Random Variables Lecture 7

Connections between Binomial and Hypergeometric

Ex. (Fisher exact test) Researcher wishes to investigate whether a certain disease is more prevalent among men than among women. A random sample of n women and m men is gathered and each person is tested for the disease.

Let X = # of women who have disease and

Y = # of men who have disease.

Then X is a binomial random variable with parameters (n, p_1) and Y is a binomial random variable with parameters (m, p_2) where p_1 and p_2 are unknown. The null hypothesis is that $p_1 = p_2 = p$. Is the null hypothesis likely? Suppose these were the results

	Women	Men	Total
Disease	k	$r-k$	r
No disease	$n-k$	$m-r+k$	$n+m-r$
Total	n	m	$n+m$

Notice that $P(X=k \mid X+Y=r) = \dfrac{P(X=k, Y=r-k)}{P(X+Y=r)}$

where $X+Y$ is binomial with parameters $(n+m, p)$.

Thus $P(X=k \mid X+Y=r) =$

$$= \frac{\binom{n}{k} p^k (1-p)^{n-k} \binom{m}{r-k} p^{r-k} (1-p)^{m-r+k}}{\binom{n+m}{r} p^r (1-p)^{n+m-r}}$$

$$= \frac{\binom{n}{k}\binom{m}{r-k}}{\binom{n+m}{r}}.$$

This is a hypergeometric probability that completely disposed of the unknown probability p.

The reason for the hypergeometric probability can be seen in the fact that $X+Y=r$ means mathematically the same as "r balls out of a total $m+n$ balls were sampled." The white balls then are women.

$P(X=k \mid X+Y=r)$ is then the probability of k white balls in a random sample of r balls.

Lets try to simulate the test with a concrete example.

	Women	Men	Total
Diseased	9	1	10
Not Diseased	3	11	14
Total	12	12	24.

If women and men are equally likely to suffer from the disease, what is the probability of observing 9 sick women out of a total of 10 sick individuals in a sample of 24?

$$P(X=9 \mid X+Y=10) = \frac{\binom{12}{9}\binom{12}{1}}{\binom{24}{10}} \approx 0.001$$

If we consider this too low, we reject the null hypothesis to conclude that women and men are not equally likely to be afflicted.

Thm: If X is binomial r.v. with parameters (n,p) and Y is binomial with parameters (m,p) then the conditional distribution of X given that $X+Y=r$ is hypergeometric with parameters (n,m,r).

On the other hand

Thm: If X is hypergeometric with parameters (w,b,n) and $N=w+b \longrightarrow \infty$ such that $p=\frac{w}{w+b}$ remains fixed then the probability mass function of X converges to that of binomial distribution with parameters (n,p).

Proof: $$P(X=k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} \quad \underline{\text{Recall}} \quad \frac{\binom{n}{k}\binom{w+b-n}{w-k}}{\binom{w+b}{w}}$$

$$= \binom{n}{k} \frac{\omega!\, b!}{(\omega+b)!} \cdot \frac{(\omega+b-n)!}{(\omega-k)!\,(b-n+k)!}$$

$$= \binom{n}{k} \frac{\omega!}{(\omega-k)!} \cdot \frac{b!}{(b-n+k)!} \cdot \frac{(\omega+b-n)!}{(\omega+b)!}$$

$$= \binom{n}{k} \frac{\omega(\omega-1)\dots(\omega-k+1)\; b(b-1)\dots(b-n+k+1)}{(\omega+b)(\omega+b-1)\dots(\omega+b-n+1)}$$

$$= \binom{n}{k} \frac{p\left(p-\frac{1}{N}\right)\dots\left(p-\frac{k-1}{N}\right)\left(1-p\right)\left(1-p-\frac{1}{N}\right)\dots\left(1-p-\frac{n-k-1}{N}\right)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\dots\left(1-\frac{n-1}{N}\right)}$$

as $N \to \infty$ we get $P(X=k) \longrightarrow \binom{n}{k} p^k (1-p)^{n-k}$

That this should happen is fairly clear, because in the limit, the hypergeometric sampling without replacement becomes indistinguishable from sampling with replacement which characterizes binomial distributions.

Geometric and Negative Binomial Random Variables

Geometric distribution A sequence of independent Bernoulli trials is run until a first success occurs. Each trial has probability of success $p \in (0,1)$. Let X = # of failures before first success. Then X has the geometric distribution of parameter p.

$$P(X=k) = (1-p)^k p.$$

Clearly $\sum_{k=0}^{\infty} P(X=k) = \sum_{k=0}^{\infty} (1-p)^k p =$

$= \frac{1}{1-(1-p)} \cdot p = 1.$

Thm: If X is geometric r.v. with parameter p, the cumulative distribution of X is given by

$$F(x) = \begin{cases} 1-(1-p)^{\lfloor x \rfloor + 1} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

Proof: $F(x) = P(X \leq x) = P(X \leq \lfloor x \rfloor) =$

$= 1 - P(X > \lfloor x \rfloor) = 1 - P(X \geq \lfloor x \rfloor + 1) =$

$= 1 - P(\lfloor x \rfloor + 1 \text{ failures at least}) = 1 - (1-p)^{\lfloor x \rfloor + 1}$

Expected Value and Variance of geometric r.v.

Observe that $E = \sum_{k=0}^{\infty} k x^k = \sum_{k=1}^{\infty} (k-1) x^k + \sum_{k=1}^{\infty} x^k =$

$= x \sum_{k=1}^{\infty} (k-1) x^{k-1} + \frac{x}{1-x} = xE + \frac{x}{1-x}$

In particular, $E = \frac{x}{(1-x)^2}$

Thus, $E[X] = \sum_{k=0}^{\infty} k(1-p)^k p = \frac{1-p}{p^2} \cdot p = \frac{1-p}{p}.$

To compute the variance we need

$$E[X^2] = \sum_{k=0}^{\infty} k^2 (1-p)^k p = p \sum_{k=1}^{\infty} (k-1+1)^2 (1-p)^k$$

$$= p \left(\sum_{k=1}^{\infty} (k-1)^2 (1-p)^k + 2 \sum_{k=1}^{\infty} (k-1)(1-p)^k + \sum_{k=1}^{\infty} (1-p)^k \right)$$

$$= (1-p) E[X^2] + 2(1-p) E[X] + 1-p$$

Thus $E[X^2] = 2\left(\frac{1-p}{p}\right)^2 + \left(\frac{1-p}{p}\right)$

and $Var(X) = E[X^2] - (E[X])^2 = \left(\frac{1-p}{p}\right)^2 + \left(\frac{1-p}{p}\right)$

$$= \frac{1-p}{p}\left(\frac{1-p}{p} + 1\right) = \frac{1-p}{p} \cdot \frac{1}{p} = \frac{1-p}{p^2}.$$

Negative Binomial distribution

In a sequence of independent Bernoulli trials with success probability p, let X = # of failures before the r^{th} success.

$$P(X=n) = \binom{n+r-1}{r-1} p^r (1-p)^n$$

Because if n failures and r successes occur (with last trial being the r^{th} success, we have $\binom{n+r-1}{r-1}$ ways of assigning the remaining successes.

We say that X has a negative binomial distribution with parameters (r, p)

To compute the expected value and variance of a negative binomial random variable with parameters (r,p), observe that we can express this variable as

$X = X_1 + \ldots + X_r$ where X_k = # number of failures between the $k-1^{th}$ and k^{th} success. Clearly the X_k are independent, identically distributed geometric random variables with parameter p.

Thus $E[X] = E[X_1 + \ldots + X_r] = E[X_1] + \ldots + E[X_r]$

$= rE[X_1] = r \cdot \frac{1-p}{p}$

By independence

$Var(X) = Var(X_1 + \ldots + X_r) = Var(X_1) + \ldots + Var(X_r) = r\,Var(X_1)$

$= r \cdot \frac{1-p}{p^2}$.

<u>Ex.</u> There are n types of toys in cereal boxes. When you buy the box, any of the n types is equally likely to be in this box. How many boxes do you need to buy on average to get the full set of n toys?

<u>Solution:</u> Let X = # of cereal boxes bought to gather all toys. Then $X = X_1 + \ldots + X_n$ where X_k = # of boxes bought after $k-1^{th}$ new toy and until k^{th} new toy.

Notice that $P(X_k = j) = \left(\frac{k-1}{n}\right)^{j-1} \cdot \frac{n-k+1}{n}$ for $j = 1, \ldots,$

$$E[X_k] = \sum_{j=1}^{\infty} j(1-p)^{j-1}p = \sum_{j=0}^{\infty} (j+1)(1-p)^j p$$

$$= \frac{1-p}{p} + 1 = \frac{1}{p} \quad \text{where } p = \frac{n-k+1}{n}. \quad \text{Thus}$$

$$E[X] = \sum_{k=1}^{n} \frac{n}{n-k+1} = n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{1}\right).$$

$$= n \sum_{i=1}^{n} \frac{1}{i}$$

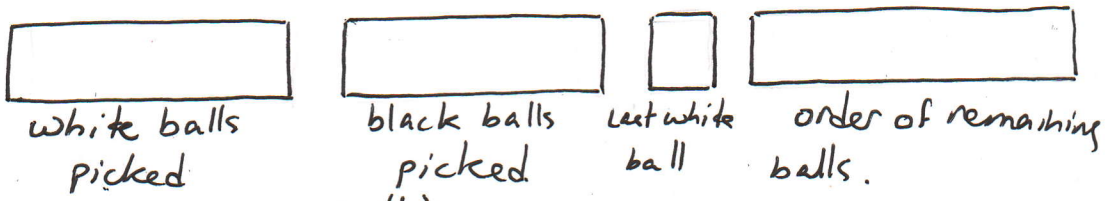
For instance if there are 5 toy types, the average number of cereal boxes is $5\left(1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\frac{1}{5}\right) = 11.41\overline{6}$.

Negative Hypergeometric random variables

Urn contains w white and b black balls. The balls are randomly drawn one by one without replacement until r white balls have been obtained. Let X = # of black balls drawn before r^{th} white ball. Then X is said to have a negative hypergeometric distribution with parameters (w, b, r).

We can compute $P(X=k)$ directly by assuming the balls numbered $1, \ldots, w$ are white and $w+1, \ldots, w+b$ are black (fixed identities)

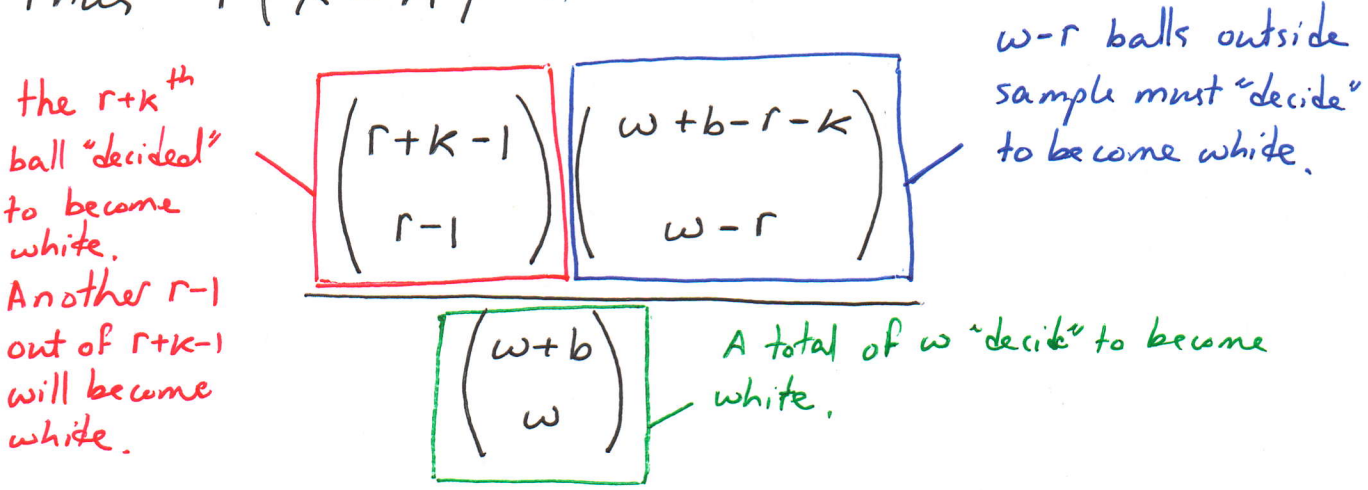
If $X = k$, we construct the following code:



$$\text{Thus } P(X=k) = \frac{\binom{w}{r}\binom{b}{k} r\,(r+k-1)!}{\binom{w+b}{r+k}(r+k)!} = \frac{\binom{w}{r}\binom{b}{k} r}{\binom{w+b}{r+k}(r+k)}$$

We can obtain a somewhat simpler formula by assuming that the sampled balls assume (randomly) their identity only after they were sampled. Thus $X = k$ means that $r+k$ balls were sampled and r decided to be white.

Thus $P(X=k) =$



These ideas are very satisfying! They remind me of many things. For instance, Tsvetayeva writes about the death of touched hope:

Когда пленясь прозрачностью медузы
Ее коснемся мы капризом рук,
Она, как пленник, заключенный в узы,
Вдруг побледнеет и погибнет вдруг.

When captivated by the transparency
of the jellyfish
we touch it with the whim of our hands,
She, like a prisoner, encased in bonds,
Suddenly turns pale and dies suddenly.

You can, of course, day dream about gender fluidity or other social justice ideas that, in the weak form I know them, I find them amusingly useful to my combinatorial thinking.

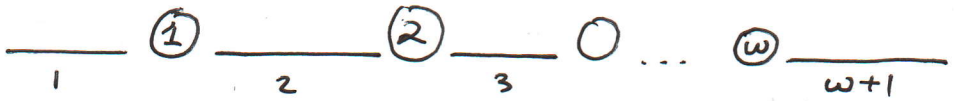
To compute the expected value of X observe that

$$X = X_1 + X_2 + \ldots + X_r$$

where X_k = # of black balls between $k-1^{th}$ and k^{th} white ball in the sample. If we imagine that the balls are arranged in a column (like inside a pez dispenser) The black balls are distributed as follows:

For each black ball (imagine them numbered $1, \ldots, b$)

This black ball is equally likely to be in any place



among the $\omega+1$ places between the white balls.

Then each $X_k = Y_{k1} + \ldots + Y_{kb}$ where

$$Y_{kj} = \begin{cases} 1 \text{ if } j^{th} \text{ black ball is in } k^{th} \text{ space} \\ \text{(between } k-1^{th} \text{ and } k^{th} \text{ white ball)} \\ 0 \text{ otherwise.} \end{cases}$$

Clearly $E[Y_{kj}] = P(Y_{kj} = 1) = \frac{1}{\omega+1}$.

Hence $E[X_k] = bE[Y_{k1}] = \frac{b}{\omega+1}$ and

$E[X] = rE[X_1] = \frac{rb}{\omega+1}$

The computation of variance is somewhat more tedious:

$$Var(X) = Var(X_1 + \ldots + X_r) = \sum_{k=1}^{r} Var(X_k) + \sum_{k=1}^{r} \sum_{\substack{j=1 \\ j \neq k}}^{r} Cov(X_k, X_j)$$

$$= r\,Var(X_1) + r(r-1)\,Cov(X_1, X_2)$$

Now $X_1 = Y_{11} + \ldots + Y_{1b}$ and these variables are independent. Hence $Var(X_1) = b\,Var(Y_{11}) =$

$$= b\left(E[Y_{11}^2] - (E[Y_{11}])^2\right) = b\left(\frac{1}{\omega+1} - \left(\frac{1}{\omega+1}\right)^2\right)$$

$$= \frac{\omega b}{(\omega+1)^2}$$

And $Cov(X_1, X_2) = E[X_1 X_2] - E[X_1]E[X_2]$

$= E[X_1 X_2] - (E[X_1])^2$.

$E[X_1 X_2] = E[(Y_{11} + \ldots + Y_{1b})(Y_{21} + \ldots + Y_{2b})]$

$= bE[Y_{11} Y_{21}] + b(b-1) E[Y_{11} Y_{22}]$

Clearly $Y_{11} Y_{21} = 0$ because black ball #1 cannot simultaneously be before the first white ball and after second white ball.

Thus $E[Y_{11} Y_{21}] = 0$

$E[Y_{11} Y_{22}] = P(Y_{11} Y_{22} = 1) = \frac{1}{(w+1)^2}$

Thus $E[X_1 X_2] = \frac{b(b-1)}{(w+1)^2}$

and $Cov(X_1, X_2) = \frac{b(b-1)}{(w+1)^2} - \frac{b^2}{(w+1)^2} = \frac{-b}{(w+1)^2}$

Thus $Var(X) = r\, Var(X_1) + r(r-1)\, Cov(X_1, X_2)$

$= r \cdot \frac{wb}{(w+1)^2} - r(r-1)\frac{b}{(w+1)^2} = \frac{br}{(w+1)^2}(w - r + 1)$

$= \frac{(w-r+1)br}{(w+1)^2}$

Check my computations!!!