The Hypergeometric Distribution

Klaus Pommerening
Fachbereich Physik, Mathematik, Informatik
der Johannes-Gutenberg-Universität
Saarstraße 21
D-55099 Mainz

February 2, 2003—English version February 14, 2015
last change February 15, 2015

The urn problem underlying the hypergeometric distribution is "drawing without replacement". Assume the urn contains n balls s of which are black, and $t = n - s$ are white. Let

$$p := \frac{s}{n}$$

be the proportion of black balls, and assume without loss of generality that $p > \frac{1}{2}$. (The case $p = \frac{1}{2}$ is not interesting, the case $p < \frac{1}{2}$ is symmetric to the considered case.)

Draw r balls ($r \leq n$) by random. The probability that exactly ν of the balls are white is

$$q_r^{(s)}(\nu) = \frac{\binom{s}{r-\nu}\binom{t}{\nu}}{\binom{n}{r}}.$$

The function

$$q_r^{(s)} : \mathbb{Z} \longrightarrow \mathbb{R}$$

is called the **hypergeometric distribution** (with parameters n, s, and r). We have $q_r^{(s)}(\nu) = 0$ for $\nu < 0$ as well as for $\nu > r$. The probability of drawing more blacks balls than white ones is

$$p_r^{(s)} = \begin{cases} \sum_{\nu=0}^{\frac{r-1}{2}} q_r^{(s)}(\nu) & \text{if } r \text{ is odd,} \\ \sum_{\nu=0}^{\frac{r}{2}-1} q_r^{(s)}(\nu) + \frac{1}{2} q_r^{(s)}(\frac{r}{2}) & \text{if } r \text{ is even,} \end{cases}$$

in case of a tie we randomly decide between black and white with probability $\frac{1}{2}$.

In the uninteresting case $p = \frac{1}{2}$ obviously all $p_r^{(s)} = \frac{1}{2}$.

Lemma 1
(i) $p_1^{(s)} = p$.
(ii) $p_2^{(s)} = p_1^{(s)}$ (if $t \geq 1$).
(iii) $p_3^{(s)} = \frac{s(s-1)}{n(n-1)} \cdot \left[3 - 2 \cdot \frac{s-2}{n-2}\right]$ (if $t \geq 2$).
(iv) $p_4^{(s)} = p_3^{(s)}$ (if $t \geq 2$).
(v) $p_r^{(s)} = 1$ for $r > 2t$.

Proof. (i) Trivial.

(ii) We draw two balls, and break the tie (in the case where we draw one ball of each type) by a random decision. Therefore the numerator is

$$\binom{s}{2} + \frac{1}{2}\binom{s}{1}\binom{t}{1} = \frac{s(s-1)}{2} + \frac{s(n-s)}{2} = \frac{s(n-1)}{2}.$$

The denominator is $\frac{n(n-1)}{2}$, and the quotient is

$$p_2^{(s)} = \frac{s(n-1)}{n(n-1)} = p.$$

(iii) Here the numerator is

$$\begin{aligned} \binom{s}{3} + \binom{s}{2} \cdot (n-s) &= \frac{s(s-1)(s-2) + 3s(s-1)(n-s)}{6} \\ &= \frac{s(s-1)}{6} \cdot [s - 2 + 3 \cdot (n-s)] \\ &= \frac{s(s-1)}{6} \cdot [3 \cdot (n-2) - 2 \cdot (s-2)]. \end{aligned}$$

The denominator is $\frac{1}{6} \cdot n(n-1)(n-2)$, hence the asserted value of $p_3^{(s)}$.

(iv) We omit the calculation since the next lemma contains a more general statement.

(v) In this case we necessarily draw a majority of black balls. ◇

Lemma 2 *If r is even and $2 \leq r \leq 2t$, then*

$$p_{r+1}^{(s)} > p_r^{(s)} = p_{r-1}^{(s)}.$$

Proof. Let $A_r^{(s)}(\nu) = \binom{n}{r} \cdot q_r^{(s)}(\nu)$ be the numerator of $q_r^{(s)}(\nu)$, and $B_r^{(s)} = \binom{n}{r} \cdot p_r^{(s)}$, the numerator of $p_r^{(s)}$.

After $r+1$ drawings we have a black majority in $B_{r+1}^{(s)}$ cases. Considering the change from r to $r+1$ we have:

- $\sum_{\nu=0}^{\frac{r}{2}-1} A_r^{(s)}(\nu)$ cases where the number of black balls is at least $\frac{r}{2}+1$ after r drawings. We have $n-r$ possibilities for the $(r+1)$-th ball, but all of these cannot change the majority. So we get

$$X_1 = (n-r) \cdot \sum_{\nu=0}^{\frac{r}{2}-1} A_r^{(s)}(\nu)$$

  cases with a black majority.

- $A_r^{(s)}(\frac{r}{2})$ cases where after r drawings we have exactly $\frac{r}{2}$ black balls. From the $n-r$ possibilities for the $(r+1)$-th ball
  - $s-\frac{r}{2}$ are black and give a black majority,
  - $t-\frac{r}{2}$ are white and give a white majority.

  Thus we get another

$$X_2 = (s-\frac{r}{2}) \cdot A_r^{(s)}(\frac{r}{2})$$

  cases with a black majority.

- In the remaining cases after r drawings we have at most $\frac{r}{2}-1$ black balls. Therefore the $(r+1)$-th ball cannot change the white majority.

This count contains each resulting set exactly $r+1$ times. Therefore

$$B_{r+1}^{(s)} = \frac{1}{r+1} \cdot (X_1 + X_2) = \frac{n-r}{r+1} \cdot \left[\sum_{\nu=0}^{\frac{r}{2}-1} A_r^{(s)}(\nu) + \frac{s-\frac{r}{2}}{n-r} \cdot A_r^{(s)}(\frac{r}{2}) \right].$$

For the coefficient of the last term we have

$$\frac{s-\frac{r}{2}}{n-r} > \frac{1}{2} \Longleftrightarrow 2s-r > n-r \Longleftrightarrow s > \frac{n}{2}.$$

(Since $r \leq 2t$ also $r < n$.) Therefore

$$B_{r+1}^{(s)} > \frac{n-r}{r+1} \cdot B_r^{(s)},$$

and the first part of the assertion, $p_{r+1}^{(s)} > p_r^{(s)}$, follows.

Analyzing the change from $r-1$ to r is somewhat more complicated. After r drawings we have a black majority in $B_r^{(s)}$ cases. Among these are:

- $\sum_{\nu=0}^{\frac{r}{2}-2} A_{r-1}^{(s)}$ cases where after $r-1$ drawings we have at least $\frac{r}{2}+1$ black balls. The $n-r+1$ possibilities for the r-th ball can't change the decision. Hence we get

$$Y_1 = (n-r+1) \cdot \sum_{\nu=0}^{\frac{r}{2}-2} A_{r-1}^{(s)}$$

  cases with black majority.

- $A^{(s)}_{r-1}(\frac{r}{2}-1)$ cases where after $r-1$ drawings we have exactly $\frac{r}{2}$ black balls. The $n-r+1$ possibilities for the r-th ball dissociate into
  - $s-\frac{r}{2}$ black ones that result in a black majority. This makes
$$Y_2 = (s-\frac{r}{2})\cdot A^{(s)}_{r-1}(\frac{r}{2}-1)$$
    additional cases.
  - $t+1-\frac{r}{2}$ white ones where we randomly decide with probability $\frac{1}{2}$. This adds another
$$Y_3 = \frac{1}{2}\cdot(t+1-\frac{r}{2})\cdot A^{(s)}_{r-1}(\frac{r}{2}-1)$$
    cases to our collection.
- $A^{(s)}_{r-1}(\frac{r}{2})$ cases where after $r-1$ drawings we have exactly $\frac{r}{2}-1$ black balls. The $n-r+1$ possibilities for the r-th ball dissociate into
  - $s+1-\frac{r}{2}$ black ones where we randomly decide with probability $\frac{1}{2}$. This gives another
$$Y_4 = \frac{1}{2}\cdot(s+1-\frac{r}{2})\cdot A^{(s)}_{r-1}(\frac{r}{2})$$
    cases.
  - $t-\frac{r}{2}$ white ones that don't disturb the white majority.
- In the remaining cases after $r-1$ drawings we have at most $\frac{r}{2}-2$ black balls. The white majority is unchanged.

Each set of drawn balls is counted exactly r times. Therefore

$$\begin{aligned} B^{(s)}_r &= \frac{1}{r}\cdot(Y_1+Y_2+Y_3+Y_4) \\ &= \frac{n-r+1}{r}\cdot\sum_{\nu=0}^{\frac{r}{2}-2} A^{(s)}_{r-1} + \frac{1}{r}\cdot(s-\frac{r}{2}+\frac{t}{2}+\frac{1}{2}-\frac{r}{4})\cdot A^{(s)}_{r-1}(\frac{r}{2}-1) \\ &\quad +\frac{1}{2r}\cdot(s-\frac{r}{2}+1)\cdot A^{(s)}_{r-1}(\frac{r}{2}) \end{aligned}$$

Since $s+\frac{t}{2} = n-\frac{t}{2}$ the coefficient of the middle term equals

$$s-\frac{r}{2}+\frac{t}{2}-\frac{r}{4}+\frac{1}{2} = n-\frac{t}{2}-r+\frac{r}{4}+1-\frac{1}{2} = (n-r+1)-\frac{1}{2}\cdot(t-\frac{r}{2}+1).$$

Hence

$$\begin{aligned} B^{(s)}_r &= \frac{n-r+1}{r}\cdot\sum_{\nu=0}^{\frac{r}{2}-1} A^{(s)}_{r-1} \\ &\quad -\frac{1}{2r}(t-\frac{r}{2}+1)\binom{s}{\frac{r}{2}}\binom{t}{\frac{r}{2}-1} + \frac{1}{2r}(s-\frac{r}{2}+1)\binom{s}{\frac{r}{2}-1}\binom{t}{\frac{r}{2}}. \end{aligned}$$

The two last terms cancel. What remains is

$$B_r^{(s)} = \frac{n-r+1}{r} \cdot B_{r-1}^{(s)}.$$

This proves the second part of the assertion. ◇

We conclude:

Proposition 1 *The probability $p_r^{(s)}$ grows monotonically with r from $p_1^{(s)} = p$ to $p_{2t+1}^{(s)} = 1$.*

If the quotients

$$\frac{rs}{n}, \frac{rt}{n}, \frac{(n-r)s}{n}, \frac{(n-r)t}{n}$$

are sufficiently large (by FISHER's rule of thumb: ≥ 5), the normal distribution approximates the hypergeometric distribution well. In particular

$$\sum_{\nu=0}^{x} q_r^{(s)}(\nu) \approx \Phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-t^2/2}\, dt \qquad (1)$$

where μ is the mean value and σ^2 is the variance of the hypergeometric distribution (with parameters n, s, and r), and Φ is the distribution function of the normal distribution. For mean value and variance we have:

Lemma 3

$$\begin{aligned} \mu &= \frac{rt}{n}, \\ \sigma^2 &= \frac{r(n-r)\cdot t(n-t)}{n^2(n-1)}. \end{aligned}$$

Proof. Take a random sample of r balls. Let $X_k : \Omega \longrightarrow \mathbb{R}$ be a random variable that assumes the value 0 if the k-th ball is black, and 1 if it is white. Then $S = X_1 + \cdots + X_r : \Omega \longrightarrow \mathbb{R}$ is a random variable that counts the number of white balls in our sample. Then $\mu = \mathrm{E}(S)$ is the expectation and $\sigma^2 = \mathrm{Var}(S)$ is the variance of this random variable.

Since $\mathrm{E}(X_k) = \frac{t}{n}$ we have $\mathrm{E}(S) = r \cdot \frac{t}{n}$.

We note that $X_k^2 = X_k$ and derive

$$\mathrm{Var}(X_k) = \mathrm{E}(X_k^2) - \mathrm{E}(X_k)^2 = \frac{t}{n} - \frac{t^2}{n^2} = \frac{t(n-t)}{n^2}.$$

Since $X_jX_k(\omega) = 1 \iff X_j(\omega) = 1$ *and* $X_k(\omega) = 1$ the probability of this event is $\frac{t(t-1)}{n(n-1)}$. This gives the expectation $\mathrm{E}(X_jX_k) = \frac{t(t-1)}{n(n-1)}$. Thus the covariance is

$$\begin{aligned} \mathrm{Cov}(X_j, X_k) &= \mathrm{E}(X_jX_k) - \mathrm{E}(X_j)\mathrm{E}(X_k) = \frac{t(t-1)}{n(n-1)} - \frac{t^2}{n^2} \\ &= \frac{t(n(t-1) - t(n-1))}{n^2(n-1)} = \frac{t(t-n)}{n^2(n-1)}. \end{aligned}$$

We deduce the variance of S:

$$\begin{aligned}\mathrm{Var}(S) &= \sum_{k=1}^{r}\mathrm{Var}(X_k) + 2\cdot\sum_{1\le j<k\le r}\mathrm{Cov}(X_j,X_k)\\ &= \frac{rt(n-t)}{n^2} + r(r-1)\cdot\frac{t(t-n)}{n^2(n-1)} = \frac{rt(n-t)}{n^2}\cdot\left[1-\frac{r-1}{n-1}\right]\\ &= \frac{rt(n-t)}{n^2(n-1)}\cdot[n-r],\end{aligned}$$

as claimed. $\diamond$

Proposition 2 (Asymptotic distribution) *The probability of a majority of black balls is*

$$p_r^{(s)} \approx \frac{1}{\sqrt{2\pi}}\cdot\int_{-\infty}^{\sqrt{r\lambda}} e^{-t^2/2}dt$$

with $\lambda = (2p-1)^2$*, under the assumption that* $p \approx \frac{1}{2}$*,* $r \ll n$*, and* r *not too small.*

[By FISHER's rule of thumb $10 \le r \le n-10$ suffices if $p \approx \frac{1}{2}$.
Note that this "proposition" lacks mathematical precision.]
Proof. We look at the upper boundary of the integral (1) for $x = \frac{r}{2}$:

$$\begin{aligned}\frac{x-\mu}{\sigma} &= \frac{(\frac{r}{2}-\frac{rt}{n})\cdot n\cdot\sqrt{n-1}}{\sqrt{r(n-r)t(n-t)}} = \frac{(rn-2rt)\sqrt{n-1}}{2\cdot\sqrt{r(n-r)t(n-t)}}\\ &= \frac{\sqrt{r}\sqrt{n-1}}{\sqrt{n-r}}\cdot\frac{s-t}{2\sqrt{st}} = \frac{\sqrt{n-1}}{\sqrt{n-r}}\cdot\sqrt{r}\cdot\frac{2p-1}{2\sqrt{p(1-p)}}\\ &\approx 1\cdot\sqrt{r}\cdot\frac{2p-1}{2\cdot\sqrt{\frac{1}{4}}} = \sqrt{r\lambda},\end{aligned}$$

as claimed. $\diamond$