Lecture 5: Poisson, Hypergeometric, and Geometric Distributions

Sta230/Mth230

Colin Rundel

January 31, 2012

Binomial Approximations

Last week we looked at the normal approximation for the binomial distribution:

- Works well when n is large
- Continuity correction helps
- Binomial can be skewed but Normal is symmetric (book discusses an additional correction for this)
- At a minimum we want $np \geq 10$ and $nq \geq 10$

What do we do when p is close to 0 or 1?

Alternative Approximation

Let $X \sim \text{Binom}(n, p)$ which we will reparameterize so that $p = \lambda/n$ for a fixed value of λ. As such, λ/n is small when n is large.

We will evaluate the Binomial distribution as $n \to \infty$.

Alternative Approximation, Cont.

$$A_n = \frac{n!}{n^k(n-k)!}$$

Alternative Approximation, Cont.

$$B_n = \left(1 - \frac{\lambda}{n}\right)^n$$

Alternative Approximation, Cont.

$$C_n = \left(1 - \frac{\lambda}{n}\right)^{-k}$$

Alternative Approximation, cont.

Let $X \sim \text{Binom}(n, p)$ we will reparameterize such that $p = \lambda/n$ for a fixed value of λ. As such, λ/n is small when n is large.

$$P(X = k|n, p = \lambda/n) = \underbrace{\left(\frac{n!}{n^k(n-k)!}\right)}_{A_n}\left(\frac{\lambda^k}{k!}\right)\underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{B_n}\underbrace{\left(1 - \frac{\lambda}{n}\right)^{-k}}_{C_n}$$

$$\lim_{n\to\infty} P(X = k|n, p = \lambda/n) = \frac{\lambda^k}{k!}\, e^{-\lambda}$$

Therefore for large n,

$$P(X = k|n, p = \lambda/n) \approx \frac{\lambda^k}{k!}\, e^{-\lambda}$$

Approximation - Mean & Variance

We defined $p = \lambda/n$ and we know that for a Binomial random variable

$$\mu = np$$
$$\sigma^2 = npq$$

Then for large n we then get,

$$\lim_{n\to\infty} \mu = \lim_{n\to\infty} n\,\frac{\lambda}{n} = \lambda$$
$$\lim_{n\to\infty} \sigma^2 = \lim_{n\to\infty} n\,\frac{\lambda}{n}\left(1 - \frac{\lambda}{n}\right) = \lambda \lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right) = \lambda$$

Poisson Distribution

Let X be a random variable reflecting the number of events in a given period where the expected number of events in that interval is λ then the probability of k occurrences ($k \geq 0$) in the interval is given by the Poisson distribution, $X \sim \text{Pois}(\lambda)$

$$P(X = k|\lambda) = f(k|\lambda) = \frac{\lambda^k}{k!}e^{-\lambda}$$

$$E[X] = \mu = \lambda$$
$$\text{Var}[X] = \sigma^2 = \lambda$$

We use this approximation to the Binomial when p is very small and n is very large since $\lambda = np$ tends to be reasonable.

Poisson Distribution - Mode

We can use the same approach that we used with the Binomial distribution Therefore k_{mode} is the smallest integer greater than $\lambda - 1$

$$k_{\text{mode}} = \begin{cases} \lambda - 1, \lambda & \text{if } \lambda = \lceil \lambda \rceil \\ \lceil \lambda \rceil - 1 & \text{otherwise} \end{cases}$$

Poisson Distribution - Example

Assume you have a sample of a stable isotope of an element, there are approximately 10^{20} atoms in this sample. If on average one of these atoms will radioactively decay every 10^{12} years ($\approx 5 \times 10^{19}$ secs).

What is the probability that 4 or fewer atoms decay in the next second?

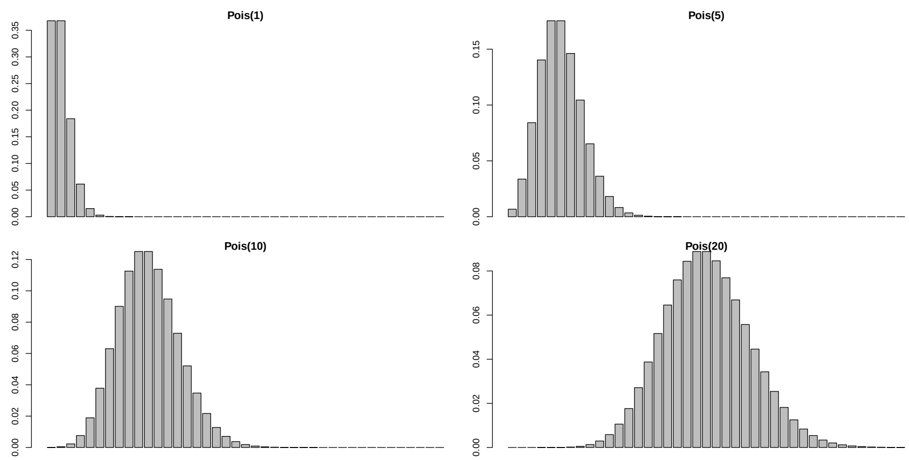
Poisson and Normal Distributions

Based on the connection between the Binomial and Poisson distributions it intuitively makes sense that we should also be able to approximate the Poisson with a Normal distribution.

For approximation to the binomial we need $np \geq 10$ and $nq \geq 10$.

What is a reasonable requirement for λ?

Poisson and Normal Distributions, cont.



Another way to look at the Binomial

Imagine we have a population is partitioned into good and bad subsets. Let G be the number of good elements in the population, B the number of bad elements, and $N = B + G$.

If we sample this population *with replacement* what is the probability that we observe g good samples and b bad samples.

This is the Binomial distribution rewritten such that

$$P(g \text{ good, } b \text{ bad in } n = b + g \text{ tries}) = \binom{n}{g} \frac{G^g\ B^b}{N^n}$$
$$= \binom{n}{g} \left(\frac{G}{N}\right)^g \left(\frac{B}{N}\right)^b = \binom{n}{g} \left(\frac{G}{N}\right)^g \left(1 - \frac{G}{N}\right)^{n-g}$$

Hypergeometric

What would change if we were *sampling without replacement*?

Hypergeometric Distribution

Let X be a random variable reflecting the number of successes in n draws without replacement from a finite population of size N with m desired items then the probability of k successes is given by the Hypergeometric distribution, $X \sim \text{Hypergeo}(N, m, n)$

$$P(X = k) = f(k|N, m, n) = \frac{\binom{m}{k}\binom{N-m}{n-k}}{\binom{N}{n}}$$

Hypergeometric Distribution - Example

You are dealt five cards, what is the probability that four of them are aces?

If we use the Hypergeometric distribution then, $N = 52$, $m = 4$, $n = 5$ and

Hypergeometric Distribution - Another Way

Let $X \sim \text{Binom}(m, p)$ and $Y \sim \text{Binom}(N - m, p)$ be independent Binomial random variables then we can define the Hypergeometric distribution as the conditional probability of $X = k$ given $X + Y = n$.

Note that $X + Y \sim \text{Binom}(N, p)$

Geometric Distribution

Let Y be a random variable reflecting the number failures of independent Bernoulli trials, with probability of success p, needed before observing the first success. Then the probability of k failures before the first success is given by the Geometric distribution, $Y \sim \text{Geo}(p)$

$$P(Y = k) = f(k|p) = p(1-p)^k$$

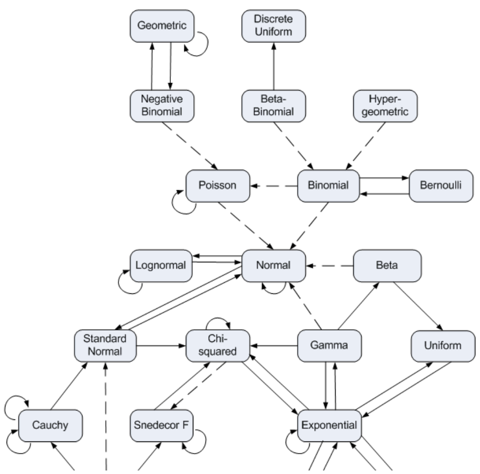
Negative Binomial Distribution

Let X be a random variable reflecting the total number of successes before the r^{th} failure where each trial is an independent Bernoulli trial with p probability of success. Then the probability of k successes is given by the Negative Binomial distribution, $X \sim \text{NB}(r, p)$

$$P(X = k) = f(k|r, p) = \binom{k + r - 1}{k} p^k (1-p)^r$$

Note that we can also think of X as being the sum of r Geometric random variable, $X = Z_1 + Z_2 + \cdots + Z_r$ where $Z_1, \cdots, Z_r \sim Geo(p)$.

Distribution Relationships



http://www.johndcook.com/distribution_chart.html

Expected Value

The expected value of a random variable is defined as follows

Discrete Random Variable:

$$E[X] = \sum_{\text{all } x} xP(X = x)$$

Continous Random Variable:

$$E[X] = \int_{\text{all } x} xP(X = x)\; dx$$

This is a natural generalization of what we do when deciding if a casino game is fair.

Properties of Expected Value

- **Constants** - $E(c) = c$ if c is constant
- **Indicators** - $E(I_A) = P(A)$ where I_A is an indicator function
- **Functions** - $E[g(X)] = \sum_{\text{all } x} g(x)\; P(X = x)$
- **Constant Factors** - $E(cX) = cE(x)$
- **Addition** - $E(X + Y) = E(X) + E(Y)$
- **Multiplication** - $E(XY) = E(X)E(Y)$ if X and Y are independent.

Expected Value of the Geometric Distribution

Let $X \sim \text{Geo}(p)$ then

Expected Value of the Negative Binomial Distribution

Previously we showed (conceptually) that if $Y \sim \text{NB}(r, p)$ then

$$Y = X_1 + X_2 + \cdots + X_r$$

where $X_1, \cdots, X_r \sim \text{Geo}(p)$ as such

$$\begin{aligned} E[Y] &= E\left[\sum_{i=1}^{r} X_i\right] \\ &= \sum_{i=0}^{r} E[X_i] \\ &= \sum_{i=0}^{r} \frac{1-p}{p} \\ &= \frac{r(1-p)}{p} \end{aligned}$$

Expected Value

Expected Value of the Unit Normal Distribution

Let $X \sim \text{N}(0, 1)$ then