Sampling from a couple of positively correlated binomial variates

Mario Catalani
Department of Economics, University of Torino, Italy
mario.catalani@unito.it

1 Introduction

Let the random variables \{X_1, X_2, X_3, X_4\} possess a multinomial distribution \(MN(n, p_1, p_2, p_3, p_4)\) (see [1], [2]). Just to fix notation in this case the joint probability function is

\[ P\{X_1 = x_1, \ldots, X_4 = x_4\} = P(x_1, \ldots, x_4) = n! \prod_{i=1}^{4} \frac{p_i^{x_i}}{x_i!}, \]

where \(\sum_{i=1}^{4} p_i = 1, \{x_i\}\) are integers \(\geq 0\) and \(\sum_{i=1}^{4} X_i = n\).

Now let \({i_j}\), \(j = 1, \ldots, 4\) define a permutation of the numbers \{1, 2, 3, 4\} and define \(W = X_{i_1} + X_{i_2}\). Then

\[ P\{W = k, X_{i_3} = x_{i_3}, X_{i_4} = n - k - x_{i_3}\} = \sum_{i=0}^{k} P\{X_{i_1} = i, X_{i_2} = k - i, X_{i_3} = x_{i_3}, X_{i_4} = n - k - x_{i_3}\} \]

\[ = \sum_{i=0}^{k} \frac{n!}{i!(k-i)!x_{i_3}!(n-k-x_{i_3})!} p_{i_1}^{i} p_{i_2}^{k-i} p_{i_3}^{x_{i_3}} p_{i_4}^{n-k-x_{i_3}} \]

\[ = \frac{n! p_{i_3}^{x_{i_3}} p_{i_4}^{n-k-x_{i_3}}}{x_{i_3}!(n-k-x_{i_3})!k!} \sum_{i=0}^{k} \frac{k!}{i!(k-i)!} p_{i_1}^{i} p_{i_2}^{k-i} \]

\[ = \frac{n! p_{i_3}^{x_{i_3}} p_{i_4}^{n-k-x_{i_3}}}{x_{i_3}!(n-k-x_{i_3})!k!} (p_{i_1} + p_{i_2})^k, \]

because of the binomial theorem.
It follows \( \{W, X_{i1}, X_{i2}\} \sim \mathcal{MN}(n, p_{i1} + p_{i2}, p_{i3}, p_{i4}) \). Because the marginals in a multinomial distribution are binomial variables with the appropriate parameters we can conclude

\[
W = X_{i1} + X_{i2} \sim \mathcal{B}(n, p_{i1} + p_{i2}).
\]

Using this result define

\[
\begin{align*}
Y_1 &= X_1 + X_3, \\
Y_2 &= X_2 + X_3.
\end{align*}
\]

Then \( Y_1 \sim \mathcal{B}(n, p_1 + p_3) \), \( Y_2 \sim \mathcal{B}(n, p_2 + p_3) \). Because in a multinomial distribution \( \text{Cov}(X_i, X_j) = -np_ip_j \) we have

\[
\text{Cov}(Y_1, Y_2) = -np_1p_2 - np_1p_3 - np_2p_3 + np_3(1 - p_3)
\]
\[
= -np_1p_2 + np_3(1 - p_1 - p_2 - p_3)
\]
\[
= -np_1p_2 + np_3p_4
\]
\[
= n(p_3p_4 - p_1p_2),
\]

so we have a positive covariance if \( p_3p_4 - p_1p_2 > 0 \).

Finally the linear correlation coefficient between \( Y_1 \) and \( Y_2 \) is given by

\[
\rho = \frac{p_3p_4 - p_1p_2}{\sqrt{p_1 + p_3}(1 - p_1)(1 - p_3)(p_2 + p_3)(1 - p_2 - p_3)}
\]

Now suppose we want to sample from two given binomial variables with a given positive correlation coefficient \( r \). Let the given variables be \( Y_1 \sim \mathcal{B}(n, \pi_1) \) and \( Y_2 \sim \mathcal{B}(n, \pi_2) \). So we are assuming given the parameters \( \pi_1, \pi_2, r \). We will use the previous framework determining the parameters \( p_1, p_2, p_3 \) as functions of the new parameters. We set

\[
p_1 + p_3 = \pi_1 \quad \text{and} \quad p_2 + p_3 = \pi_2.
\]

Then

\[
r = \frac{p_3p_4 - p_1p_2}{\sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}}.
\]

Then

\[
p_3p_4 - p_1p_2 = r \sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}.
\]

Inserting here \( p_4 = 1 - p_1 - p_2 - p_3, p_1 = \pi_1 - p_3 \) and \( p_2 = \pi_2 - p_3 \) we get

\[
p_3 = \pi_1\pi_2 + r \sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}.
\]
Then
\[ p_1 = \pi_1 - \pi_1 \pi_2 - r \sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}, \] (3)
\[ p_2 = \pi_2 - \pi_1 \pi_2 - r \sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}. \] (4)

The only problem is that we have to make sure that both \( p_1 \) and \( p_2 \) turn out to be positive. So we have to impose
\[ \pi_1(1 - \pi_2) > r \sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}, \] (5)
and
\[ \pi_2(1 - \pi_1) > r \sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}. \] (6)

This impose restrictions on the attainable upper bound for \( r \). Indeed it has to be
\[ r < \min \left( \sqrt{\frac{\pi_1(1 - \pi_2)}{\sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}}, \sqrt{\frac{\pi_2(1 - \pi_1)}{\sqrt{\pi_1(1 - \pi_1)\pi_2(1 - \pi_2)}}} \right). \]

Write \( \pi_1 = \beta \pi_2, \beta > 0 \). Then the above equation becomes
\[ r < \min \left( \sqrt{\frac{\beta(1 - \pi_2)}{1 - \beta \pi_2} \frac{1 - \beta \pi_2}{\beta(1 - \pi_2)}} \right). \]

If \( \beta > 1 \) the minimum is
\[ \sqrt{\frac{1 - \beta \pi_2}{\beta(1 - \pi_2)}}. \]

If \( \beta < 1 \) the minimum is
\[ \sqrt{\frac{\beta(1 - \pi_2)}{1 - \beta \pi_2}}. \]

If \( \beta = 1 \) the minimum is 1. So when \( Y_1 \) and \( Y_2 \) are identically distributed there is no constraint to impose on \( r \).

Now, when \( \beta > 1 \), if we take the derivative of the minimum with respect to \( \beta \), if we write
\[ f(\beta) = \frac{1 - \beta \pi_2}{\beta(1 - \pi_2)^2}, \]
we have
\[ f'(\beta) = -\frac{1 - \pi_2}{\beta^2(1 - \pi_2)^2} < 0, \]
so that the derivative of the minimum is negative: the minimum is a decreasing function of $\beta$.
When $\beta < 1$, if we take the derivative of the minimum with respect to $\beta$, if we write
\[ f(\beta) = \frac{\beta(1 - \pi_2)}{1 - \beta \pi_2}, \]
we have
\[ f'(\beta) = \frac{(1 - \pi_2)(1 + \pi_2(1 - \beta))}{\beta^2(1 - \pi_2)^2} > 0, \]
so that the derivative of the minimum is positive: the minimum is an increasing function of $\beta$.

In Table 1 we present the upper bound of $r$ for some couples of values of $\pi_1$ and $\pi_2$. Let us note that this upper bound is the same if we interchange $\pi_1$ with $\pi_2$ and if we substitute $\pi_1$ with $1 - \pi_1$ and $\pi_2$ with $1 - \pi_2$.

2 Regression Function

Let us recall two consequences of the binomial theorem:
\[ \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} = na(a + b)^{n-1}. \] (7)
\[ \sum_{i=0}^{n} i^2 \binom{n}{i} a^i b^{n-i} = na(a + b)^{n-2}[a + b + (n-1)a]. \] (8)
Because of Equation 1 the joint probability function of $X_1$ and $X_2 + X_3$ is given by
\[ P\{X_1 = k, X_2 + X_3 = h, X_4 = n - k - h\} = \frac{n!p_1^k p_4^{n-k-h}}{k!(n-k-h)!h!(p_2 + p_3)^h}. \]
Because of Equation 2 the conditional distribution of $X_1$ given $X_2 + X_3 = h$ is
\[ P\{X_1 = k | X_2 + X_3 = h\} = \frac{p_1^k p_4^{n-k-h}(n-h)!}{k!(n-k-h)!(1-p_2-p_3)^{n-h}}. \]
Then the conditional expectation of $X_1$ given $X_2 + X_3 = h$ is given by, noticing that the above conditioning event implies $0 \leq X_1 \leq n - h$,
\[ E(X_1 | X_2 + X_3 = h) = \frac{1}{(1-p_2-p_3)^{n-h}} \sum_{k=0}^{n-h} \binom{n-h}{k} p_1^k p_4^{n-k-h} \]
\[ = \frac{1}{(1-p_2-p_3)^{n-h}}(n-h)p_1(p_1 + p_4)^{n-h-1}, \]
where we used Equation 7. Now, since \( p_1 + p_4 = 1 - p_1 - p_2 \), we get
\[
E(X_1|X_2 + X_3 = h) = \frac{p_1}{1 - p_2 - p_3}(n - h).
\]
Finally the conditional expectation of \( X_1 \) given \( X_2 + X_3 \) is
\[
E(X_1|X_2 + X_3) = \frac{p_1}{1 - p_2 - p_3}(n - X_2 - X_3). \tag{9}
\]
Along the same lines we evaluate now the joint probability function of \( X_3 \) and \( X_2 + X_3 \). We obtain
\[
P\{X_3 = h, X_2 + X_3 = k\} = \sum_{i=0}^{n-k} P\{X_1 = i, X_2 = k - h, X_3 = h, X_4 = n - k - i\}
\]
\[
= \sum_{i=0}^{n-k} \frac{n!}{i!(k-h)!h!(n-k-i)!} p_1^{i} p_2^{k-h} p_3^{h} p_4^{n-k-i}
\]
\[
= \frac{n!}{k!(n-k)!} \sum_{i=0}^{n-k} \frac{i}{i!(n-k-i)!} p_1^{i} p_4^{n-k-i}
\]
\[
= \frac{k!}{k!(n-k)!} \sum_{i=0}^{n-k} \frac{i}{i!(n-k-i)!} p_1^{i} p_4^{n-k-i}
\]
\[
= \binom{k}{h} \frac{n!}{(k-h)!h!} p_2^{k-h} p_3^{h} (p_1 + p_4)^{n-k}.
\]
Then the conditional distribution of \( X_3 \) given \( X_2 + X_3 \) is
\[
P\{X_3 = h|X_2 + X_3 = k\} = \binom{k}{h} \frac{p_2^{k-h} p_3^{h}}{(p_2 + p_3)^k}.
\]
Hence the conditional expectation of \( X_3 \) given \( X_2 + X_3 = k \) is given by, noticing that the above conditioning event implies \( 0 \leq X_3 \leq k \),
\[
E(X_3|X_2 + X_3 = k) = \sum_{h=0}^{k} h \binom{k}{h} \frac{p_2^{k-h} p_3^{h}}{(p_2 + p_3)^k}
\]
\[
= \frac{p_2^{k}}{(p_2 + p_3)^k} \sum_{h=0}^{k} h \binom{k}{h} \left( \frac{p_3}{p_2} \right)^h
\]
\[
= \frac{p_2^{k}}{(p_2 + p_3)^k} k \frac{p_3}{p_2} \left( 1 + \frac{p_3}{p_2} \right)^{k-1}
\]
\[
= \frac{kp_3}{p_2 + p_3}.
\]

5
Finally the conditional expectation of $X_3$ given $X_2 + X_3$ is

$$E(X_3 | X_2 + X_3) = \frac{p_3}{p_2 + p_3} (X_2 + X_3).$$

(10)

It follows

$$E(Y_1 | Y_2) = E(X_1 + X_3 | X_2 + X_3)$$

$$= E(X_1 | X_2 + X_3) + E(X_3 | X_2 + X_3)$$

$$= \frac{p_1}{1 - p_2 - p_3} (n - X_2 - X_3) + \frac{p_3}{p_2 + p_3} (X_2 + X_3)$$

$$= \alpha n + (\beta - \alpha) Y_2,$$

(11)

where we wrote $\alpha = \frac{p_1}{1 - p_2 - p_3}$ and $\beta = \frac{p_3}{p_2 + p_3}$. We can conclude that the regression function is linear.

### 3 Conditional Variance

Since

$$P\{X_1 = h, X_2 + X_3 = k, X_3 = r, X_4 = n - k - h\} = \frac{n!}{h!(k - r)!r!(n - k - h)!} p_1^{h} p_2^{k-r} p_3^{r} p_4^{n-k-h},$$

the joint conditional probability function of $X_1$ and $X_3$ given $X_2 + X_3$ is obtained as

$$P\{X_1 = h, X_3 = r | X_2 + X_3 = k\} = \binom{k}{r} \binom{n-k}{h} \frac{p_1^{h} p_2^{k-r} p_3^{r} p_4^{n-k-h}}{(p_2 + p_3)^k (1 - p_2 - p_3)^{n-k}}.$$  

Since

$$P\{X_1 = h | X_2 + X_3 = k\} = \binom{n-k}{h} \frac{p_1^{h} p_4^{n-k-h}}{(1 - p_2 - p_3)^{n-k}},$$

and

$$P\{X_3 = r | X_2 + X_3 = k\} = \binom{k}{r} \frac{p_2^{k-r} p_3^{r}}{(p_2 + p_3)^k},$$

we see that $X_1$ and $X_3$ are conditionally independent given $X_2 + X_3$. It follows that

$$Var(X_1 + X_3 | X_2 + X_3) = Var(X_1 | X_2 + X_3) + Var(X_3 | X_2 + X_3).$$

6
We obtain
\[
\mathbf{E}(X_1^2|X_2 + X_3 = h) = \frac{1}{(1 - p_2 - p_3)^{n-h}} \sum_{k=0}^{n-h} k^2 \binom{n-h}{k} p_1 p_4^{n-k-h}
\]
\[
= \frac{1}{(1 - p_2 - p_3)^{n-h}} (n-h)p_1(p_1 + p_4)^{n-h-2} \times [p_1 + p_4 + (n-h-1)p_1]
\]
\[
= \frac{1}{(p_1 + p_4)^2} (n-h)p_1[p_1 + p_4 + (n-h-1)p_1],
\]
where we used Equation 8 and the fact that \(1 - p_2 - p_3 = p_1 + p_4\). Thus
\[
\mathbf{V}ar(X_1|X_2 + X_3 = h) = \frac{p_1 p_4}{(p_1 + p_4)^2} (n-h),
\]
that is
\[
\mathbf{V}ar(X_1|Y_2) = \frac{p_1 p_4}{(p_1 + p_4)^2} (n-Y_2).
\]
Analogously
\[
\mathbf{E}(X_3^2|X_2 + X_3 = h) = \frac{p_2^h}{(p_2 + p_3)^h} \sum_{k=0}^{h} k^2 \binom{h}{k} \left(\frac{p_3}{p_2}\right)^k
\]
\[
= \frac{p_2^h}{(p_2 + p_3)^h} h \binom{p_3}{p_2} \left(1 + \frac{p_3}{p_2}\right)^{h-2} \times [1 + \frac{p_3}{p_2} + (h-1)\left(\frac{p_3}{p_2}\right)]
\]
\[
= \frac{h p_3}{(p_2 + p_3)^2} [p_2 + p_3 + (h-1)p_3],
\]
where we used Equation 8. It follows
\[
\mathbf{V}ar(X_3|X_2 + X_3 = h) = \frac{h p_2 p_3}{(p_2 + p_3)^2},
\]
that is
\[
\mathbf{V}ar(X_3|Y_2) = \frac{p_2 p_3}{(p_2 + p_3)^2} Y_2.
\]
Finally
\[
\mathbf{V}ar(Y_1|Y_2) = \gamma + \delta Y_2,
\]
where we set \(\gamma = \frac{np_1 p_4}{(p_1 + p_4)^2}\) and \(\delta = \frac{p_2 p_3}{(p_2 + p_3)^2} - \frac{p_1 p_4}{(p_1 + p_4)^2}\). We can conclude that the regression function is linear but not homoscedastic.
Table 1. Upper bound of $r$ for selected values of $\pi_1$ and $\pi_2$.

| $\pi_1$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
|---------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0.1     | 1   | 0.667 | 0.509 | 0.409 | 0.333 | 0.272 | 0.218 | 0.167 | 0.111 |
| 0.2     | 0.667 | 1   | 0.767 | 0.612 | 0.500 | 0.408 | 0.327 | 0.250 | 0.167 |
| 0.3     | 0.509 | 0.767 | 1   | 0.802 | 0.655 | 0.534 | 0.428 | 0.327 | 0.218 |
| 0.4     | 0.409 | 0.612 | 0.802 | 1   | 0.816 | 0.667 | 0.534 | 0.408 | 0.272 |
| 0.5     | 0.333 | 0.500 | 0.655 | 0.816 | 1   | 0.816 | 0.655 | 0.500 | 0.333 |
| 0.6     | 0.272 | 0.409 | 0.534 | 0.667 | 0.816 | 1   | 0.802 | 0.612 | 0.409 |
| 0.7     | 0.218 | 0.327 | 0.428 | 0.534 | 0.655 | 0.802 | 1   | 0.767 | 0.509 |
| 0.8     | 0.167 | 0.250 | 0.327 | 0.409 | 0.500 | 0.612 | 0.767 | 1   | 0.667 |
| 0.9     | 0.111 | 0.167 | 0.218 | 0.272 | 0.333 | 0.409 | 0.509 | 0.667 | 1   |

References

[1] Fishman, G.S., *Monte Carlo. Concepts, Algorithms, and Applications*, Springer, New York, 1996.

[2] Johnson, N.L. and S. Kotz, *Distributions in Statistics: Continuous Multivariate Distributions*, Wiley, New York, 1972.