James Green-Armytage
Econ 229, fall 2015

Overview of Probability Distributions

Part A: Discrete distributions

Preliminary definitions

• The probability mass function (PMF) of the distribution, $f(x_i)$, tells us the probability with which x will take on each value x_i. Necessarily, these probabilities sum to one; i.e. $\sum_i[f(x_i)] = 1$.

• The cumulative distribution function (CDF), which we denote as $F(x_i)$, tells us the probability with which x will be less than or equal to each value x_i.

• Let $\mu = E[x]$ be the true mean, or expected value of the distribution. We calculate this as $\boxed{\mu = \sum_i[x_i \cdot f(x_i)]}$.

• Let $\sigma^2 = E[(x-\mu)^2]$ be the true variance of the distribution. We calculate this as $\boxed{\sigma^2 = \sum_i[(x_i-\mu)^2 \cdot f(x_i)]}$. The true standard deviation, σ, is the square root of the true variance.

A.1. Bernoulli

• Description: A Bernoulli random variable is a "success" with probability p, and a "failure" with probability $1-p$.

• PMF: $\boxed{f(1) = p;\ f(0) = 1-p}$.

• CDF: $\boxed{F(0) = 1-p;\ F(1) = 1}$.

• Mean: $\boxed{\mu = p}$.

• Variance: $\boxed{\sigma^2 = p(1-p)}$.

• Note: "Success" can be defined as the occurrence of some event, e.g. flipping a coin and getting heads, rolling a die and getting a 6, etc. A "failure", then, is just when the specified event doesn't occur.

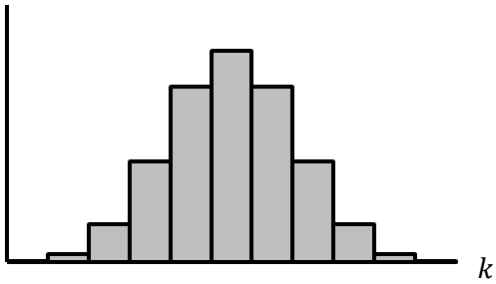
A.2. Binomial

• Description: Given n independent draws from a Bernoulli distribution, the binomial distribution is concerned with the total number of successes.

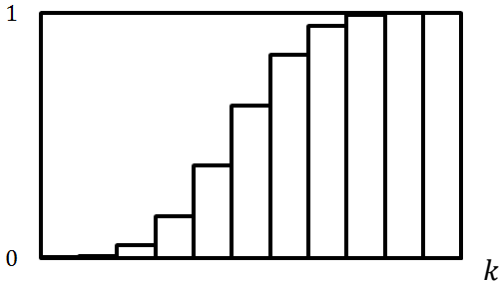
• PMF: $\boxed{f(k) = \binom{n}{k}p^k(1-p)^{n-k}}$. Here, $f(k)$ gives the probability of getting exactly k successes in n independent trials, if each success happens with probability p.

• Mean: $\boxed{\mu = np}$.

• Variance: $\boxed{\sigma^2 = np(1-p)}$.



Binomial PMF with $n = 10, p = 1/2$



Binomial CDF with $n = 10, p = 1/2$

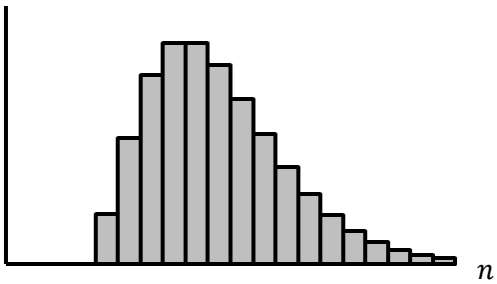
• Excel command: `=BINOMDIST(k, n, p, cumulative)`. Here, `cumulative = false` returns the value of PMF given the indicated parameters, while `cumulative = true` returns the CDF value.

• Note: Unless n is very small, the binomial distribution has approximately the same shape as a normal distribution.

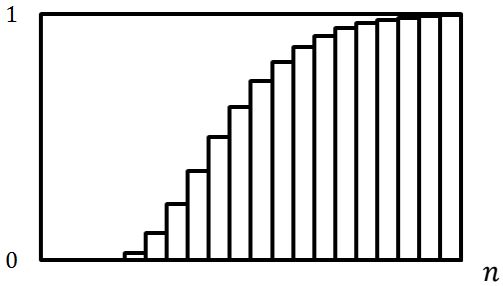
A.3. Negative Binomial

• Description: Given n independent draws from a Bernoulli distribution, the binomial distribution is concerned with when the k^{th} success will be achieved.

• PMF: $\boxed{f(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}}$. Here, $p(n)$ gives the probability of getting the k^{th} success on exactly the n^{th} trial.



Negative binomial PMF with $k = 5, p = 1/2$



Negative binomial CDF with $k = 5, p = 1/2$

• Excel command: `=NEGBINOMDIST(n - k, n, p)`. Note that in Excel, the first argument is the number of *failures* before k successes are achieved; since n is the number of trials and k is the number of successes, the number of failures is $n - k$.

• Note, if we set the parameter $k = 1$, we get the geometric distribution, which is concerned with how many trials will be needed to achieve just one success. Plugging in 1 for k in the PMF of the negative binomial distribution, we can simplify to the PMF of the geometric distribution. This is $f(n) = p(1-p)^{n-1}$, which gives the probability of getting the first success on the n^{th} trial.

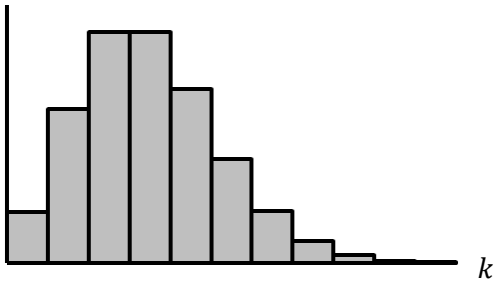
A.4. Poisson

• Description: Given an event that occurs λ times per time unit (e.g. per hour) *on average*, the Poisson distribution is concerned with how many such events will occur in a *particular* time unit. Here we assume that the different occurrences of the event are independent of each other.

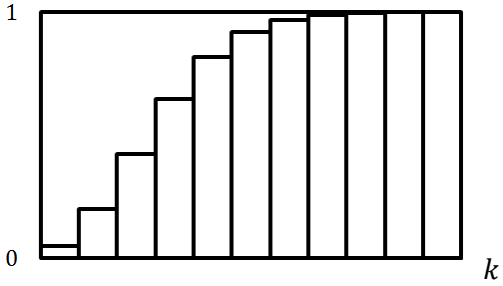
• PMF: $\boxed{f(k) = \frac{\lambda^k}{k!} e^{-\lambda}}$. This gives the probability that the event will occur k times in the specified time interval. Here, e is Euler's number, the mathematical constant equal to $\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n$.

• Mean and variance: $\boxed{\mu = \lambda}$. $\boxed{\sigma^2 = \lambda}$.

• Excel command: `=poisson(k, μ, cumulative)`



Poisson PMF with $\lambda = 3$



Poisson CDF with $\lambda = 3$

A.5. Discrete Uniform

• Description: This is a random variable that has an equal probability of taking on any of the integer values between a and b, inclusive. We assume that a and b are integers, and that $b > a$.

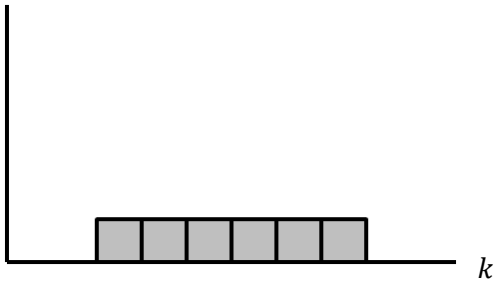
• Define n as the number of possible values that the variable can take on: $n = b - a + 1$.

• PMF: $\boxed{f(k) = 1/n}$. The probability of getting any one of the n values is $1/n$.

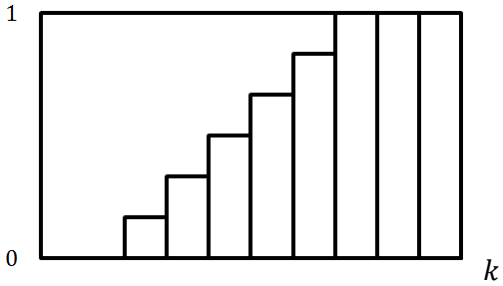
• CDF: $\boxed{F(k) = \frac{k-a+1}{n}}$. This is the probability of getting a value less than or equal to k.

• Mean: $\boxed{\mu = \frac{a+b}{2}}$. The expected value is the mean of the distribution's two endpoints.

• Variance: $\boxed{\sigma^2 = \frac{n^2-1}{12}}$.



Discrete uniform PMF with $a = 3$, $b = 8$



Discrete uniform CDF with $a = 3$, $b = 8$

• Excel simulation: The PMF and CDF of a discrete uniform can be calculated using standard operations. Also, it is possible to randomly generate draws from this distribution using the code

```
=roundup(   rand( ) * n + a - 1   ,   0   )
```

Part B: Continuous Distributions

Preliminary definitions

• Again, the cumulative distribution function (CDF), $F(\xi)$, tells us the probability with which x will be less than or equal to some particular value ξ. That is, $F(\xi) = prob(x \leq \xi)$. But now the probability that x will be exactly equal to any particular value is zero, so we can write this equivalently as $\boxed{F(\xi) = prob(x < \xi)}$.

• The probability density function (PDF) of a continuous distribution is analogous to the PMF of a discrete distribution, but the idea is slightly more abstract. Whereas the PMF of a discrete distribution tells us the probability with which the random variable will take on each particular value, the PDF of a continuous distribution is a function such that the area underneath the curve for some specified range of ξ values gives the probability that x will take on a value in that range. We can write this using calculus as $\boxed{prob(a < \xi < b) = \int_a^b f(\xi)\, d\xi}$.

• Similarly, we can describe the relationship between the PDF and the CDF of a continuous distribution using calculus as follows: $\boxed{f(\xi) = F'(\xi)}$, and $\boxed{F(\xi) = \int_{-\infty}^{\xi} f(x)\, dx}$.

• The formula for the true mean or expected value is analogous to the one for discrete distributions, except that it uses integration rather than summation: $\boxed{\mu = \int_{-\infty}^{\infty} x \cdot f(x)\, dx}$.

• The formula for the true variance is similarly analogous: $\boxed{\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\, dx}$. Again, the true standard deviation, σ, is simply the square root of this.

B.1. Continuous uniform

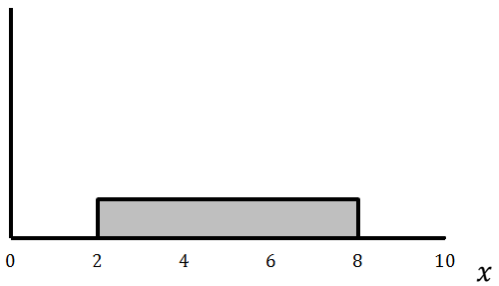
• Description: This is a random variable that has an equal probability of falling anywhere between two real numbers, a and b. Unlike the previous distribution, we no longer assume that the value it takes on is an integer. Define n as the range of the distribution, $n = b - a$.

• PDF: $\boxed{f(x) = 1/n, \text{for } x \in [a, b]}$. $f(x) = 0$ for $x < a$. $f(x) = 0$ for $x > b$.

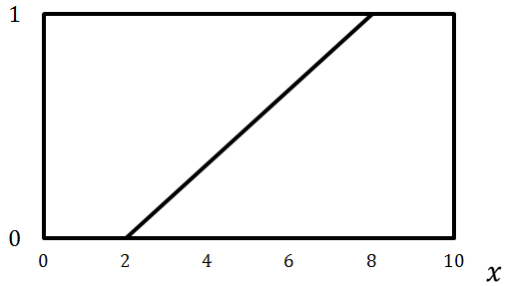
• CDF: $\boxed{F(x) = \frac{x-a}{n}, \text{for } x \in [a, b]}$. $F(x) = 0$ for $x < a$. $F(x) = 1$ for $x > b$.

• Mean and variance: $\boxed{\mu = \frac{a+b}{2}}$, and $\boxed{\sigma^2 = \frac{n^2}{12}}$.

• Excel simulation: The command `rand( )` generates random draws from a continuous uniform distribution with $a = 0$ and $b = 1$. Therefore, you can generate random draws from a uniform distribution with other values of a and b using the code `= rand( ) * n + a`.



Continuous uniform PDF with $a = 2, b = 8$



Continuous uniform CDF with $a = 2, b = 8$

B.2. Normal

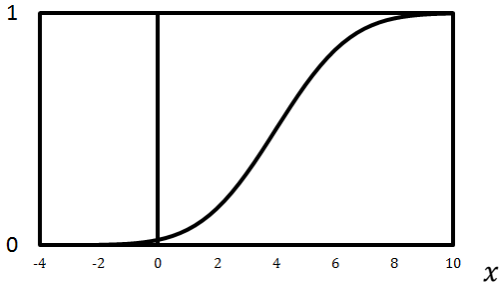
• Description: The normal (or Gaussian) distribution is the most commonly-studied distribution in statistics. It has a symmetrical, bell-shaped PDF. It has two parameters: μ and σ.

• PDF: $\boxed{f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}}$. Here, e and π are the two most well-known transcendental numbers, i.e. Euler's number, and the ratio of a perfect circle's circumference to its diameter. If you plot this function for any given values of μ and σ, you will get a bell-shaped curve.

• CDF: $\boxed{F(x) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right]}$, where $\text{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2}\, dt$.

• Mean and variance: The PDF and CDF have already been defined in terms of these parameters, $\boxed{\mu}$ and $\boxed{\sigma^2}$.

Normal PDF with $\mu = 4$, $\sigma = 2$



Normal CDF with $\mu = 4$, $\sigma = 2$

• Excel has many commands pertaining to normal distributions. For example…

`=normdist(x, μ, σ, cumulative)` returns the value of the PDF with the specified parameters if `cumulative = FALSE`, or the value of the CDF if `cumulative = true`.

`=norminv(F, μ, σ)` returns the value of x such that the CDF value $F(x)$ is equal to the indicated probability, F.

`=normsdist(z)` is a shortcut for the `normdist` function with $\mu = 1$, $\sigma = 1$ (i.e. the standard normal distribution), and `cumulative = true`.

`=normsinv(F)` is a shortcut for the `norminv` function with $\mu = 1$ and $\sigma = 1$.

One can generate random draws from a normal distribution with parameters μ and σ using the code `=norminv( rand( ), μ, σ )`.

• By the central limit theorem, the distribution of a sum or average of n independent draws from a distribution (not necessarily a normal one!) converges toward a normal distribution as n increases. With this in mind, we should have some intuition for why the normal distribution is so carefully studied, and why many real-life data are approximately normally distributed.