Module 4: Discrete Distributions

For any domain there are infinitely many distributions. The most common and famous distributions get a name; Binomial, Negative Binomial, Geometric, Hypergeometric, Poisson, etc. In this section we focus on Binomial and Poisson distributions.

The **Bernoulli Distribution** is a special member of the distribution family. It is the simplest example of a **Binomial distribution,** with only two domains (aka **dicothomous** distribution). A experiment which only has two domains is called a Bernoulli experiment. Ex. the number of students who get an A on a test, whether a person has a disease or not.

If we have two Bernoulli independent trials with equal probability of a positive result, **we refer to that probability as pi (not 3.14)**

X_1 = { 1 if outcome +, 0 if outcome - }   and X_2 = { 1 if outcome +, 0 if outcome - }

Then, $X = X_1 + X_2$

The variable X above is a random variable with domain of {0, 1, 2} as it is a result of the two trials. The distribution is an example of a Binomial (2, pi) distribution.

More generally, if X_i are n Bernoulli independent trials with probability of a positive result equals pi

$$X = \sum_i X_i$$

The domain of X a Binomial (n, pi) is {0, 1, 2... n}. When n=1 the binomial reduces to Bernoulli

For k in domain {0, 1, 2, ...n}:

$$P\{X = k\} = \binom{n}{k} \pi^k (1-\pi)^{n-k}$$

This is only for = and not eqaullity

- Where $({}^n{}_k)$ = (n!) / (k! * (n-k)!)   Where n! = 1 * 2 * 3 * ... n and 0! = 1
- Mean = μ = E [X] = nπ

- Variance = σ^2 = Var [X] = $n\pi(1 - \pi)$

Note that variance is a function of mean, Mean > Variance and for a fixed n the variance is maximum at pi = .5

We can construct the standard Z score with:

$$Z = \frac{X - E[X]}{\sqrt{Var[X]}} = \frac{X - \mu}{\sigma} = \frac{X - n\pi}{\sqrt{n\pi(1 - \pi)}}$$

We can use the Standard Normal Distribution to approximate a binomial distribution when n is large (say > 25), this is an example of the Central Limit Theorem. The **Central Limit Theorem** states if you take the sum of a large number of independent, identically distributed variables you can approximate the outcome under a normal distribution. This is the basis of inference in current applied statistics.

Poisson Distribution

Named after the French mathematician who derived it; the first application was the description of the number of deaths as a result of horse kicking in the Prussian army. It can be used to model the number of events occurring within a given time interval. The probability density (mass) function is:

$$P\{X = k\} = \frac{\lambda^k \exp\{-\lambda\}}{k!}$$

where λ is the mean of the distribution (mean number of events); λ determines the shape of the distribution. Other properties which make Poisson distribution popular:

- The mean and variance are both equal to λ
- The sum of independent Poisson variables is also (!) Poisson variable with mean equal to sum of the individual means
- Poisson distribution provides an approximation for the Binomial distribution
- The standard Z score:

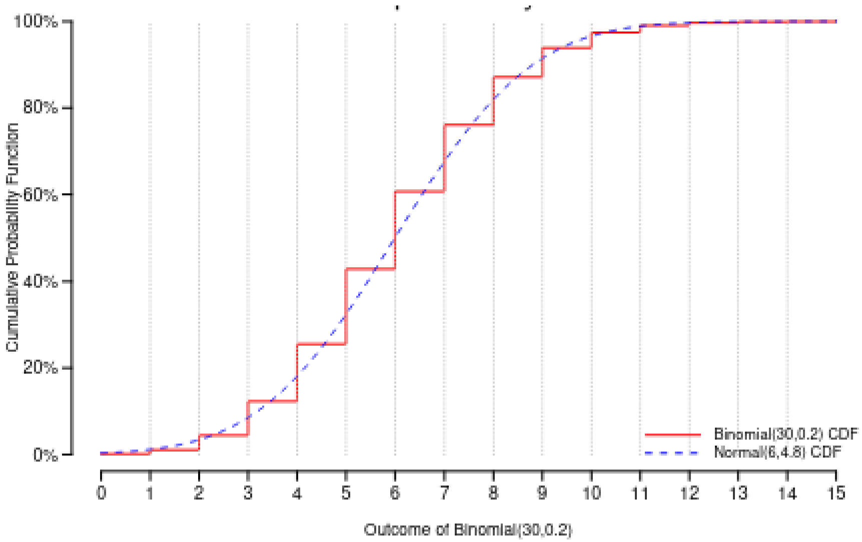
$$Z = \frac{X - E[X]}{\sqrt{Var[X]}} = \frac{X - \mu}{\sigma} = \frac{X - \lambda}{\sqrt{\lambda}}$$

If n is large and pi is small, the the Binomial distribution with parameters n and pi can be approximated by a Poisson distribution with mean parameter n*(pi). From there probability calculations with Poisson reduce to probability calculations for a standard normal distribution.

When converting a discrete binomial distribution to a continuous distribution we must add correction for continuous conversions:

Probability	Corrected
$P(X \leq x)$	$P(X \leq x + 0.5)$
$P(X < x)$	$P(X < x - 0.5)$
$P(X > x)$	$P(X > x + 0.5)$
$P(X \geq x)$	$P(X \geq x - 0.5)$

Think about it like this:



In a binomial distribution probability can only accumulate at discrete times {1, 2, 3,...} but since a normal distribution is continuous, you have to account for weather or not you want to include the point.

Useful R functions:

Function	Normal	Binomial	Poisson
Probability Density/Mass	`dnorm(x, mean, sd)`	`dbinom(x, size, prob)`	`dpois(x, lambda)`
Cumulative Distribution	`pnorm(q, mean, sd)`	`pbinom(q, size, prob)`	`ppois(q, lambda)`
Quantile	`qnorm(p, mean, sd)`	`qbinom(p, size, prob)`	`qpois(p, lambda)`
Random Variable	`rnorm(n, mean, sd)`	`rbinom(n, size, prob)`	`rpois(n, lambda)`

The general rule with functions in R:

- A single point (Probability Density Function) starts with d
- Cumulative Distribution starts with p
- Quantile starts with q
- Random variables start with r

Revision #8
Created 18 August 2022 13:56:29 by Elkip
Updated 19 August 2022 00:51:22 by Elkip