produced, the number of customers that a bank teller helps per hour, or the number of telephone calls processed by a switching system per hour. It is denoted as $X \sim Poi(\lambda)$.

$$p(k) = P(X = k) = \frac{\lambda^k \, e^{-\lambda}}{k!}, \; k = 0, 1, 2, \ldots, \; \lambda > 0$$

$$E[X] = \lambda$$

$$Var[X] = \lambda$$

```
>> poi = @(l, k) l.^k .* exp(-l) ./ factorial(k);

>> k = 0:10;
>> lambda = 3;
>> p = poi(lambda, k);
>> stem(k, p)
>> title('Poisson Distribution, \lambda = 3')
>> xlabel('k')
>> ylabel('p(k)')
```

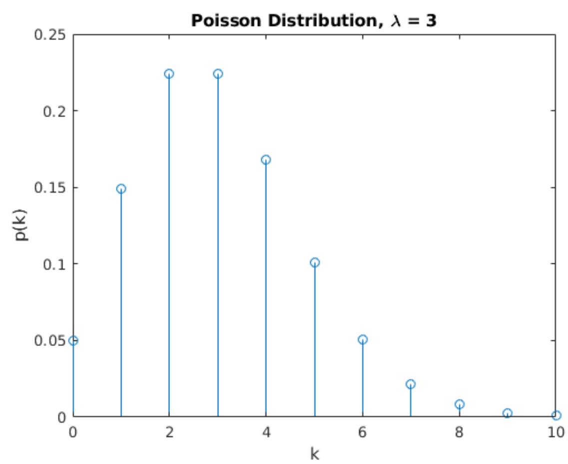
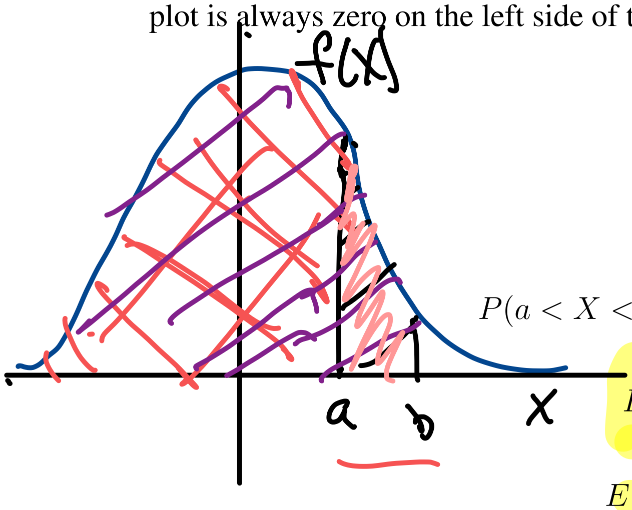


Fig. 4.2: Poisson Distribution

4.4.2 Continuous Distributions

Since a random variable following a continuous distribution can take on any real number, the probability of the variable being a specific value is zero, $P(X = a) = 0$. Instead a probability is specified only for a range of values ($P(a < X < b)$). Therefore, continuous variables do not have probability mass function, instead we use a *probability density function* ($PDF = f(x)$). Probabilities are calculated as the area of an interval under the PDF – that is, an integral. The total area under a PDF must be one. Another useful tool for computing probabilities is a *cumulative* *distribution function* (CDF), which tells us the probability $CDF = P(X < a) = F(a)$. A CDF plot is always zero on the left side of the plot and one to the right.



$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

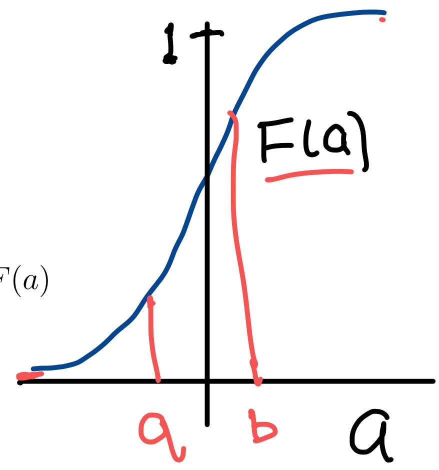
$$F(a) = \int_{-\infty}^{a} f(x)\,dx$$

$$P(a < X < b) = \int_{a}^{b} f(x)\,dx = F(b) - F(a)$$

$$E[X] = \int_{-\infty}^{\infty} x\,f(x)\,dx$$

$$E\left[X^n\right] = \int_{-\infty}^{\infty} x^n\,f(x)\,dx$$

$$Var[X] = E\left[(X - E[X])\right] = E\left[X^2\right] - E[X]^2$$



Uniform random variable

A random variable with a *Uniform distribution* is equally likely to take any value within an interval $[a, b]$.

P(X>4)
= 1 - F(4)

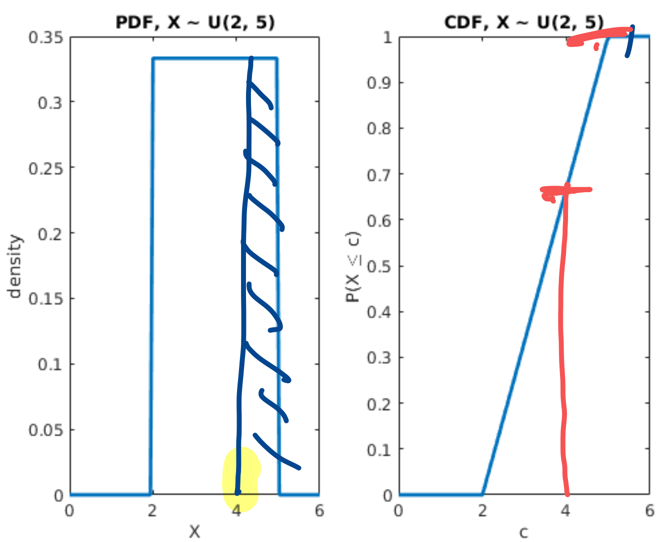


Fig. 4.3: Uniform Distribution

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0, & \text{otherwise} \end{cases}$$

$$F(c) = \begin{cases} 0, & c < a \\ \frac{c-a}{b-a}, & a \le c \le b \\ 1, & c > b \end{cases}$$

$$E[X] = \frac{a+b}{2}$$

$$Var[X] = \frac{(b-a)^2}{12}$$

```
function F = unifcdf(c, varargin)
% UNIFCDF - CDF of continuous uniform distribution U(a, b)
%
%  Inputs: c - variable in question -- may be a scalar or row vector
%          a - minimum value of distribution
%          b - maximum value of distribution
%        unifcdf(c) is the same as unifcdf(c, 0, 1)

    if length(varargin) == 0
        a = 0;
        b = 1;
    else
        a = varargin{1};
        b = varargin{2};
    end
    F = zeros(1, length(c));
    inRange = (c >= a) & (c <= b);
    F(inRange) = (c(inRange) - a)/(b - a);
    F(c > b) = 1;
end
```

Exponential random variable

The *exponential distribution* is related to the Poisson distribution. Exponential random variables are often used to model time measurements, such as the length of time between events or the duration of events.

In the telecommunications industry, the number of voice circuits that a system requires is measured by a metric called Erlangs, where the number of calls per hour is modeled with a Poisson distribution and the length of the calls is modeled with an exponential distribution.

$$f(x) = \begin{cases} \lambda\, e^{-\lambda\, x}, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

$$F(a) = \begin{cases} 1 - e^{-\lambda\, a}, & a > 0 \\ 0, & a \leq 0 \end{cases}$$

$$E[X] = \frac{1}{\lambda}$$

$$Var[X] = \frac{1}{\lambda^2}$$

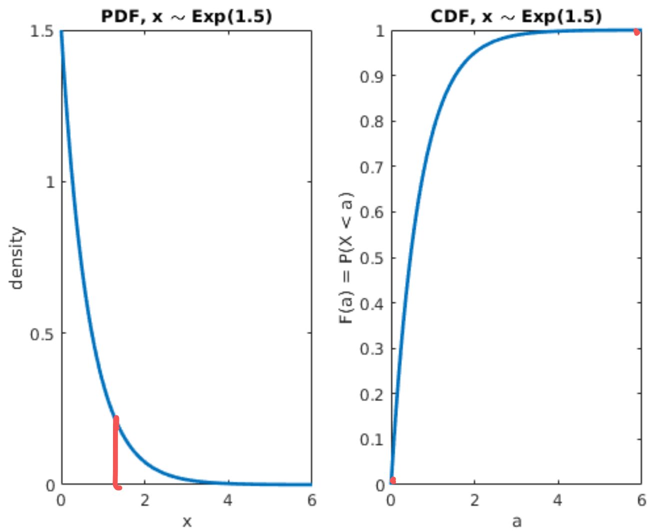


Fig. 4.4: Exponential Distribution

Normal random variable

The *normal distribution*, also called the *Gaussian distribution* models random variables that naturally occur in nature and society. Random variables with a normal distribution include measurements (length, width, height, volume, weight, etc.) of plants and animals, scores on standardized tests, income levels, air pollution levels, etc. Its probability distribution function has the familiar *bell shaped curve* that is shaped by the variable's mean and standard deviation. We denote the distribution as $X \sim N(\mu, \sigma)$.

$$f(x) = \frac{1}{\sigma\sqrt{2\,\pi}} e^{\frac{-(x-\mu)^2}{2\,\sigma^2}}$$

To simplify things, we will map the random variable to a *standard normal distribution* with zero mean and unit variance, $Y \sim N(0, 1)$.

$$y = \frac{x - \mu}{\sigma}$$

$$f(y) = \frac{1}{\sqrt{2\,\pi}} e^{\frac{-y^2}{2}}$$

Computing the CDF requires numerical integration because there is not a closed form integral solution to the PDF. Fortunately, there is a built-in MATLAB function that we can use to compute the integral. See the documentation for functions `erf()` and `erfc()`. The *Statistics and Machine Learning Toolbox* includes a function to compute the CDF, but implementing our own function using `erfc()` is not difficult.

We desire a function called `normcdf(a)` that behaves as follows for distribution $Y \sim N(0, 1)$.

$$\text{normcdf(a)} = P(Y < a) = F(a) = \frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{a} e^{\frac{-y^2}{2}}\,dy$$

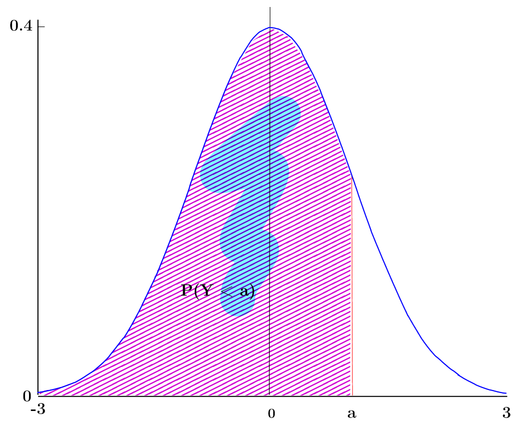


Fig. 4.5: The shaded area is P(Y < a).

The `erfc(b)` function gives us the following definite integral.

$$\text{erfc(b)} = \frac{2}{\sqrt{\pi}} \int_{b}^{\infty} e^{-y^2}\,dy.$$

So, if we multiply by 1/2, change the sign of the definite integral boundary-value to reflect that we are integrating from the boundary to infinity rather than from negative infinity to the boundary, and divide the boundary-value by $\sqrt{2}$ because the squared integration variable in the `erfc` calculation is not divided by 2 as we need, then the integration will compute an equivalent area to what we need.

Note that because of the symmetry of the PDF curve about the mean, $P(Y < a)$ is the same as $P(Y > -a)$.

$$\text{normcdf(a)} = F(a) = \frac{1}{2}\left(\frac{2}{\sqrt{\pi}} \int_{-a/\sqrt{2}}^{\infty} e^{-y^2}\,dy\right) = \frac{1}{2}\text{erfc}(-a/\sqrt{2})$$

You may have noticed that we lost a scalar multiplier of $1/\sqrt{2}$ from the original CDF equation in the CDF equation using the `erfc` function. This is because the integration variable in the `erfc` function is not divided by two. Since we need to use numeric integration instead of analytic integration, we can not see exactly how the $1/\sqrt{2}$ cancelled out of the equation. But I performed the *Numeric Integration* and verified that the $1/\sqrt{2}$ multiplier is not used to get the correct answer.

$$\frac{1}{\sqrt{2\,\pi}} \int_{-\infty}^{\infty} e^{\frac{-y^2}{2}}\,dy = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-y^2}\,dy. = 1$$

```
function F = normcdf(a, varargin)
% NORMCDF - cumulative distribution function for normal␣
 ↪distribution
```

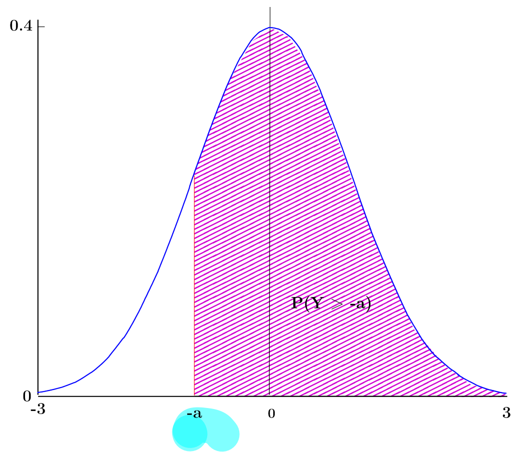


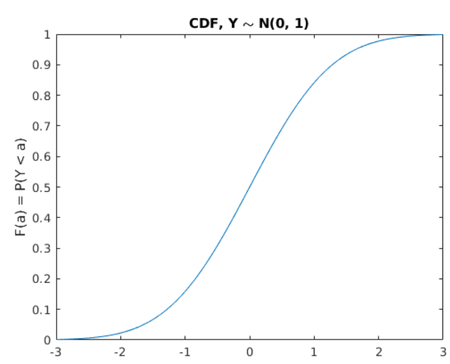
Fig. 4.6: The shaded area is P(Y > -a) = P(Y < a).

```
%   Inputs: a  - variable in question -- may be a scalar or row␣
↪vector
%             mu - mean of distribution
%             sigma - standard deviation
%       normcdf(a) is the same as normcdf(a, 0, 1) - standard normal
%
%   Output: Probability P(X < a)

    if length(varargin) ~= 0
        mu = varargin{1};
        sigma = varargin{2};
        a = (a - mu)./sigma;
    end
    F = erfc(-a/sqrt(2))/2;
end
```

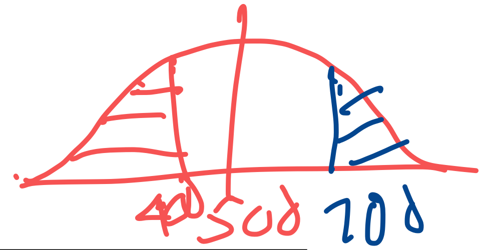
To plot the CDF:

```
>> a = linspace(-3,3);
>> plot(a, normcdf(a))
```



Probability Example:

You learn that the lot of calves to be auctioned at the local livestock sale have a mean weight of 500 pounds with a standard deviation of 150 pounds. What fraction of the calves likely weigh



1.less than 400 pounds?

2.more than 700 pounds?

3.between 450 and 550 pounds?

```
>> % less than 400 pounds
>> less400 = normcdf(400, 500, 150)
less400 =
    0.2525
>> % more than 700 pounds
>> more700 = 1 - normcdf(700, 500, 150)
more700 =
    0.0912
>> % between 450 and 550 pounds
>> middle100 = normcdf(550, 500, 150) - normcdf(450, 500,␣
 ↪150)
middle100 =
    0.2611
```