Inference, Models and Simulation for Complex Systems CSCI 7000-001
Lecture 0 23 August 2011
Prof. Aaron Clauset

# 1 A brief primer on probability distributions

## 1.1 Probability distribution functions

A probability density function (pdf) is defined as any function that satisfies the equation

$$1 = C \int_{-\infty}^{+\infty} p(x) \mathrm{d}x \ . \tag{1}$$

In many cases, the range of the function is somewhat less than $(-\infty, +\infty)$ (we'll revisit this point in a moment).

Generally, to convert some function $p(x)$ into a pdf, we simply need to normalize it. (NB: not all functions can be normalized; think back to infinite sums and series from calculus.) To normalize a function, we simply must identify the constant $C$ that makes Eq. (1) true.

For instance, consider an exponential function $p(x) = \mathrm{e}^{-\lambda x}$, defined on the interval $[0, +\infty)$[1]:

$$1 = C \int_{0}^{+\infty} \mathrm{e}^{-\lambda x} \mathrm{d}x \ .$$

Solving this equation for $C$ yields $C = \lambda$, and thus the pdf for an exponential distribution is $\Pr(x) = \lambda \mathrm{e}^{-\lambda x}$.

As we'll see in the next lecture, it can be useful to define a pdf over a more limited interval, e.g., the "tail" interval $[x_{\min}, +\infty)$.[2] Solving the appropriate version of Eq. (1) reveals that this change only affects the normalization constant $C$:

$$\begin{aligned} \Pr(x) &= \lambda \mathrm{e}^{\lambda x_{\min}} \mathrm{e}^{-\lambda x} \\ &= \lambda \mathrm{e}^{-\lambda (x - x_{\min})} \ . \end{aligned}$$

[1] Formally, our function must be defined on the entire interval $(-\infty, +\infty)$, e.g., in the case of the exponential, we would define it piecewise such that $p(x) = 0$ for $(-\infty, 0)$ and $p(x) = \mathrm{e}^{-\lambda x}$ for $[0, +\infty)$. But this notation is cumbersome. It's simpler and equally clear to assume that unless otherwise stated, the function is defined as given on the specified range and 0 everywhere else.

[2] This is called the "tail" of the distribution because it isolates the part of the distribution corresponding to low-probability events, which, if you squint at the graph, kind of looks like a dog's tail. Or something.

### 1.2 Cumulative distribution functions

The cumulative distribution function (cdf) is defined as the fraction of density that falls below some particular value $x$. Mathematically, we say

$$\begin{aligned}\Pr(X < x) &= C\int_{-\infty}^{x} p(y)\mathrm{d}y \\ &= \lambda \mathrm{e}^{\lambda x_{\min}} \int_{x_{\min}}^{x} \mathrm{e}^{-\lambda y}\mathrm{d}y \\ &= 1 - \mathrm{e}^{-\lambda(x - x_{\min})} \quad .\end{aligned}$$

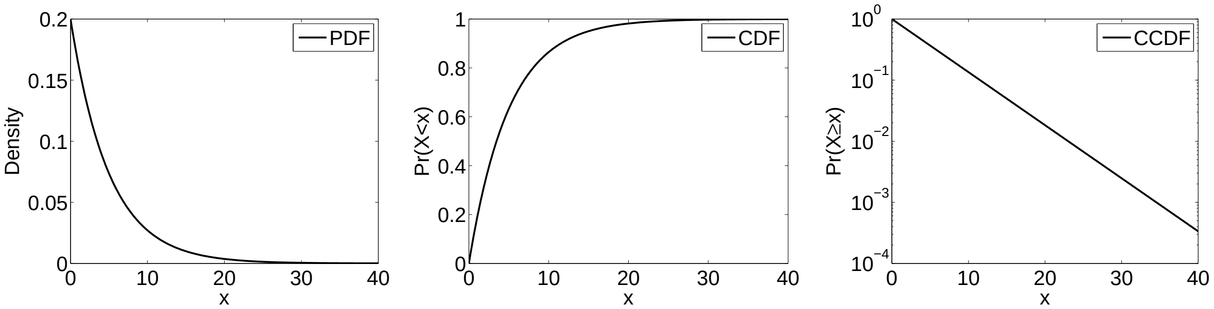
Here, we're using the notation that $X$ denotes a *random variable*: a value whose distribution is given by $\Pr(x)$. Note that the form of the cdf is very similar to the form of the pdf for the exponential; this is a special property of exponential distributions (and a few other functions).

### 1.3 Complementary cumulative distribution functions

Another useful tool will be the complementary cumulative distribution function (ccdf), which is defined as $1 - \text{cdf} = 1 - \Pr(X < x) = \Pr(X \geq x)$. Mathematically, we say

$$\begin{aligned}\Pr(X \geq x) &= C\int_{x}^{\infty} p(y)\mathrm{d}y \\ &= 1 - \lambda \mathrm{e}^{\lambda x_{\min}} \int_{x_{\min}}^{x} \mathrm{e}^{-\lambda y}\mathrm{d}y \\ &= \mathrm{e}^{-\lambda(x - x_{\min})} \quad .\end{aligned}$$

To illustrate all three types of distribution functions, the next page shows the pdf, cdf and ccdf for the exponential distribution, with $x_{\min} = 0$.



### 1.4 Independent and identically distributed (iid) variables

Throughout this class, we will model complex systems using probabilistic models. These are helpful because data produced by complex systems often exhibit fluctuations and variability. Probabilistic models are a natural way to represent this variability.

One of the assumptions we will often make about our data is that observations are independent and identically distributed (iid). This means that if we are given a set of data $\{x_i\}$, each of these $x_i$ observations is an independent draw from a fixed ("stationary") probabilistic model. Independence means that

$$\Pr(x_1) \text{ and } \Pr(x_2) = \Pr(x_1)\Pr(x_2) \ \ .$$

That is, the probability of observing two values $x_1$ and $x_2$ is simply the probability of observing $x_1$ multiplied by the probability we observe $x_2$. This implies what's called *conditional independence*, that is,

$$\Pr(x_2 \,|\, x_1) = \Pr(x_2) \ \ .$$

To give you some intuition about what this means, consider the familiar process of rolling dice. If we assume that we have 2 regular 6-sided dice and that they are "fair," then each of the 6 values occurs with equal probability. If we throw the dice together, the values they display are iid random variables. If we throw the dice separately, the values they display are iid random variables. If we throw a long sequence of the dice, all of the values we observe are iid random variables.

### 1.5 Central Limit Theorem

In some places in the class, we will implicitly or explicitly invoke the *central limit theorem*,[3] which is a fundamental result from probability theory.

Suppose that we are given a sample of $n$ iid random variables, denoted $\{X_1, X_2, \ldots, X_n\}$, each with expected value (average) $\mu$ and variance $\sigma^2$. Let $S_n$ be the *sample average*[4] of the values,

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \ \ .$$

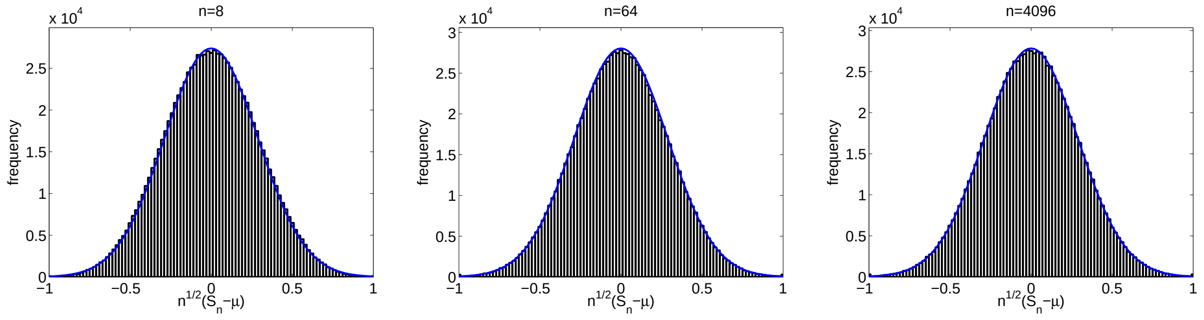
[3] In the lecture about Lévy flights, we will likely meet the generalized version of the central limit theorem, which considers random variables with potentially infinite variance. The classic version assumes that the variance is finite, which is often, but not always, a reasonable assumption.

[4] We call it the sample average because it's the average of a sample of observations (random variables). The true average is defined by the underlying generative process. That is, if $p(x)$ is the probability of $x$, then the (non-sample = true) average is defined as $\langle x \rangle = \bar{x} = \int_{-\infty}^{+\infty} x\, p(x) \mathrm{d}x$.

Note that because the $X_i$ are random variables, the sample average $S_n$ is also a random variable. The central limit theorem states that, in the asymptotic (thermodynamic) limit of $n \to \infty$, the random variable $\sqrt{n}(S_n - \mu)$ converges in distribution to a Normal (Gaussian) distribution $\mathrm{N}(0, \sigma^2)$, i.e., a Normal distribution with zero-mean and $\sigma^2$ variance.[5]

That is, in the limit of large sample sizes (infinite data), the error in our estimate (the difference between the sample average $S_n$ and the true average) converges on 0 at a rate equal to $\sqrt{n}$. (We'll revisit this point below when we discuss "standard error" estimates.)

The central limit theorem is an easy thing to demonstrate using a simple numerical simulation. Consider a uniform distribution $X \sim \mathrm{U}(0, 1)$. This distribution has an expected value $\mathrm{E}(X) = 1/2$. The following figures show the distribution of $10^6$ estimates of the rescaled error in our estimate $\sqrt{n}(S_n - \mu)$, for three sizes of samples $n = \{10^0, 10^1, 10^3\}$.



Comfortingly, the distribution is indeed Normally distributed, even for small sample sizes. (It's hard to see, but these three histograms are in fact different from each other.)

Here's the Matlab code that produces these figures:

```
v = [8 64 4096];     % sample sizes to try
m = 10^6;            % number of trials in distribution

for n=v
    ms = zeros(m,1);
    for j=1:m
        ms(j) = mean(rand(n,1));
    end;
```

[5] The Normal distribution is defined $\Pr(x) = 1/(\sigma\sqrt{2\pi})\mathrm{e}^{(x-\mu)^2/2\sigma^2}$.

```
    figure;          % make a pretty figure
    h=hist(sqrt(n).*(ms-0.5),(-1:0.02:1));
    g=bar((-1:0.02:1),h); hold on;
    plot((-1:0.02:1),(max(h).*0.73).*pdf('norm',(-1:0.02:1),0,0.2887),'b-','LineWidth',2);
    hold off;
    set(g,'BarWidth',1.0,'FaceColor','none','LineWidth',2);
    set(gca,'XLim',[-1 1],'YLim',[0 1.1*max(h)],'FontSize',16);
    title(strcat('n=',num2str(n)),'FontSize',16);
    xlabel('n^1^/^2(S_n-\mu)','FontSize',16);
    ylabel('frequency','FontSize',16);
end;
```

### 1.5.1 Standard errors and error bars

The central limit theorem also underlies the way we calculate *standard error* estimates, which is a notion of uncertainty and which is often used to calculate error bars on estimates.

That is, whenever we calculate a sample average from some data, because we assume the data are random variables, our average is only an estimate of the true average value, and we should also report our uncertainty. This is what people mean when they say $\hat{\theta} \pm$ s.e., where $\hat{\theta}$ denotes the estimated value ($\theta$, with no hat, is the true value) and s.e. is conventionally a standard error, where $\hat{\sigma}$ is the sample standard deviation.

The formula for the standard error is simply

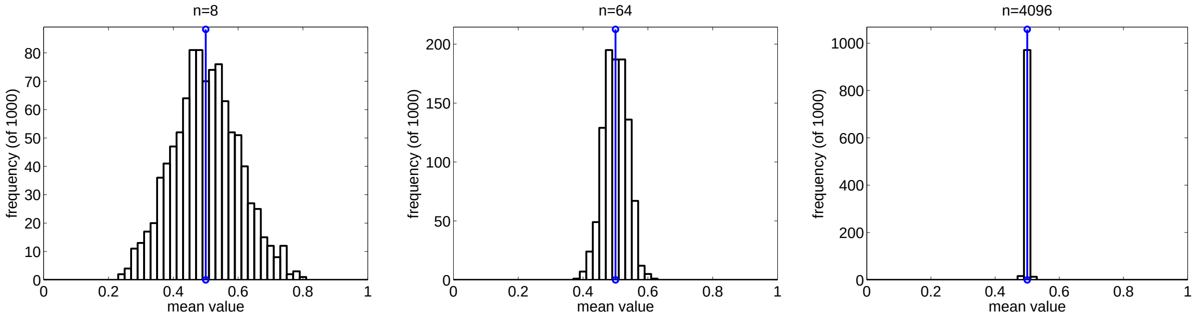
$$\text{s.e.} = \frac{\hat{\sigma}}{\sqrt{n}} \ .$$

The $\sqrt{n}$ here is, in fact, the same $\sqrt{n}$ from the central limit theorem. Do you see why?

## 1.6 Law of Large Numbers

In the above example, the factor of $\sqrt{n}$ is important as it rescales the distribution of sample means so that it remains stationary (fixed), as the number of observations in the sample increases. If we eliminate it, we can see the impact of the law of large numbers, another fundamental result from probability theory. The "strong" version of this law states that in the limit of infinite data $n \to \infty$, the sample mean converges on the expected value $S_n \to \mu$. That is, with more and more data, our estimates should become more and more accurate.

Here's a simple demonstration of this, using the same simulation as above but now showing the raw distribution of sample means $S_n$.



Indeed, as the sample size increases, the distribution of the mean values becomes increasingly "concentrated" around the true expected value $\mathrm{E}(X) = 1/2$.

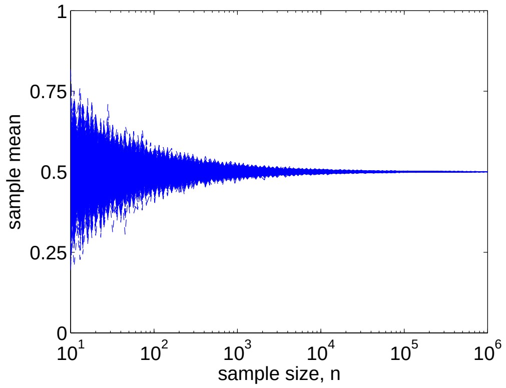
Here's the Matlab code that produces these figures:

```
v = [8 64 4096];     % sample sizes to try
m = 1000;            % number of trials in distribution

for n=v
    ms = zeros(m,1);
    for j=1:m
        ms(j) = mean(rand(n,1));
    end;

    figure;          % make a pretty figure
    h=hist(ms,(0:0.02:1));
    g=bar((0:0.02:1),h); hold on;
    plot([0.5 0.5],[0 1.09*max(h)],'bo-','LineWidth',2); hold off;
    set(g,'BarWidth',1.0,'FaceColor','none','LineWidth',2);
    set(gca,'XLim',[0 1],'YLim',[0 1.1*max(h)],'FontSize',16);
    title(strcat('n=',num2str(n)),'FontSize',16);
    xlabel('mean value','FontSize',16);
    ylabel('frequency (of 1000)','FontSize',16);
end;
```

Another way to show the law of large numbers is to plot the sample mean as a function of sample size, like the next figure. Here, I'm computing $S_n$ many times at each value of $n$ in order to illustrate the way the variance changes with $n$. Notice that for large $n$, the estimate is extremely accurate.



## 1.7 Caveats

There are a few caveats, of course. The central limit theorem is quite general and holds even when some of its assumptions (iid variables with finite mean and variance) are violated. If the independence assumption is violated, the convergence rate is decreased, meaning that our estimates are less accurate than we might believe, if we were to assume independence. Even if we have non-stationary processes, some statistical calculations may still be reliable, so long as the non-stationary effects are not too severe. The law of large numbers also holds when some of its assumptions are violated, e.g., it holds even when the variance of the generating distribution is infinite, which has the impact of, again, slowing down the convergence.

The take-home message here is that these are powerful and useful assumptions to make in analyzing and modeling data from complex systems, but they are mainly starting points. Complex systems often exhibit mixed or non-stationary processes, feedback loops or long-range "memory," and all of these effects can reduce the accuracy of the iid assumptions. This is not to say that you should not start with the simplest and strongest assumptions (iid random variables); rather, you should start there but then think carefully about how to improve your models relaxing the unrealistic assumptions.