we have $M = 10268$,

$$\Pr\left(9952\frac{3}{4} + 268 \leq X \leq 10047\frac{1}{4} + 268\middle|p = 1.055\right) = \frac{1}{2}$$

or

$$\Pr\left(10220\frac{3}{4} \leq X \leq 10315\frac{1}{4}\middle|p = 1.055\right) = \frac{1}{2}$$

In reality, the probability is 0.4962. In general, we should observe that the quartiles lie at

$$\mu = 0.4725\sqrt{N}.$$

and that the corresponding interquartile range is

$$\frac{2Na}{a+b} \pm 0.4725\sqrt{N}$$

Thus, assuming as he does $2N = 600000$ offspring and $a : b = 1.055$, we have

$$\frac{2Na}{a+b} \pm 0.4725\sqrt{N} = \frac{600000 * 1055}{2055} \pm 0.4725\sqrt{300000} = 308029 \pm 259$$

Example 6 (§ 22): Bernoulli exhibits a table of births for the ten year period 1721–1730 in which these interquartile ranges are implicitly constructed. For example, in the year 1721, $b = 8940$ girls were born and $a = 9430$ boys thereby giving a total of $2N = 18370$.

The quartiles lie at a distance of $0.4725\sqrt{9185} = 45$ from the mean. Under the hypothesis that $a : b = 1.055$, the expected number of boys born is $M = 9431$ which exceeds the observed value by but 1 and so falls well within the interquartile range. For this reason, Bernoulli marks this case as NB.

Overall the ratio of boys to girls in this data is $a : b = 1.040$. Under this hypothesis, the expected number of boys born is $M = 9365$ which falls short of the observed number by 65.