COMMON DISCRETE DISTRIBUTIONS

Binomial (p, n)

Experiment: n independent trials of 'roll of a die' with two possibilities, Success and Failure (probability p and $q = 1 - p$ respectively).

Sample space has 2^n simple events of the $FSS...SF$ type, each having the probability of $p^i q^{n-i}$ where i is the # of Ss.

X is defined as # of Ss.

$$
\begin{aligned}
f(i) &= \binom{n}{i} p^i q^{n-i} \qquad i = 0..n \\
P(z) &= \sum_{i=0}^{n} f(i) z^i = (q + pz)^n \\
&\quad \text{- use } (a+b)^n \text{ formula to prove} \\
\mu &= n(q+pz)^{n-1} p\big|_{z=1} = np \\
\mathrm{Var}(X) &= n(n-1)(q+pz)^{n-2} p^2\big|_{z=1} + np - n^2 p^2 = npq
\end{aligned}
$$

Special case when $n = 1$ is called 'Bernoulli distribution'

Geometric p

Same type of experiment, done till getting the first S, X defined as the number of *trials*.

$$f(i) = p \cdot q^{i-1} \qquad i = 1, 2, 3, ...$$

$$
\begin{aligned}
P(z) &= pz \sum_{i=1}^{\infty} (qz)^{i-1} = \frac{pz}{1 - qz} \\
\mu &= P'(z)|_{z=1} = \frac{1}{p} \\
\mathrm{Var}(X) &= P''(z)|_{z=1} + \mu - \mu^2 = \frac{q}{p^2}
\end{aligned}
$$

$X - 1$ (counting only the failures) has the so-called 'modified' geometric distribution.

Negative Binomial (p, k) counts the trials till the k^{th} Success - obviously a sum of k independent RVs of the previous type.

$$
\begin{aligned}
P(z) &= \left(\frac{pz}{1 - qz} \right)^k \\
\mu &= \frac{k}{p} \\
\mathrm{Var}(X) &= \frac{kq}{p^2} \\
f(i) &= p \cdot \binom{i-1}{k-1} p^{k-1} q^{i-k} \qquad i = k, k+1, k+2, ...
\end{aligned}
$$

Can be easily adjusted for the 'modified' case (of counting failures only).

Hypergeometric (N, K, n)

Experiment: From N physical objects (cards, marbles, etc.), K of which are 'special' in some sense (spades, aces, red marbles, etc.), select randomly and *without replacement* n; X is the # of special objects in your sample.

Sample space consists of $\binom{N}{n}$ 'orderless' selections, of which $\binom{K}{i}\cdot\binom{N-K}{n-i}$ contain exactly i special objects (use *multiplication principle*).This implies

$$f(i) = \frac{\binom{K}{i}\cdot\binom{N-K}{n-i}}{\binom{N}{n}} \qquad i = 0, 1...n$$

The range can actually be narrower (eg. when $K < n$), but luckily the binomial coefficients take care of it (becoming 0 in any such case)!

The formula for $P(z)$ is tricky (we won't use it) - we can still easily build $P(z)$ numerically. That means it is now more difficult to find the mean and variance. This is how we can do it: Put each of the selected marbles under a cup *before* observing its colour; our X then 'splits' into

$$X = X_1 + X_2 + ... + X_n$$

where all the X_i have Bernoullli-type distribution, but they are NOT independent! This yields

$$\mathbb{E}(X) = \sum_{i=1}^{n} \mathbb{E}(X_i) \overset{\text{sym}}{=} n \cdot \mathbb{E}(X) = n \cdot \frac{K}{N}$$

and

$$\mathrm{Var}(X) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} \mathrm{Cov}(X_i, X_j) =$$

$$n \cdot \mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1, X_2) =$$
$$n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

Note the similarity with the binomial *npq* formula, except for the last 'correction' factor, which makes the 2 formulas identical when $n = 1$ (check) and makes the last formula equal to 0 when $n = N$ (check). Also note that X would have Binomial distribution with $p = \frac{K}{N}$ if this sampling were done WITH replacement!

Poisson Λ

Experiment: customers are arriving at a store (library, gas station, etc.) randomly and independently of each other, at an *average* rate of λ per hour. X is the # of customers arriving during a specific time interval of length T.

As an approximation, we can subdivide the time interval into n equal-length subintervals and assume that during each of these a customer arrives with a (tiny) probability of $p_n = \frac{\lambda T}{n}$ (note that this makes the corresponding expected value equal to $\Lambda \overset{\text{def}}{=} \lambda \cdot T$). This implies that

$$P_n(z) = \left(1 - \frac{\Lambda}{n} + \frac{\Lambda}{n} z\right)^n \underset{n\to\infty}{\longrightarrow} \exp\left(\Lambda(z-1)\right)$$

(this 'model' becomes perfect only in the $n \to \infty$ limit). From $P(z)$ we can get everything else:

$$\begin{aligned} \mu &= \Lambda \\ \text{Var}(X) &= \Lambda^2 + \Lambda - \Lambda^2 = \Lambda \end{aligned}$$

and, from

$$e^{-\Lambda} \cdot e^{\Lambda z} = \left(1 + \Lambda z + \frac{\Lambda^2 z^2}{2!} + \frac{\Lambda^3 z^3}{3!} + \frac{\Lambda^4 z^4}{4!} + ...\right) \cdot e^{-\Lambda}$$

we get

$$f(i) = \frac{\Lambda^i}{i!} \cdot e^{-\Lambda} \qquad\qquad i = 0, 1, 2, ...$$

Note that the *sum* of 2 (or more) independent Poisson RVs is also Poisson (with $\Lambda = \Lambda_1 + \Lambda_2$) - clear from PGF.

Binomial and Hypergeometric extended to MULTIVARIATE (we do trivariate only)

Binomial becomes **Multinomial** by assuming that in each trial there are 3 possibilities (winning, losing and tying a game) with probabilities p_1, p_2 and p_3 respectively (they have to add up to 1).

It's easy to see how to extend the samples space (to consist of 3^n simple events), implying that

$$\begin{aligned} \Pr(X &= i \cap Y = j \cap Z = k) = \binom{n}{i,j,k} p_1^i p_2^j p_3^k \\ & \quad\text{whenever } i,j,k \geq 0 \quad \text{ and } \quad i+j+k=n \\ \text{ie. } i &= 0..n, \quad j = 0..n-i \quad \text{ and } \quad k = n-i-j \end{aligned}$$

where X represends the # of wins, etc. The marginal distribution of X is clearly $\mathcal{B}(p_1, n)$, etc., the only new formula we need is

$$\text{Cov}(X, Y) = -np_1p_2$$

Proof:

$$\begin{aligned} &\text{Cov}(X_1 + X_2 + ... + X_n, Y_1 + Y_2 + ... + Y_n) = \\ &\sum_{i=1}^{n} \text{Cov}(X_i, Y_i) = n \cdot \text{Cov}(X_1, Y_1) \end{aligned}$$

(finish in class).

Multivariate Hypergeometric

Now we assume that there is K_1 red, K_2 blue and K_3 green marbles (in a box of $N = K_1 + K_2 + K_3$). By a similar extension of the sample space we get

$$f_{x,y,z}(i,j,k) = \frac{\binom{K_1}{i}\binom{K_2}{j}\binom{K_3}{k}}{\binom{N}{n}}$$

for any possible combination of i, j and k

Again, all the marginals are clearly of the univariate hypergeometric type, the only extra formula (badly) needed is

$$\text{Cov}\,(X,Y) = -n \cdot \frac{K_1}{N} \cdot \frac{K_2}{N} \cdot \frac{N-n}{N-1}$$

(again, note the parallel with the multinomial formula, except for the extra correction term). This time

$$\begin{aligned} \text{Cov}(X_1 + X_2 + ... + X_n, Y_1 + Y_2 + ... + Y_n) &= \\ n \cdot \text{Cov}(X_1, Y_1) + n(n-1)\text{Cov}(X_1, Y_2) &= .... \end{aligned}$$