Monte Carlo Method: Probability

John Burkardt (ARC/ICAM)
Virginia Tech

.....

Math/CS 4414:
"The Monte Carlo Method: PROBABILITY"
https://people.sc.fsu.edu/~jburkardt/presentations/monte_carlo_probability.pdf

.....

ARC: Advanced Research Computing
ICAM: Interdisciplinary Center for Applied Mathematics

26-28-30 October 2009

VirginiaTech

Monte Carlo Method: Probability

- **Overview**
- Discrete Probability
- Continuous Probability
- Fitting Distributions to Data

VirginiaTech

Overview


Help
13 36 24 3 15
24
4
24
$1
$101
AMOUNT TO
BET PER CLICK
Practice $696.00
min max current
Inside $1 $250 $202
Outside $1 $250 $102
Repeat
Spin
Remove
ClearAll
SUN PALACE CASINO
1st 12
2nd 12
3rd 12
1-18
Even
Odd
19-36
2to1
2to1
2to1

Overview

This is the first of several talks on the Monte Carlo Method (MCM).

The Monte Carlo Method uses random numbers to try to determine the answer to problems. This seems like a peculiar way to do mathematics!

Although many mathematical problems have efficient and accurate algorithms for their solution, there are times when the problem is too big, too hard, too irregular for such an approach.

The Monte Carlo method can always give you an approximate answer, and if you are willing to work a little harder, it can improve that approximation.

VirginiaTech

Overview

The Monte Carlo Method is based on principles of probability and statistics.

To begin our discussion, we will look at some basic ideas of probability; in particular, the idea of how the behavior of a system can be described by a curve called the probability density function, and how the properties of that curve can help us to understand a system, or to simulate its behavior.

VirginiaTech

Overview

Historically, the birth of probability arose when a gambler wrote to Fermat, asking him whether it was possible to settle up the bets in a particular dice game that had gotten interrupted. Fermat had some ideas, but he wrote to Pascal and between them they worked out methods that are still in use today.

VirginiaTech

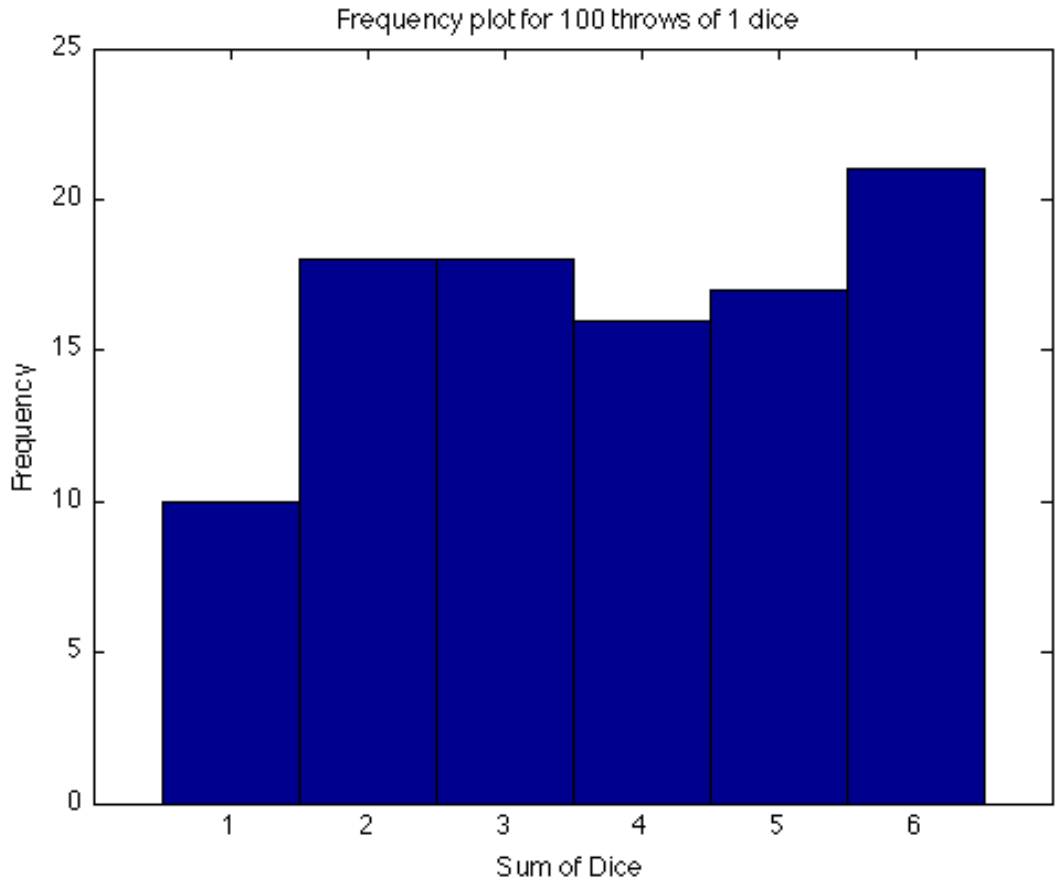
Monte Carlo Method: Probability

- Overview
- **Discrete Probability**
- Continuous Probability
- Fitting Distributions to Data

VirginiaTech

Discrete Probability: Frequency Plot For 1 Die

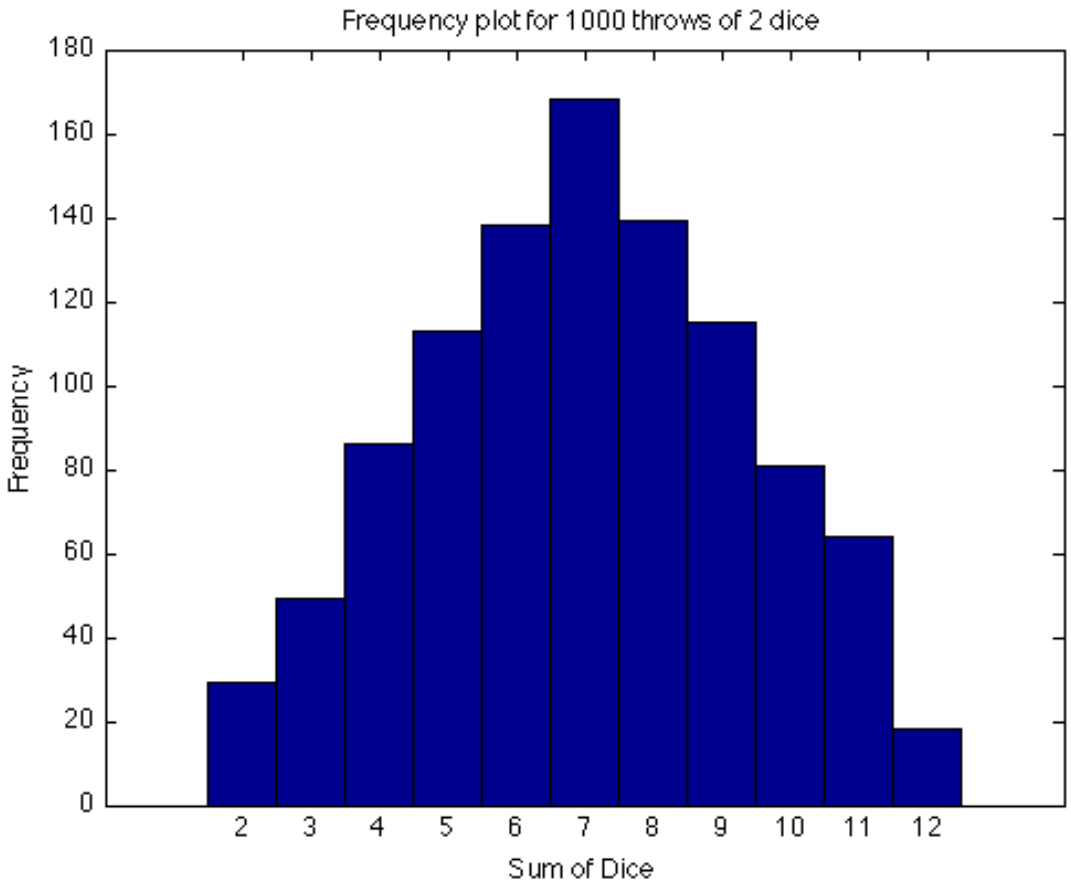
If we roll a die hundreds of times, keep track of the results, and make a bar graph, we have constructed a **frequency plot**:



VirginiaTech

Discrete Probability: Frequency Plot For 2 Dice

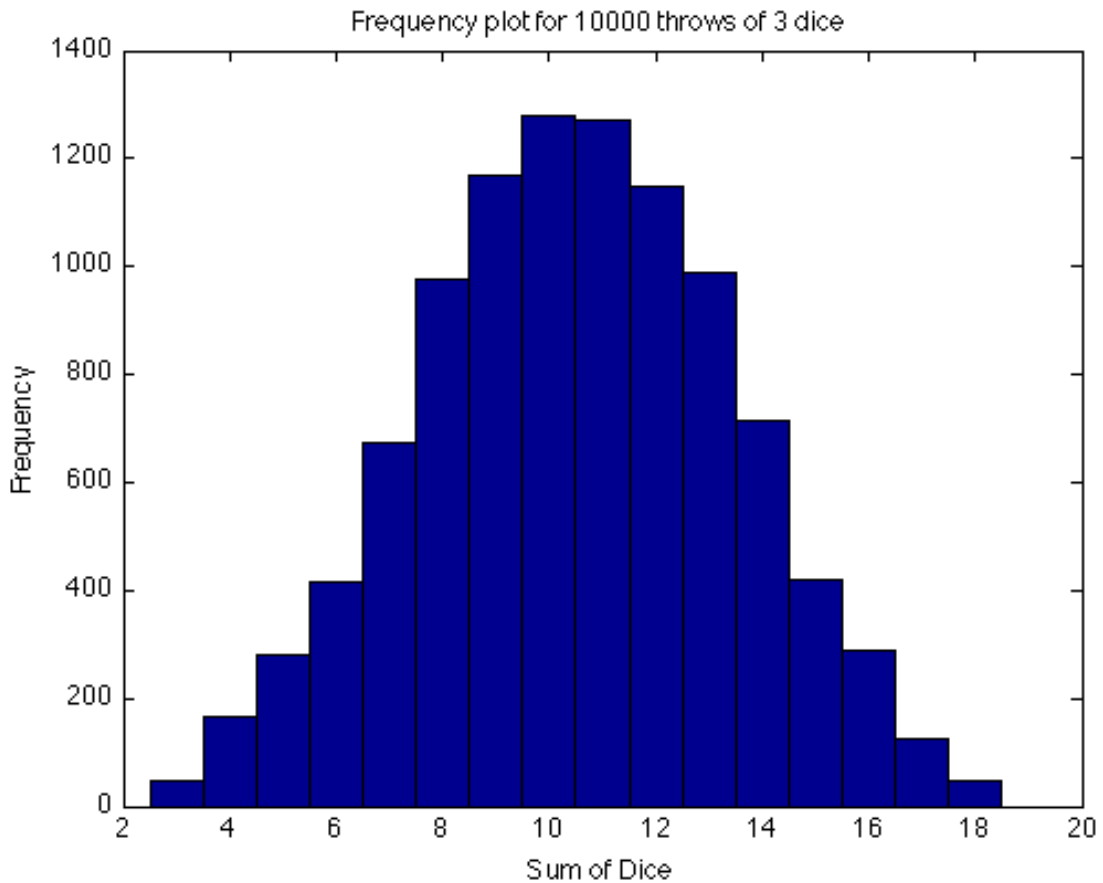
Suppose now that we roll two dice. The individual outcomes on each die are still equally likely, but now the total of the two dice shows a nonuniformity: there is only one way to score 2, but many ways to score 7:



VirginiaTech

Discrete Probability: Frequency Plot For 3 Dice

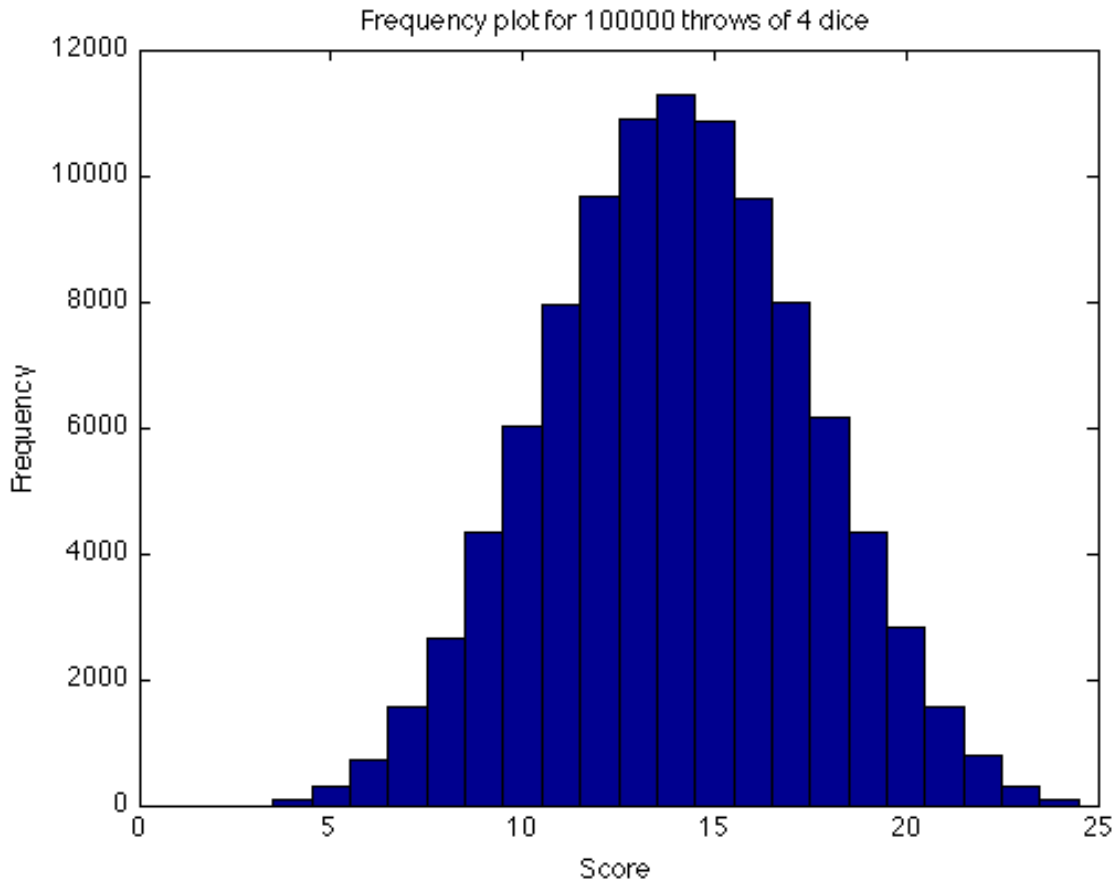
We can use 3 dice:



VirginiaTech

Discrete Probability: Frequency Plot For 4 Dice

By the time we use 4 dice, the plot is looking very much as though there is an underlying function $f(x)$ that is influencing the shape.



VirginiaTech

Discrete Probability: Hints of a Normal Distribution

In fact, this plot suggests the normal curve, or "bell shaped distribution", even though we expect to see that curve only for continuous variables.

This is an example of the following true statement:

The sum of several uniform random variables behaves like a normal random variable.

We are seeing this already with a sum of just 4 uniform random variables.

We will come back to this idea when we are ready to talk about continuous variables!

VirginiaTech

Discrete Probability: Frequency Plot to Probability

A frequency plot is simply a record of what happened. However, if we keep records long enough, we may see that the frequency plot can be used to make statements about what is likely to happen in the future. For two dice, we can say that a score of 7 is very likely. We can even guess that it is about twice as likely as a 3.

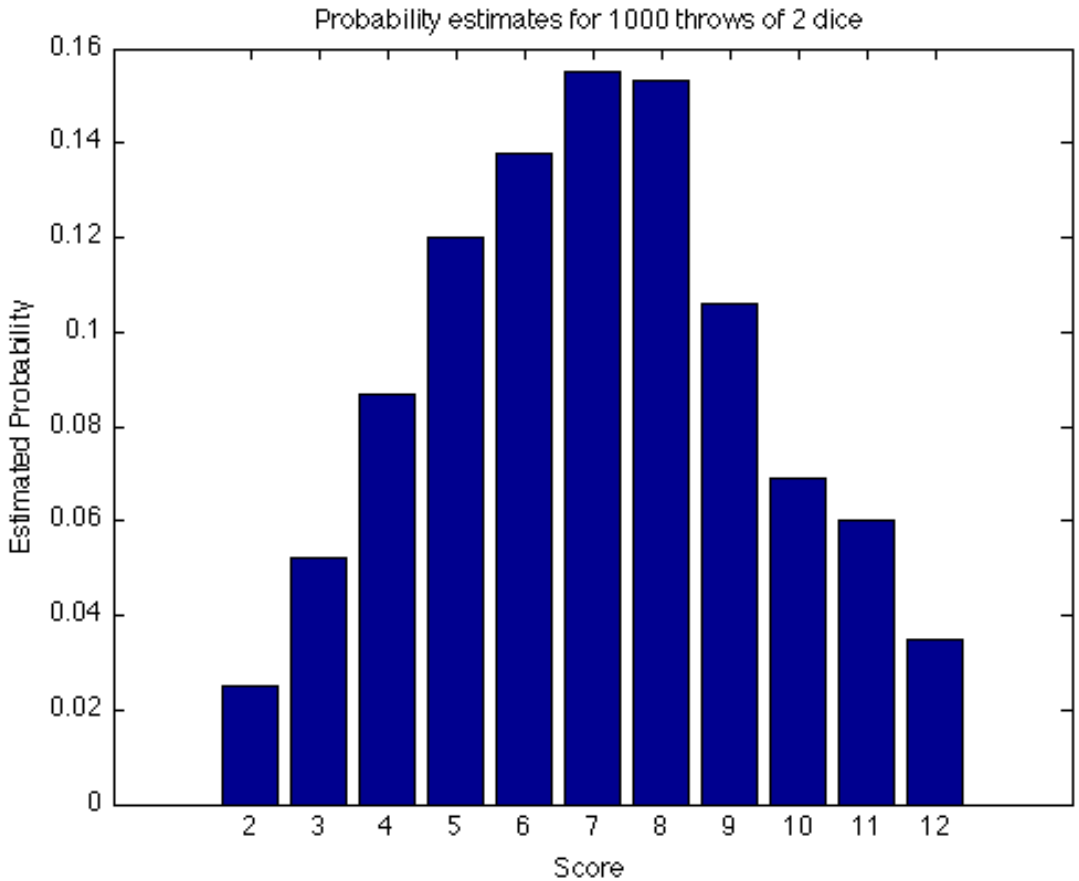
This suggests that there is an underlying *probability* that influences which outcomes occur. In cases where we are simply collecting data, we can turn our frequencies into estimates of probability by normalizing by the total number of cases we observed:

$$\text{estimated probability of result \#i} = \frac{\text{frequency of result \#i}}{\text{total number of results}}$$

VirginiaTech

Discrete Probability: Probability Plot for 2 Dice

We redraw the plot with the vertical scale the estimated probability.



VirginiaTech

Discrete Probability: Probability Table for 2 Dice

For a pair of fair dice, the outcome is the pair of values:

	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
2	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
2	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

The probability of the outcome (3,5), for instance, is $\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$, which we get by multiplying the probability of a 3 with the first die times the probability of a 5 with the second die.

VirginiaTech

Discrete Probability: Probability Table for 2 Dice

Even though each outcome is equally likely, when we take the *sum* of the dice, there are several outcomes that produce the same sum, so, considering the sum as the outcome we're actually interested in, we get unequal probabilities.

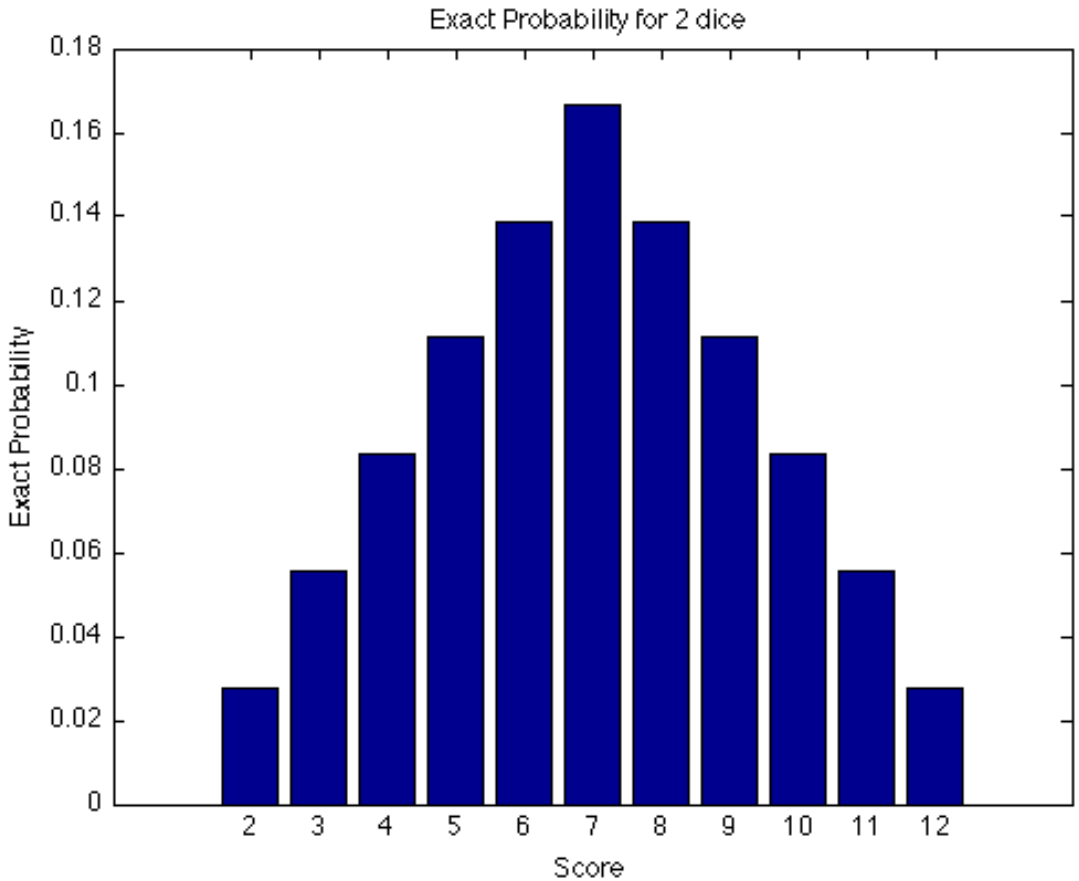
For a pair of fair dice, the exact probability of each total is:

2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
.03	.06	.08	.11	.14	.17	.14	.11	.08	.06	.03

VirginiaTech

Discrete Probability: Exact Probability Plot for 2 Dice

Here is the exact probability plot for two dice:



VirginiaTech

Discrete Probability: Related Problems

I will not explain the details, but you should be able to repeat these calculations for the following variations:

- Suppose we use 2 dice, but they are loaded? Now what is the probability of each pair of dice values? What are the probabilities of a given total?
- Suppose we use 3 fair dice instead of 2?

VirginiaTech

Discrete Probability: Probability Density Functions

This is our first example of a *probability density function* or PDF, which assigns a probability $p(x)$ to each outcome x in our set X of all possible outcomes.

It's a special case, since there are only finitely many possible outcomes; we call this a *discrete* problem.

For the case of two dice, the number of outcomes is **N**=11.

We have two obvious requirements for a discrete PDF:

1. $0 \leq p(x_i)$ for each x_i
2. $\sum_{i=1}^{N} p(x_i) = 1$

VirginiaTech

Discrete Probability: Mean and Variance

Two important characteristics of a PDF are its *mean* and *variance*.

For a discrete PDF, these quantities are easy to calculate:

$$\text{ave} = \sum_{i=1}^{N} p(x_i) * x_i$$

$$\text{var} = \sum_{i=1}^{N} p(x_i) * (x_i - \text{ave})^2$$

Notice that we do not divide by **N**; that is taken care of through the values of $p(x)$!

VirginiaTech

Discrete Probability: Mean and Variance

For example, for our problem with 2 dice, using the exact probabilities:

the average is:

$$\text{ave} = \frac{1}{36} * 2 + \frac{2}{36} * 3 + \frac{3}{36} * 4 + \ldots + \frac{1}{36} * 12 = 7$$

the variance is:

$$\text{var} = \frac{1}{36} * (2-7)^2 + \frac{2}{36} * (3-7)^2 + \frac{3}{36} * (4-7)^2 + \ldots + \frac{1}{36} * (12-7)^2 = 5.8\ldots$$

VirginiaTech

Discrete Probability: Expected Value

Especially with dice games, there may be a payoff associated with certain rolls. The **expected value** computes the average payoff you would get, if you played many games.

If the payoff for outcome x is $v(x)$, then the expected value is

$$E(v(x)) = \sum_{i=1}^{N} p(x)v(x)$$

If we roll two dice, and receive \$10 if the sum is divisible by 3, \$20 if it is divisible by 4, and nothing for other rolls, then the nonzero payoffs are for 3, 4, 6, 8, 9 and 12:

$$\begin{aligned} E(v(x)) = &\frac{2}{36} * \$10 + \frac{3}{36} * \$20 + \frac{5}{36} * \$10 + \frac{5}{36} * \$20 \\ &+ \frac{4}{36} * \$10 + \frac{1}{36} * \$30 = \$7.50 \end{aligned}$$

VirginiaTech

Discrete Probability: the CDF

If we have a discrete system with a known PDF, the value $pdf(x_i)$ is the probability that the outcome x_i will occur.

But suppose we wanted to know the chances of rolling a 7 or less, using two dice. This is the probability that the outcome is less than or equal to 7; it is so important that it has its own name, the *cumulative density function* or **CDF**.

$$\begin{aligned} \mathrm{cdf}(x) &= \text{prob outcome is less than or equal to } x \\ &= \sum_{y \le x} \mathrm{pdf}(y) \end{aligned}$$

VirginiaTech

Discrete Probability: the CDF

cdf(x) = probability that the outcome is less than or equal to **x**.

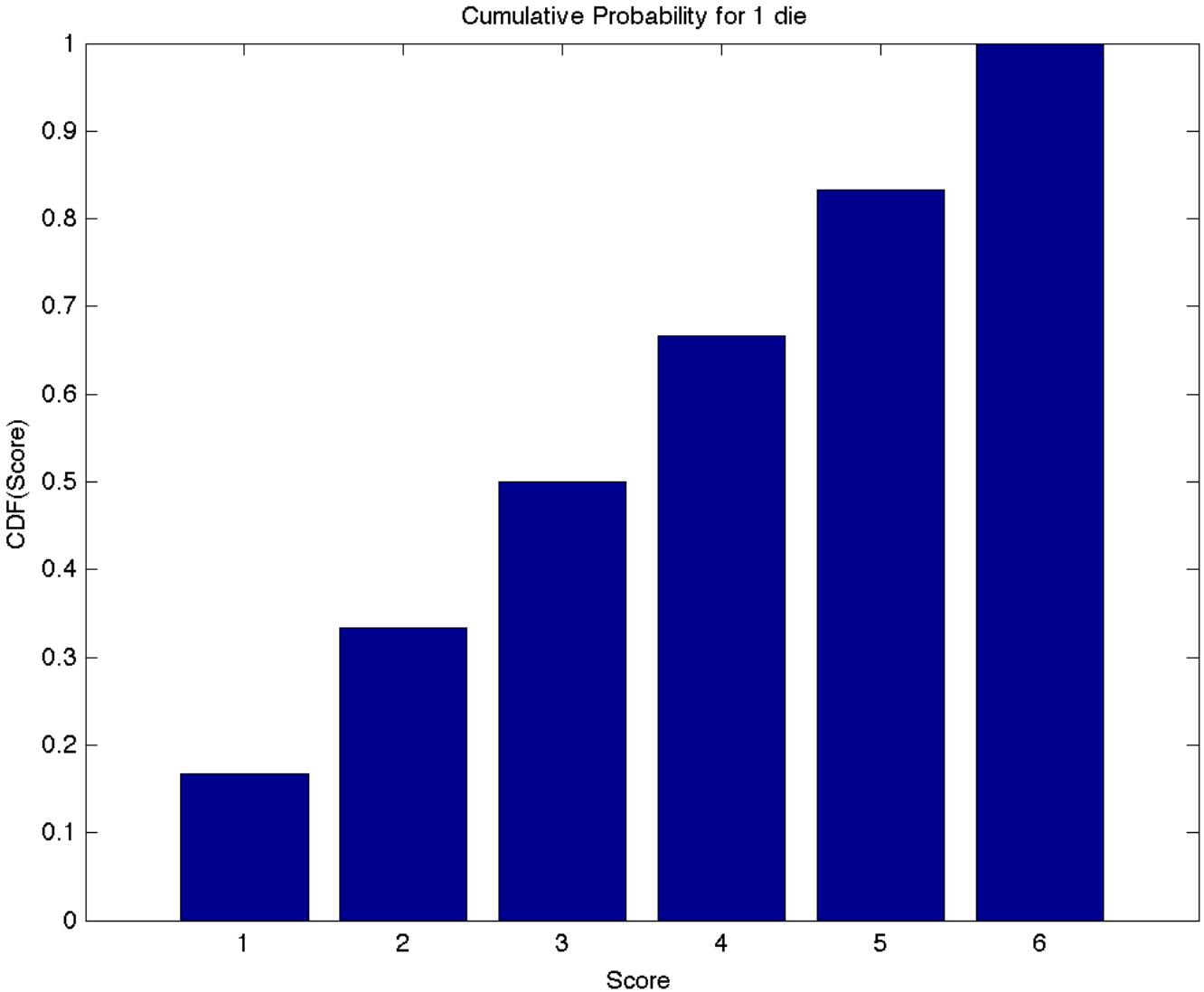
This implies:

- cdf(x) is a piecewise constant function defined for all x;
- cdf(x) = 0 for x less than smallest possible outcome;
- cdf(x) = 1 for x greater than or equal to largest outcome;
- cdf(x) is essentially the discrete integral of pdf(x);
- cdf(x) is monotonic (and hence invertible);
- the probability that x is between x_1 and x_2 ($x_1 < x \leq x_1$) is $\mathrm{cdf}(x_2) - \mathrm{cdf}(x_1)$.

VirginiaTech

Discrete Probability: CDF Plot for 1 Die

Here is the CDF for 1 die:



Discrete Probability: CDF Table for 2 Dice

Recall for a pair of fair dice, the exact probability of each total is:

2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$
.03	.06	.08	.11	.14	.17	.14	.11	.08	.06	.03

For a discrete case like this, it's easy to make the corresponding cumulative density function table:

2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$
.03	.08	.16	.28	.42	.58	.72	.83	.92	.97	1.00

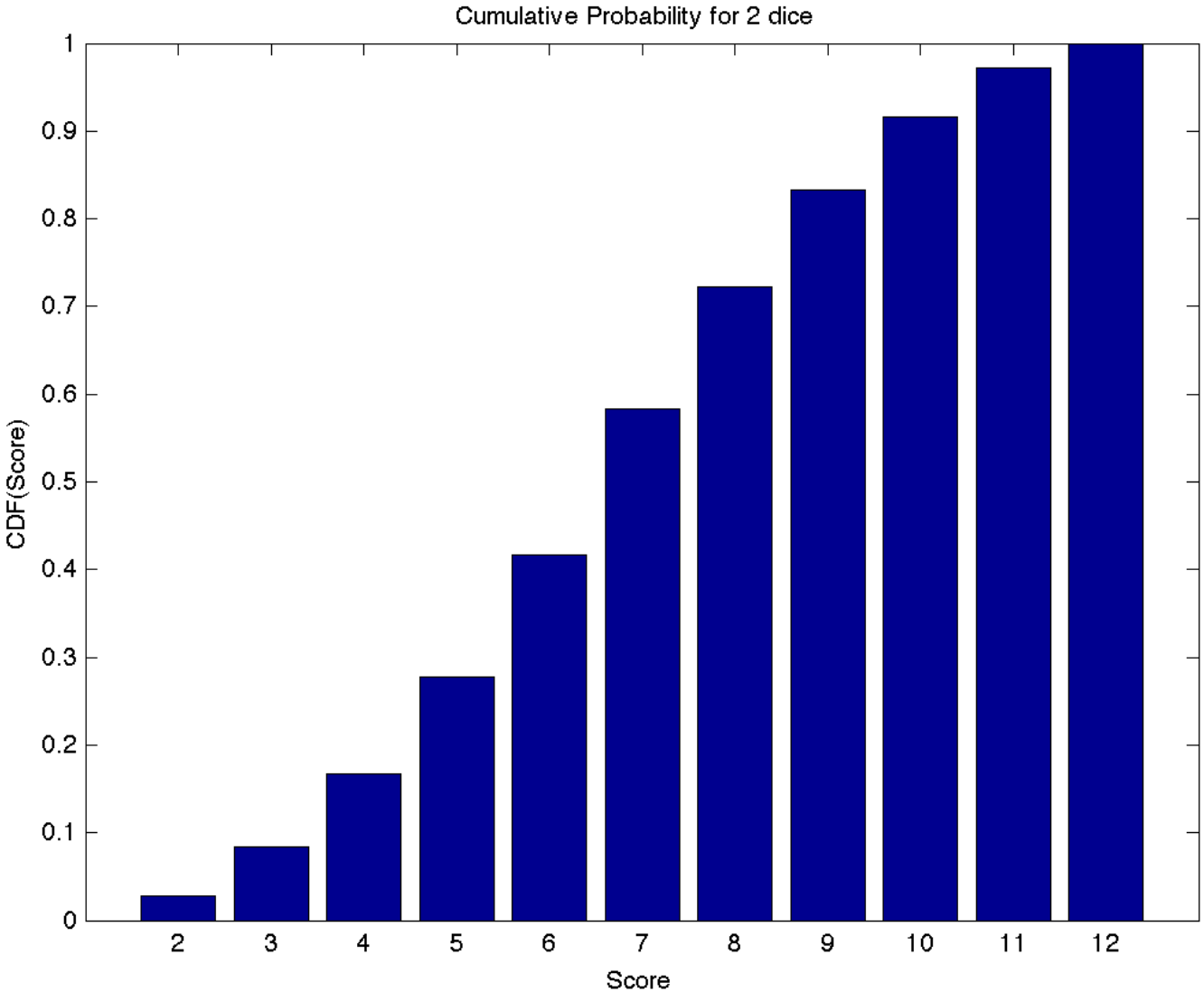
The CDF is actually defined for all x.

The table tells us that $\mathbf{cdf}(4.3) = 0.16$ and $\mathbf{cdf}(15) = 1$.

VirginiaTech

Discrete Probability: CDF Plot for 2 Dice

Here is the CDF for 2 dice:



iaTech

Discrete Probability: Program for CDF Plot

```
figure ( 1 )
x = 2 : 12;
pdf = [1,2,3,4,5,6,5,4,3,2,1]/36;
bar ( x, pdf )
xlabel ( 'Score' )
ylabel ( 'PDF(Score)' )
title ( 'PDF for 2 dice' );

figure ( 2 )
x = 2 : 12;
pdf = [1,2,3,4,5,6,5,4,3,2,1]/36;
cdf = cumsum ( pdf )
bar ( x, cdf )
xlabel ( 'Score' )
ylabel ( 'CDF(Score)' )
title ( 'CDF for 2 dice' );
```

VirginiaTech

Discrete Probability: Using the CDF for Simulation

The CDF can be used to simulate the behavior of a discrete system.

Suppose a system has **M** possible outcomes, and we want to simulate its behavior. We simply pick a random number **r**, and search for the outcome x_i with the property that $\text{cdf}(x_{i-1}) < r \leq \text{cdf}(x_i)$.

We will have to treat the first case ($i = 1$) specially.

VirginiaTech

Discrete Probability: Using the CDF for Simulation

Here is another method, which starts at case $i = 1$, and returns x_i as the outcome as soon as it finds the value of i for which $r <= \text{cdf}(x_i)$: the first

1. Let $r = \text{rand}()$;
2. Initialize $i = 1$;
3. Begin loop;
4. Set $x = x_i$;
5. if $r \leq \text{cdf}(x_i)$ return x_i;
6. else $i = i + 1$;
7. End loop.

VirginiaTech

Discrete Probability: Using the CDF for Simulation

If we are working with MATLAB, we can use some special features:

- **cumsum** can compute the CDF from the PDF:
  cdf = cumsum (pdf);
- **find** returns the entries in an array where something is true:
  index = find (cdf(i-1) $<$ r & r $\leq$ cdf(i));
- **length** tells us how many entries are in a vector.

VirginiaTech

Discrete Probability: Using the CDF for Simulation

```
function two_dice ( n )
%
%  Simulate N throws of two fair dice.
%
  pdf = [0,1,2,3,4,5,6,5,4,3,2,1] / 36;
%
%  Compute the CDF:
%
  cdf = cumsum ( pdf );
%
%  Throw N times:
%
  r = rand ( n, 1 );
%
%  Each R corresponds to an X for which CDF(X-1) < R <= CDF(X).
%
  for x = 2 : 12
    match = ( cdf(x-1) < r & r <= cdf(x) );
    score ( match ) = x;
  end
%
%  Compute the frequency and estimated probability of each score X.
%
  for x = 1 : 12
    match = find ( score == x );
    freq(x) = length ( match );
  end

  for x = 1 : 12
    pdf_est(x) = freq(x) / n;
  end
```

VirginiaTech

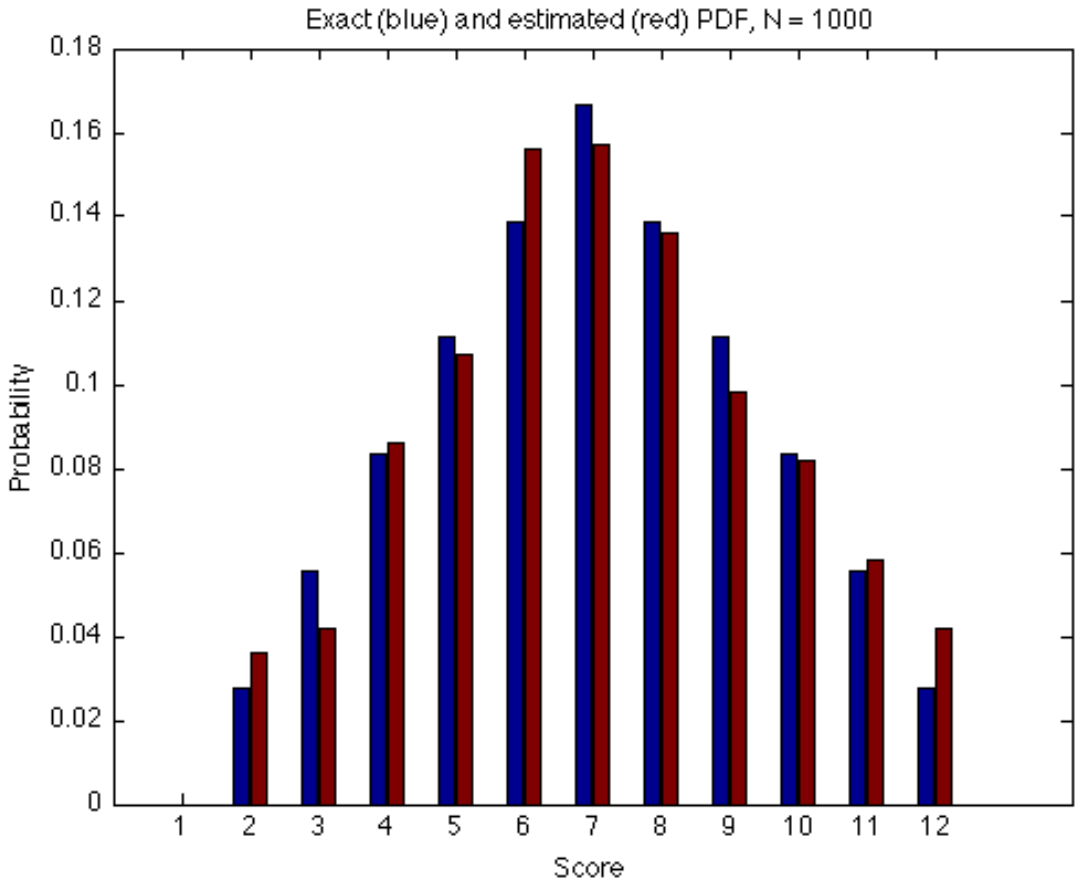
Discrete Probability: Using the CDF for Simulation

```
%
%  Plot the estimated PDF versus the PDF
%
  x = 1 : 12;

  y(1:12,1) = pdf(1:12);
  y(1:12,2) = pdf_est(1:12);

  bar ( x, y )
  title_string = sprintf ( 'Exact (blue) and estimated (red) PDF, N = %d', n );
  title ( title_string )
  xlabel ( 'Score' )
  ylabel ( 'Probability' )

  return
end
```

VirginiaTech

Discrete Probability: Comparison of Plots

Compare the exact and estimated PDF's for two fair dice:



VirginiaTech

Discrete Probability: Assignment

A factory makes dice whose six sides have the PDF:
$\frac{1}{10}, \frac{0}{10}, \frac{1}{10}, \frac{2}{10}, \frac{4}{10}, \frac{2}{10}$.
Write a program that simulates rolling TWO such dice 1,000 times.
You must compute the 6x6 table of probabilities, add up the probabilities to get the PDF for each score of 2 through 12.

For 10 points, turn in 3 numbers and one plot:

1. What is the average when rolling ONE of these dice?
2. What is the variance when rolling ONE of these dice?
3. What is the average when rolling TWO of these dice?
4. Plot the exact and estimated PDF for two dice;

The program **fair_dice.m** (in Scholar) may be useful to you.

Due on MONDAY, 2 November (hand in, or email, or Scholar).

VirginiaTech

Monte Carlo Method: Probability

- Overview
- Discrete Probability
- **Continuous Probability**
- Fitting Distributions to Data

VirginiaTech

Continuous Probability

Since most variables have a continuous range, it's important to be able to extend the ideas of probability that we considered for discrete problems.

Let's start with the frequency plots or bar graphs, which we called probability density functions or PDF's. It's easy to look at the bar graph and imagine replacing it by a smooth, continuous curve.

However, before, it was easy to say that the height of the curve was a count of how many times that value was encountered, or, if we normalized it, the relative frequency of that value. For a continuous problem, we have an infinite number of possible values, so, in a sense, every value must have a relative frequency of 0!

Continuous Probability

If we really measure exactly the output of some continuous random process, then we won't get nice bar graphs. We'll probably never get the same value twice, so we just have lots of dots on the axis.

We can *see* that the dots seem to be denser in some areas than others, but in this form, the information is hard to analyze.

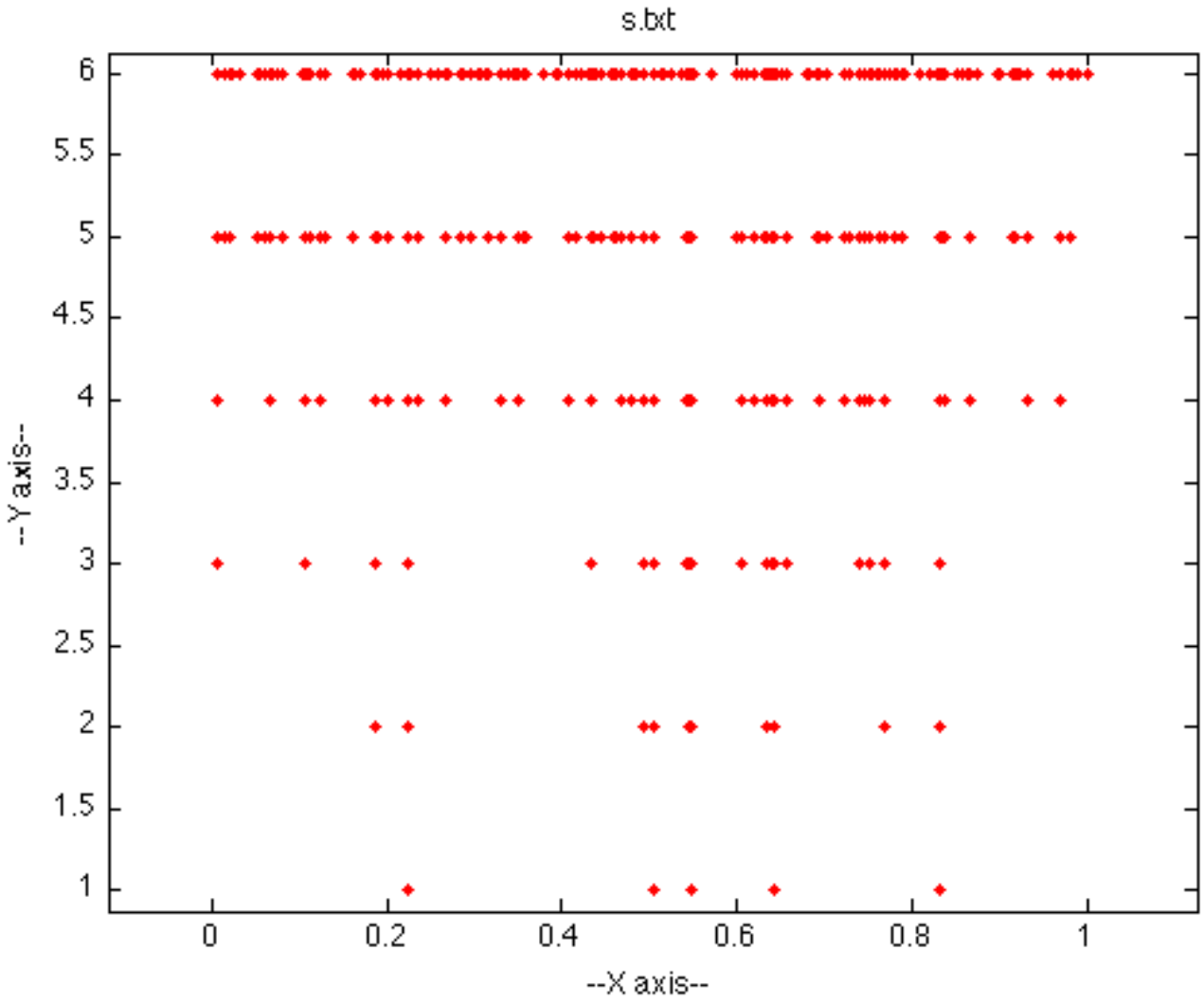
We need a method to convert this kind of raw data so that it is simpler to digest and analyze and plot!

VirginiaTech

Continuous: Uniform Samples

n = [5,10,20,40,80,160]

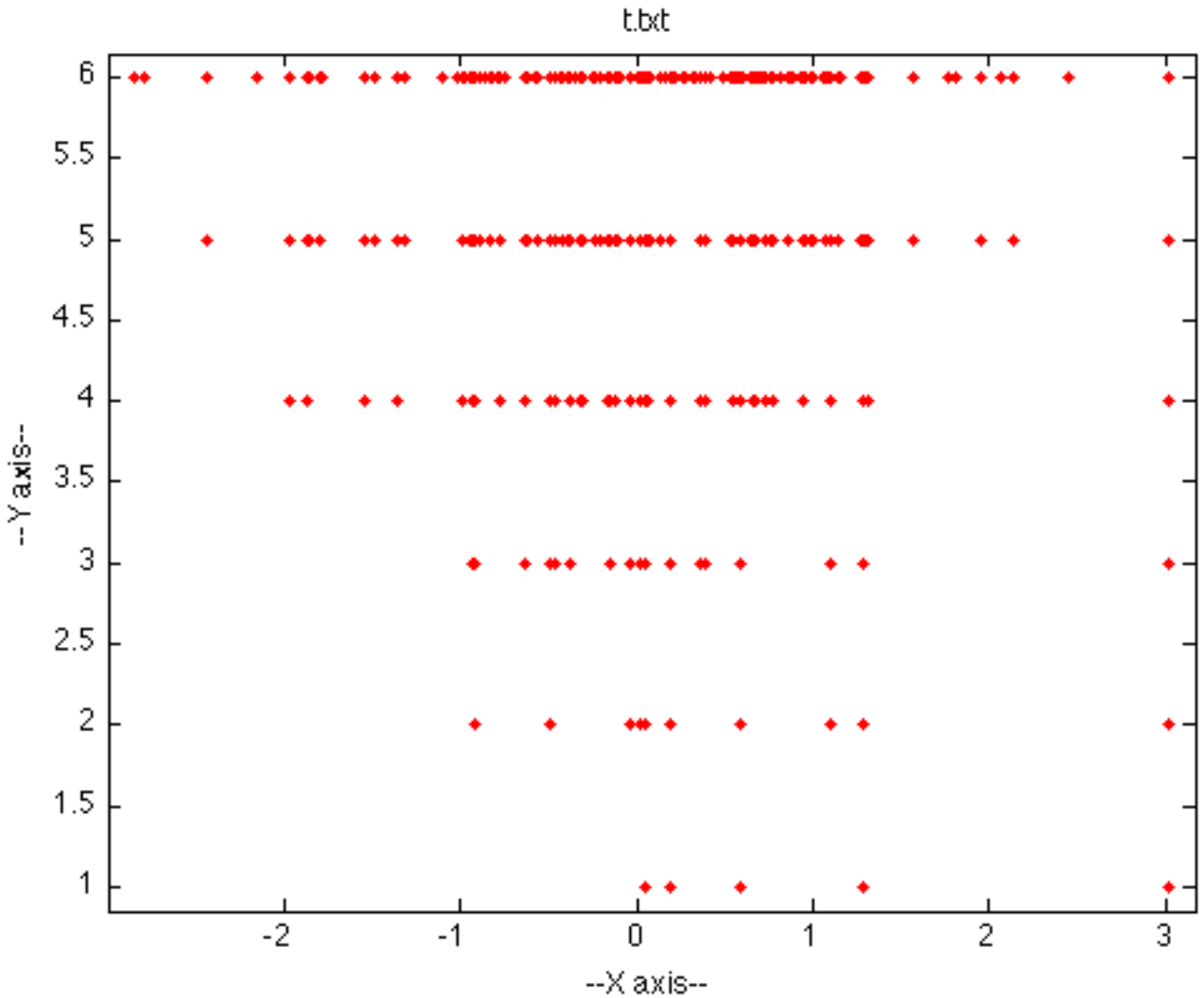
r = rand (n,1)



niaTech

Continuous: Normal Samples

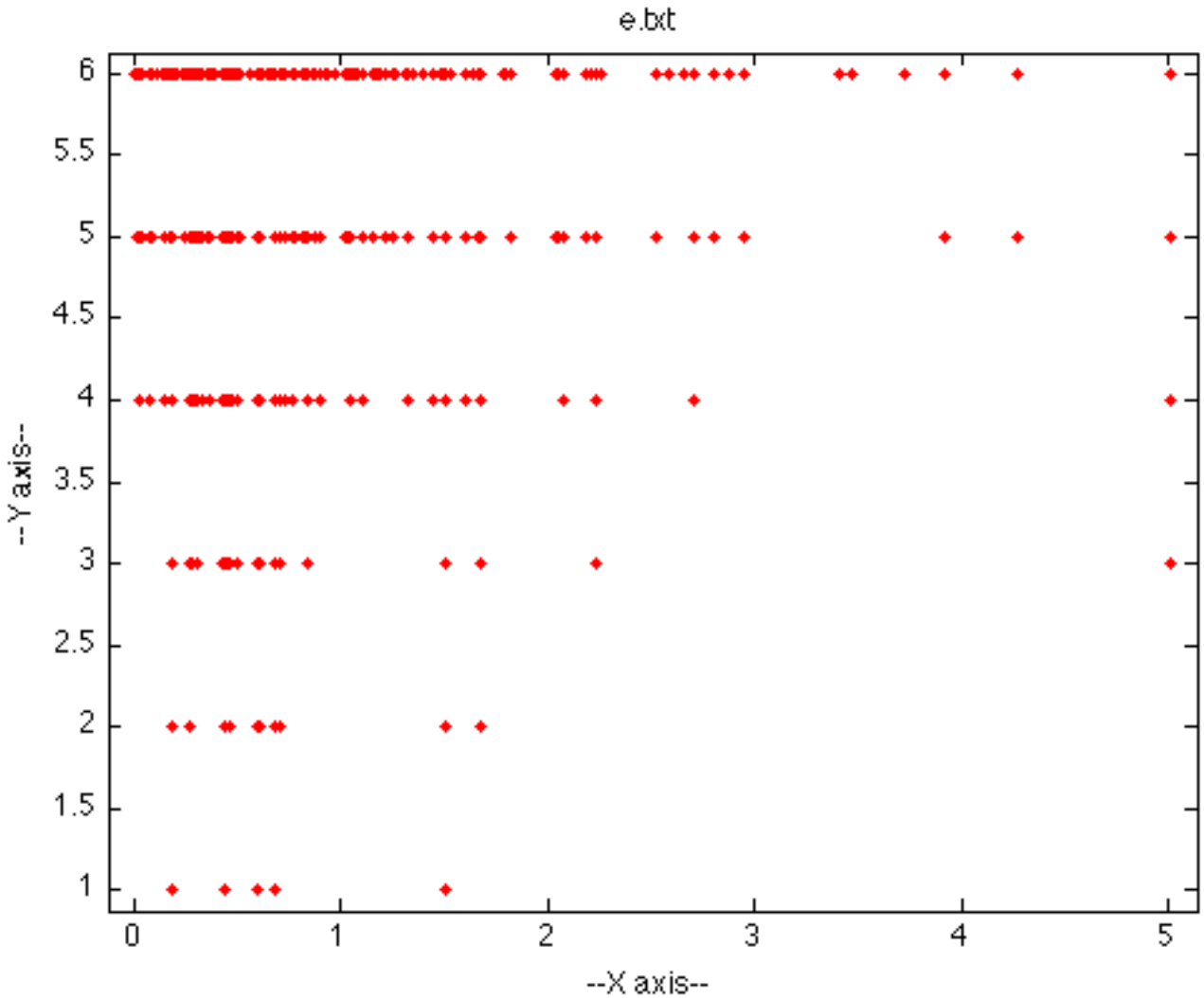
n = [5,10,20,40,80,160]
r = randn (n,1)



niaTech

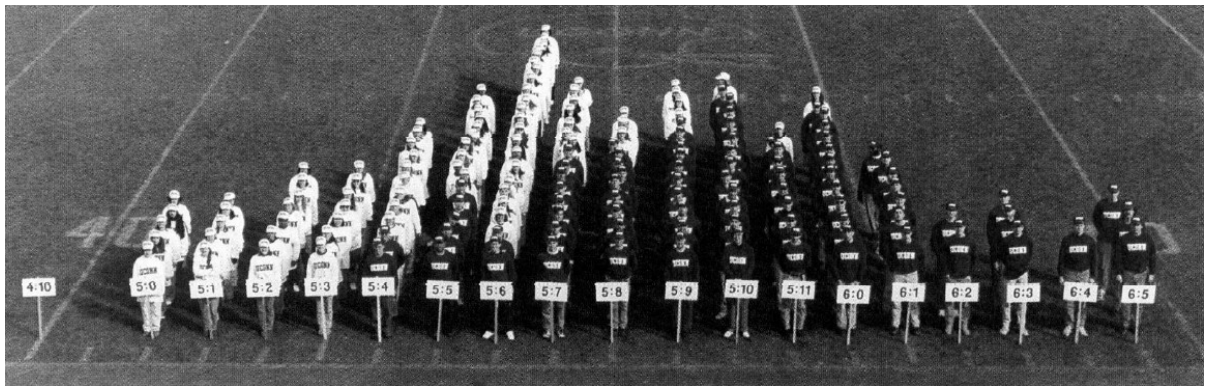
Continuous: Exponential Samples

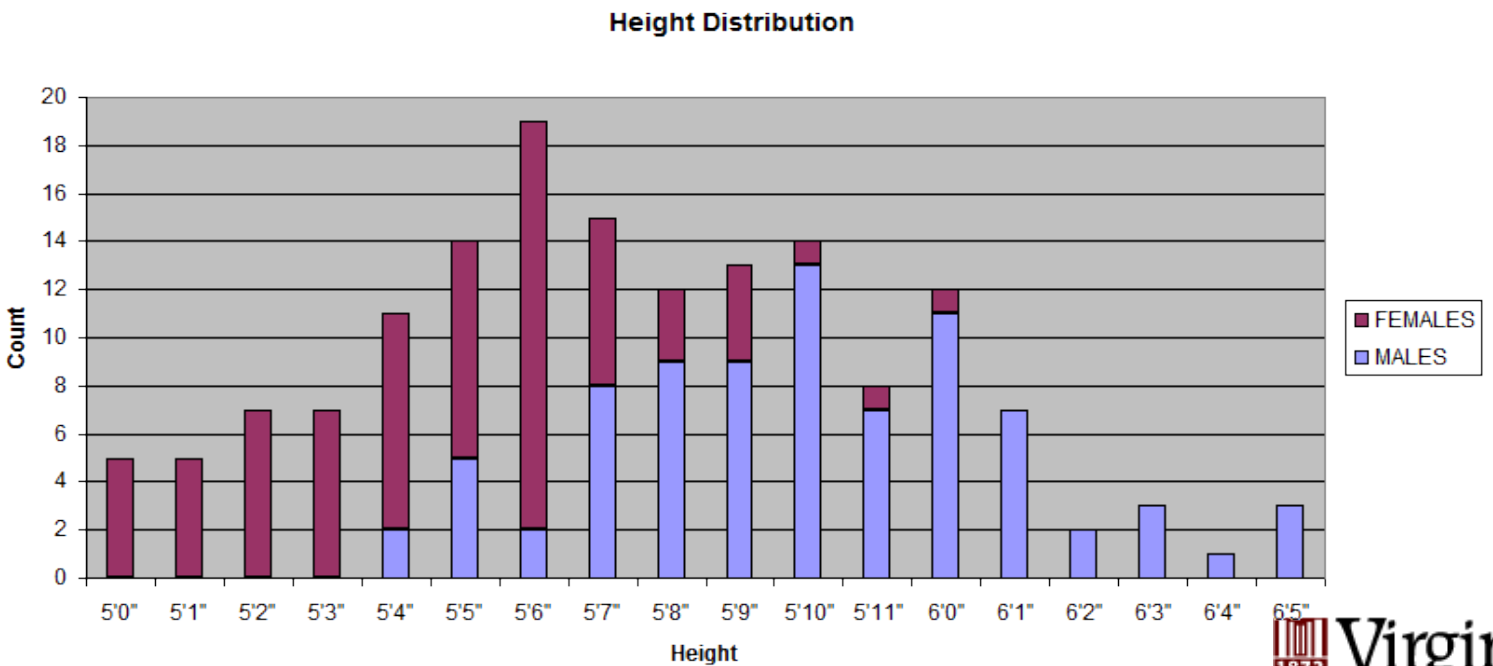
n = [5,10,20,40,80,160]
r = -log(rand (n,1));



niaTech

Continuous: Continuous Data is Histogrammed





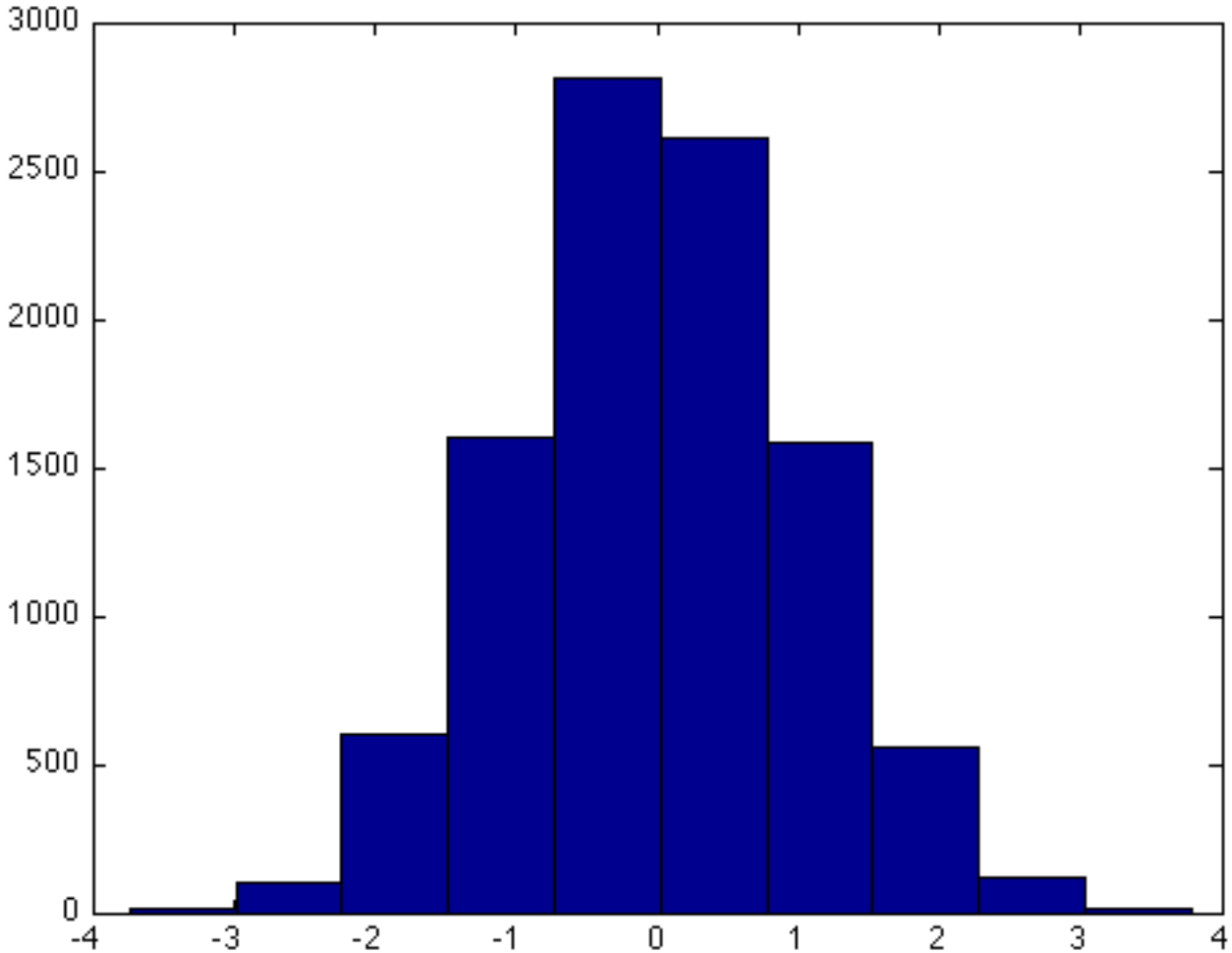
VirginiaTech

Continuous Probability

Although the outcomes are continuous variables, we only see finitely many of them. "Plotting" their values produces a string of dots whose density suggests where the most likely outcomes are. (Strictly speaking, a continuous variable has a probability density, not a probability!)

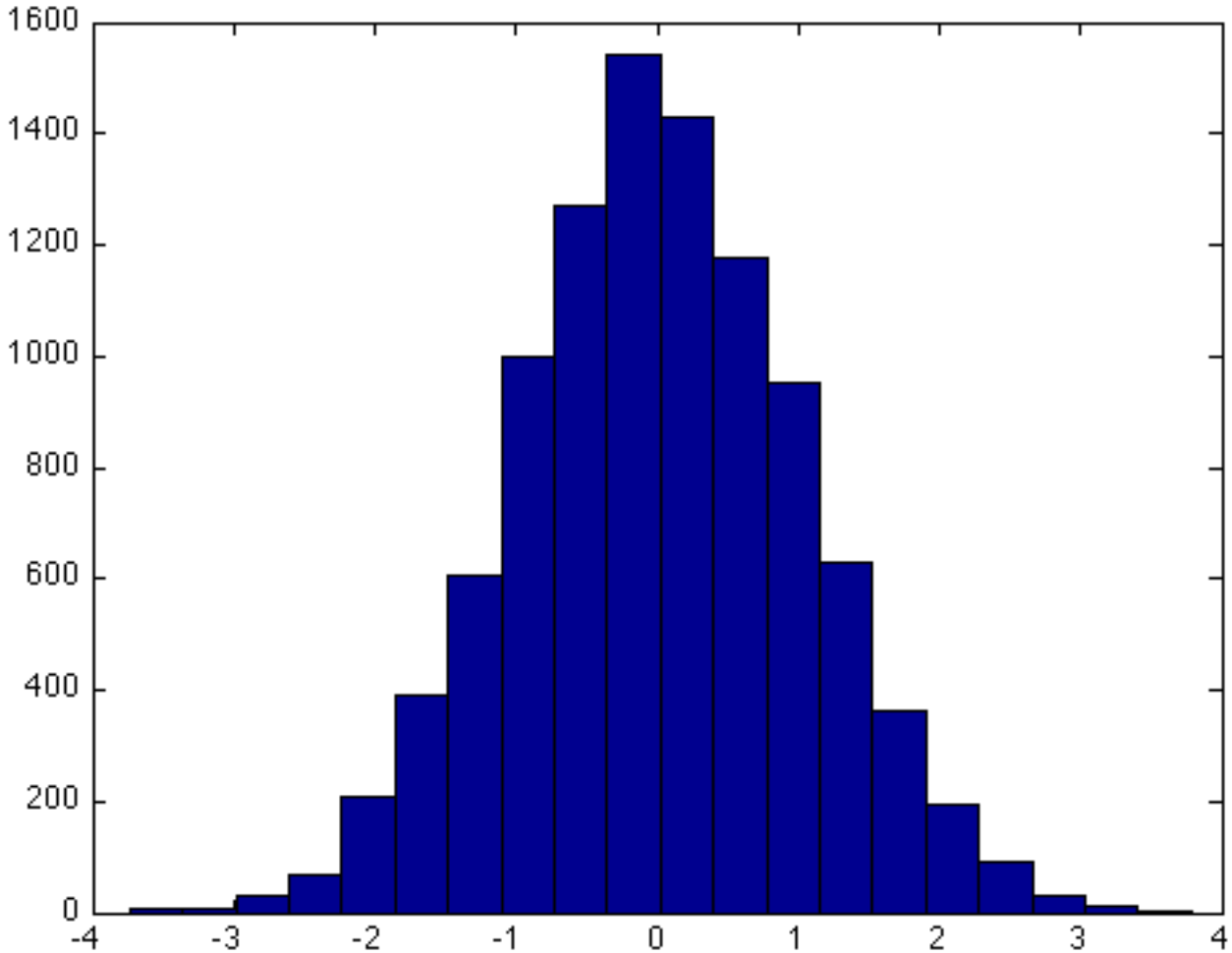
If we have sample outcomes, and want to get a feeling for the underlying continuous probability density function, we have to estimate the density. We can do that by grouping the outcomes into bins of equal width. This is called a histogram.

Continuous: 10 Bins of 10,000 Normal Samples



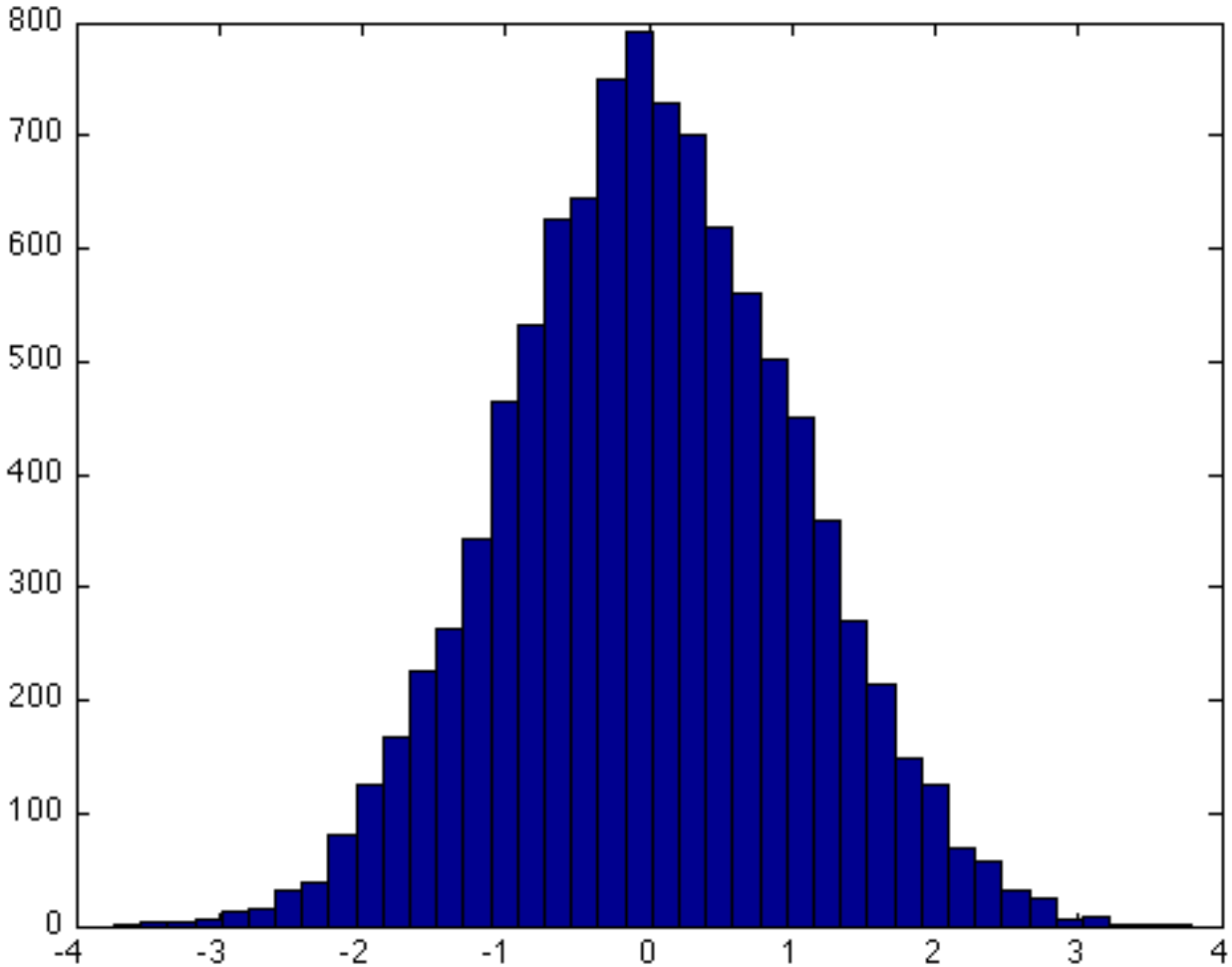
VirginiaTech

Continuous: 20 Bins of 10,000 Normal Samples



VirginiaTech

Continuous: 40 Bins of 10,000 Normal Samples



VirginiaTech

Continuous Probability: The Histograms Suggest a Smooth Function

Just like we did in the discrete case, we start to get a feeling that there's a mathematical function that's trying to express itself here.

As we increase the number of bins, we think we see a smooth curve emerging. At some point, we don't have enough data to keep doing that. However, we now have faith that there's something called a probability density function for continuous variables, and that if we don't know its form, we can estimate it, or at least approximate a graph of it, by using histograms.

This is similar to the way we used frequency counts in the discrete case to estimate the probability function.

Continuous Probability

To begin with, let's suppose that for our continuous problems, the set of outcomes **X** that we are interested in is an interval $[a, b]$, a half-infinite interval like $[a, +\infty)$, or the whole real line $(-\infty, +\infty)$, but in any case, some connected subset of $\mathbb{R}$.

A PDF function $p(x)$ for such a set of outcomes must have the basic properties that

1. $p(x)$ is defined for all x in X;
2. $0 \leq p(x)$ for all x in X;
3. $\int_X p(x)\, dx = 1$.

in other words: *defined, nonnegative* and *positive*.

VirginiaTech

Continuous Probability

We need to be more careful about interpreting a PDF graph for the continuous case.

The height suggests that some values are more likely than others. If $p(x)$ is twice the height of $p(y)$, then x is twice as likely as y.

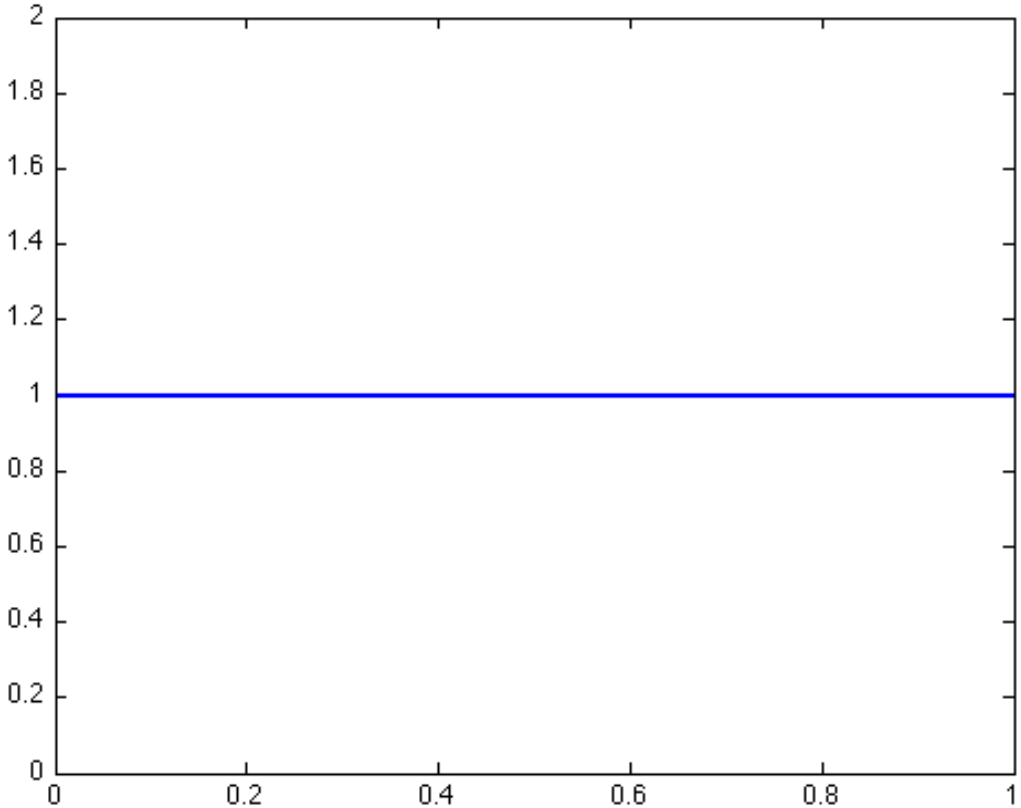
But the actual value of $p(x)$ does not represent a probability. The closest thing to a probability is the expression for the infinitesimal probability $p(x)\,dx$. We only get a nonzero value when we integrate these infinitesimal probabilities over some interval:

$$\text{probability}(5 \leq x \leq 7) = \int_5^7 p(x)\,dx$$

VirginiaTech

Continuous Probability: Uniform PDF

The set of outcomes X is the interval $0 \leq x \leq 1$.
The PDF function is $p(x) = 1$;
The average is $\frac{1}{2}$;
The variance is $\frac{1}{12}$.



VirginiaTech

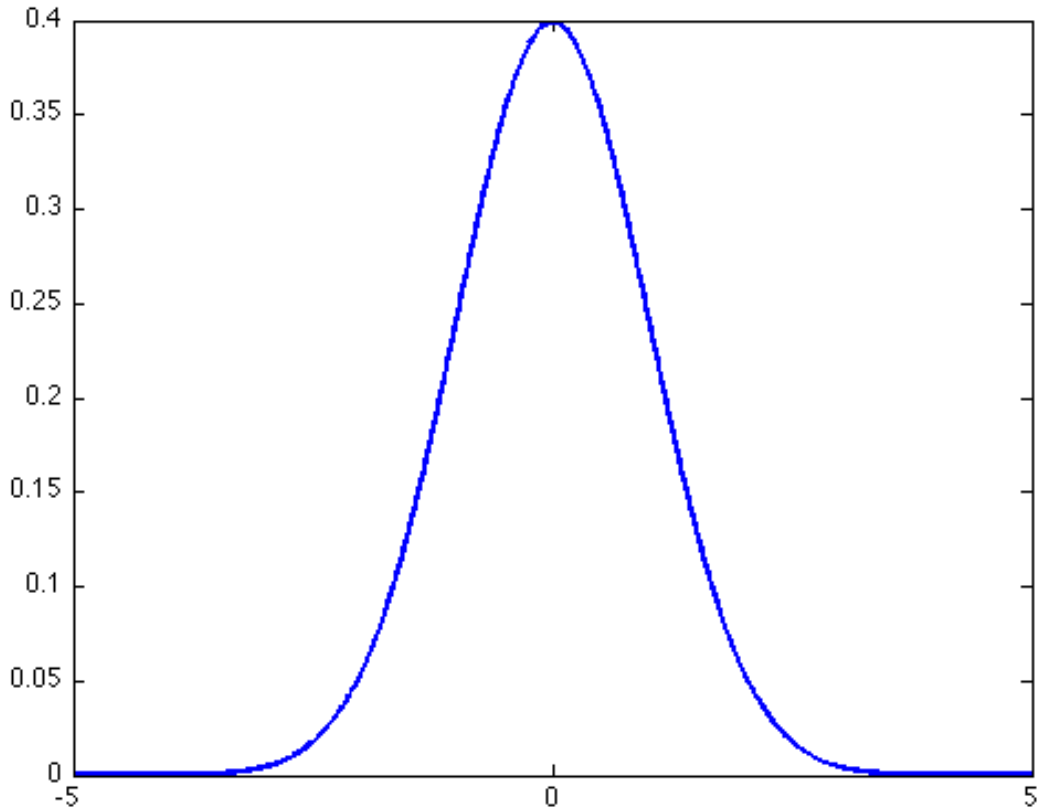
Continuous Probability: Normal PDF

The set of outcomes X is $-\infty < x < \infty$.

The PDF function is $p(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$;

The average is 0;

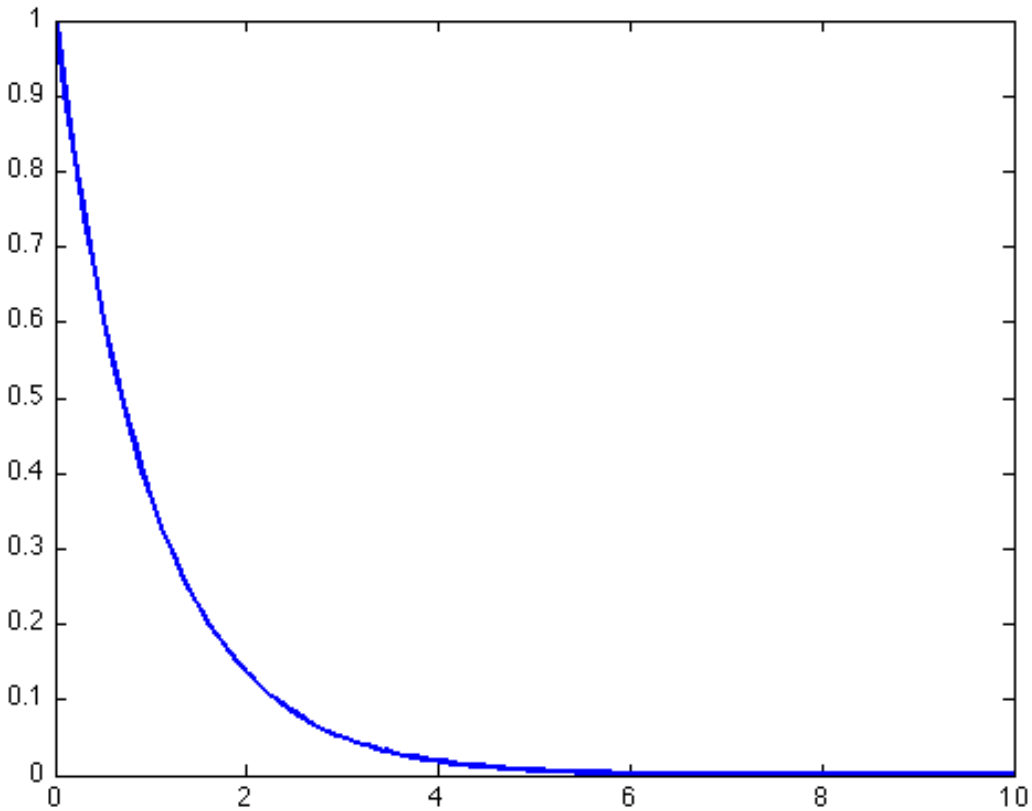
The variance is 1.



VirginiaTech

Continuous Probability: Exponential PDF

The set of outcomes X is $0 \leq x < \infty$.
The PDF function is $p(x) = e^{-x}$;
The average is 1;
The variance is 1.



VirginiaTech

Continuous Probability: Mean and Variance

Just as for the discrete case, the PDF for a continuous variable has a *mean* and *variance*. However, now we must replace sums by integrals!

$$\text{ave} = \int_a^b p(x) * x \, dx$$

$$\text{var} = \int_a^b p(x) * (x - \text{ave})^2 \, dx$$

Here **a** and **b** are the left and right intervals of the variable, either or both of which might be infinite.

VirginiaTech

Continuous Probability: Mean and Variance

To compute the mean value for the uniform PDF:

$$\begin{aligned}\text{ave} &= \int_a^b p(x) * x \, dx \\ &= \int_0^1 1 * x \, dx \\ &= \frac{x^2}{2}\Big|_0^1 \\ &= \frac{1}{2}\end{aligned}$$

You can compute the variance for the uniform PDF!
You can compute the mean value for the exponential PDF!

VirginiaTech

Continuous Probability: Expected Value

Just as for the discrete case, if there is a payoff or some other value $v(x)$ associated with each outcome x, then we can compute the **expected value**, that is, the average payoff.

$$E(v(x)) = \int_a^b p(x) * v(x)\, dx$$

Note that the average or mean is simply the expected value of x itself.

Continuous Probability: the CDF

We said that the PDF can answer questions such as: *what is the probability that* **x** *is between 5 and 7?*:

$$\text{probability}(5 <= x <= 7) = \int_5^7 \text{pdf}(x)\, dx$$

This idea is so important that a new function is defined, the **CDF** or *cumulative density function*. It answers the question *what is the probability that the variable is less than or equal to x?*:

$$\text{cdf}(x) = \int_a^x \text{pdf}(t)\, dt$$

Here **a** is the lower limit of the range of the variable.

VirginiaTech

Continuous Probability: the CDF

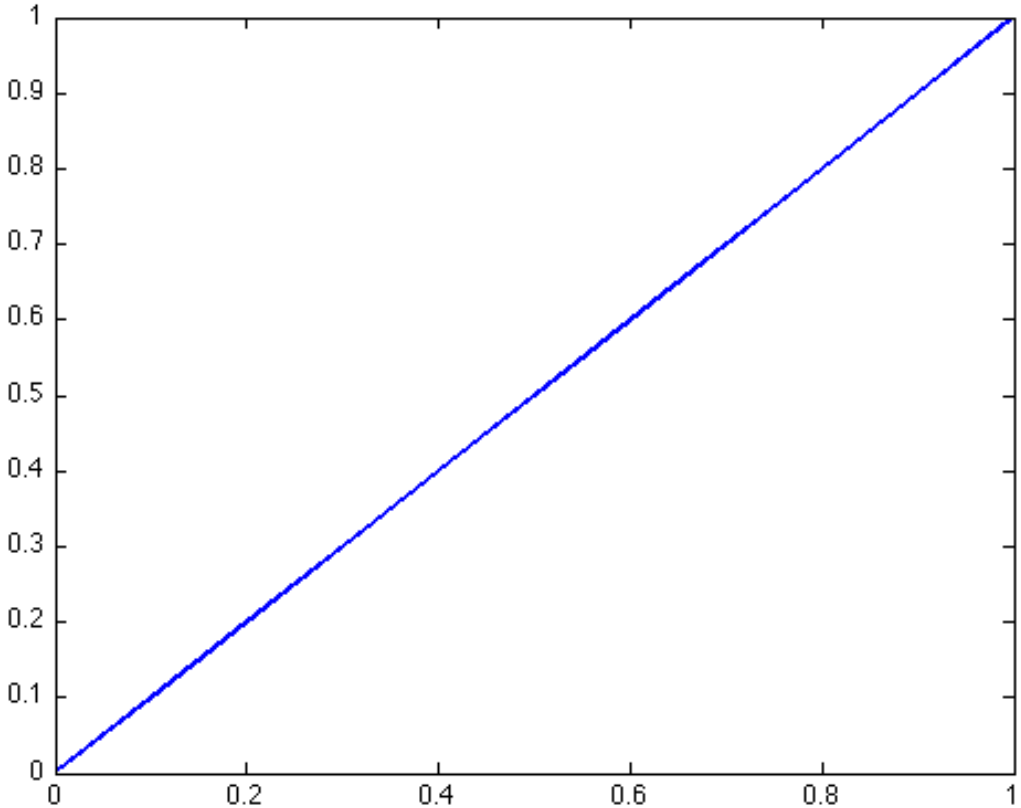
Facts about the CDF for a continuous variable x:

- **cdf(x)** is differentiable (assuming **pdf(x)** is continuous!)
- $\frac{d}{dx}\text{cdf}(x) = \text{pdf}(x)$, (assuming **pdf(x)** is continuous);
- **cdf(x)** is monotone increasing (because $0 \leq \text{pdf}(x)$);
- **cdf(x)** is almost always invertible;
- **cdf(a)** $= 0$ and **cdf(b)** $= 1$. But if either endpoint is infinite, then this statement is actually a statement about limits.

VirginiaTech

Continuous Probability: Uniform CDF

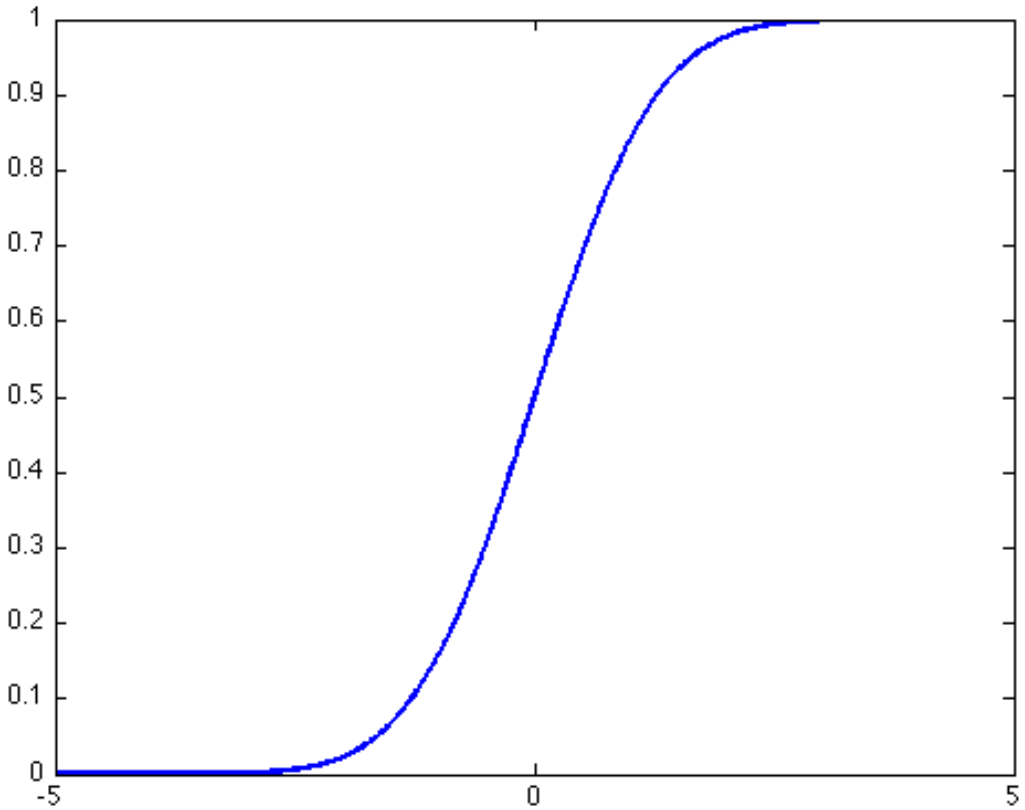
The uniform CDF is $\mathrm{cdf}(x) = x$



VirginiaTech

Continuous Probability: Normal CDF

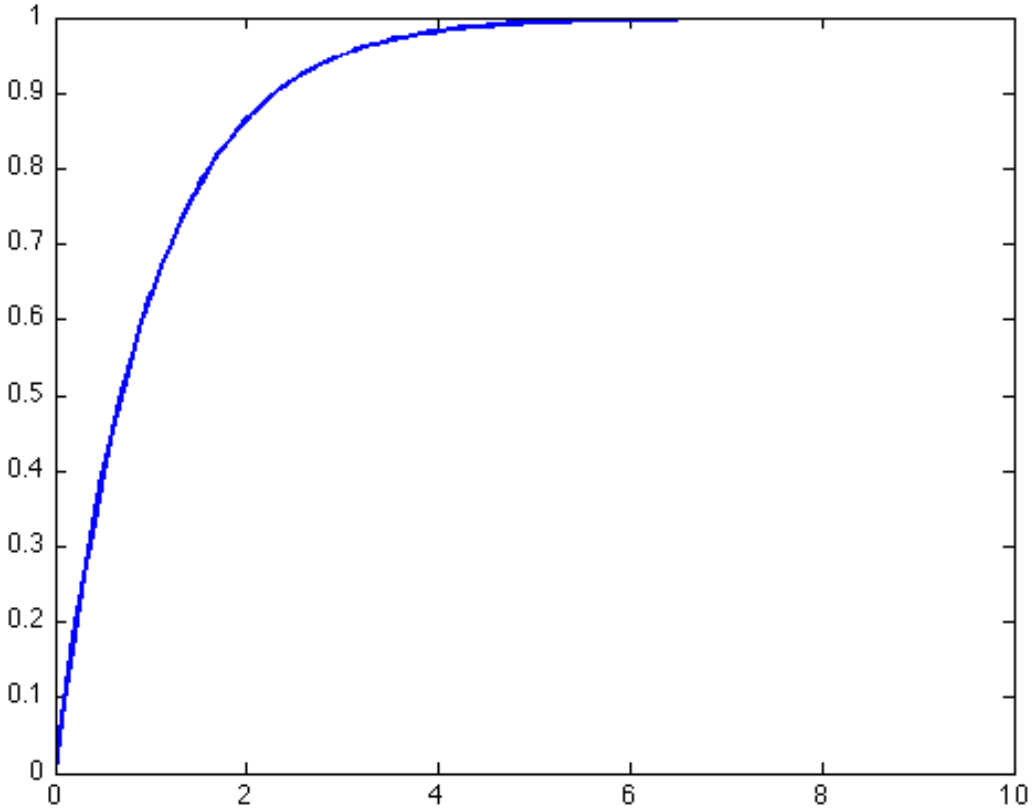
The normal CDF is $\mathrm{cdf}(x) = \frac{1}{2}(1 + \mathrm{erf}(\frac{x}{\sqrt{2}}))$;



VirginiaTech

Continuous Probability: Exponential CDF

The exponential CDF is $\mathrm{cdf}(x) = 1 - e^{-x}$



VirginiaTech

Monte Carlo Method: Probability

- Overview
- Discrete Probability
- Continuous Probability
- **Fitting Distributions to Data**

VirginiaTech

Fitting Distributions to Data

The formula I gave you for the normal PDF and CDF is for the "standard" functions, in which the mean is taken to be 0 and the variance is 1.

The mean is often symbolized by μ. People usually work with the square root of the variance, called the standard deviation, and symbolize this by σ.

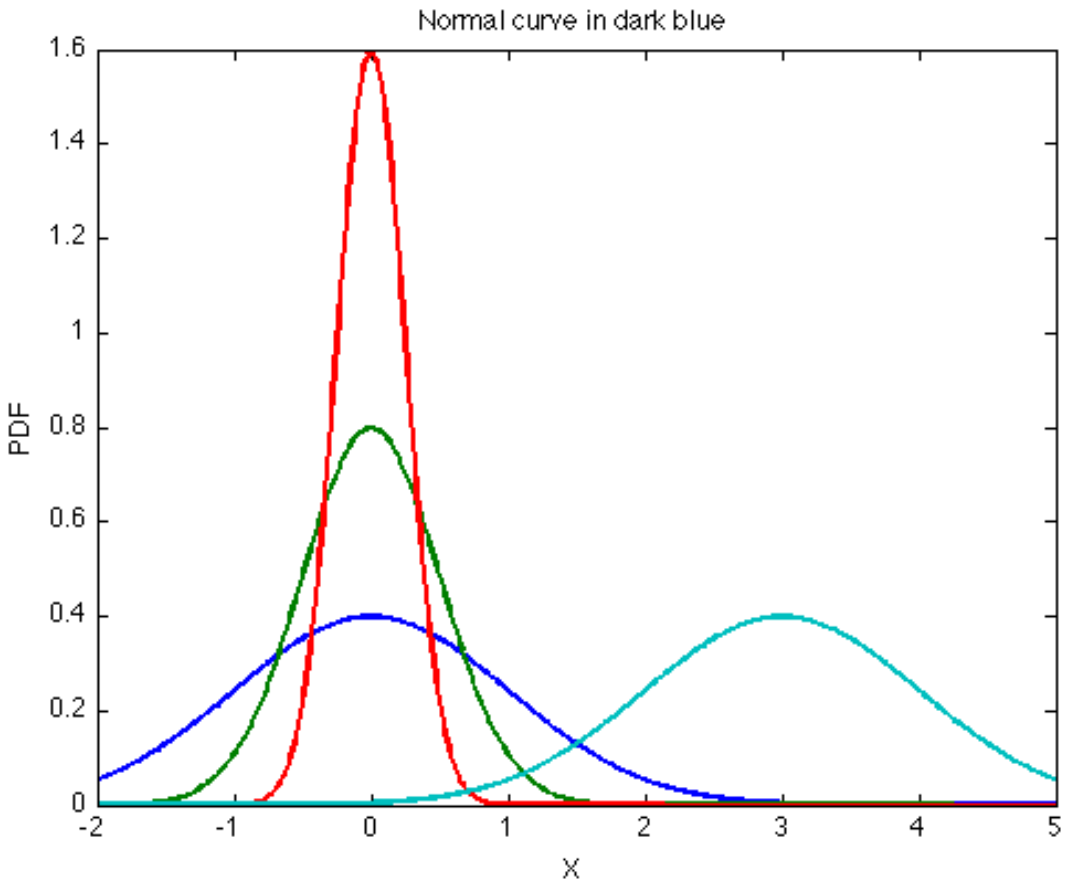
By changing the mean and deviation, a family of curves can be generated that are similar to the normal curves. Often, simply by choosing the right values of these two parameters, the normal curve can come reasonably close to matching statistical data.

Seeking parameters which minimize the difference is called *fitting a curve to the data.*

VirginiaTech

Fitting Distributions to Data

MATLAB: y = pdf ('normal', x, mean, std)



VirginiaTech

Fitting Distributions to Data

The family of normal curves has the following formulas:

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

$$\mathrm{cdf}(x, \mu, \sigma) = \frac{1}{2}(1 + \mathrm{erf}(\frac{x-\mu}{\sigma\sqrt{2}}))$$

If we find some data that might be approximated by some version of the normal curve, then we have two parameters to play around with.

VirginiaTech

Fitting Distributions to Data

Here is data on height measurements **H** in inches for women between the ages of 20 and 29. Instead of listing the exact height of each woman, the data has been summarized as *percentiles* labeled **C**.

Luckily, percentiles are simply another way of describing a CDF.

C	5	10	15	25	50	75	85	90	95
H	59.9	60.5	61.4	62.3	64.1	65.8	67.0	67.5	68.4

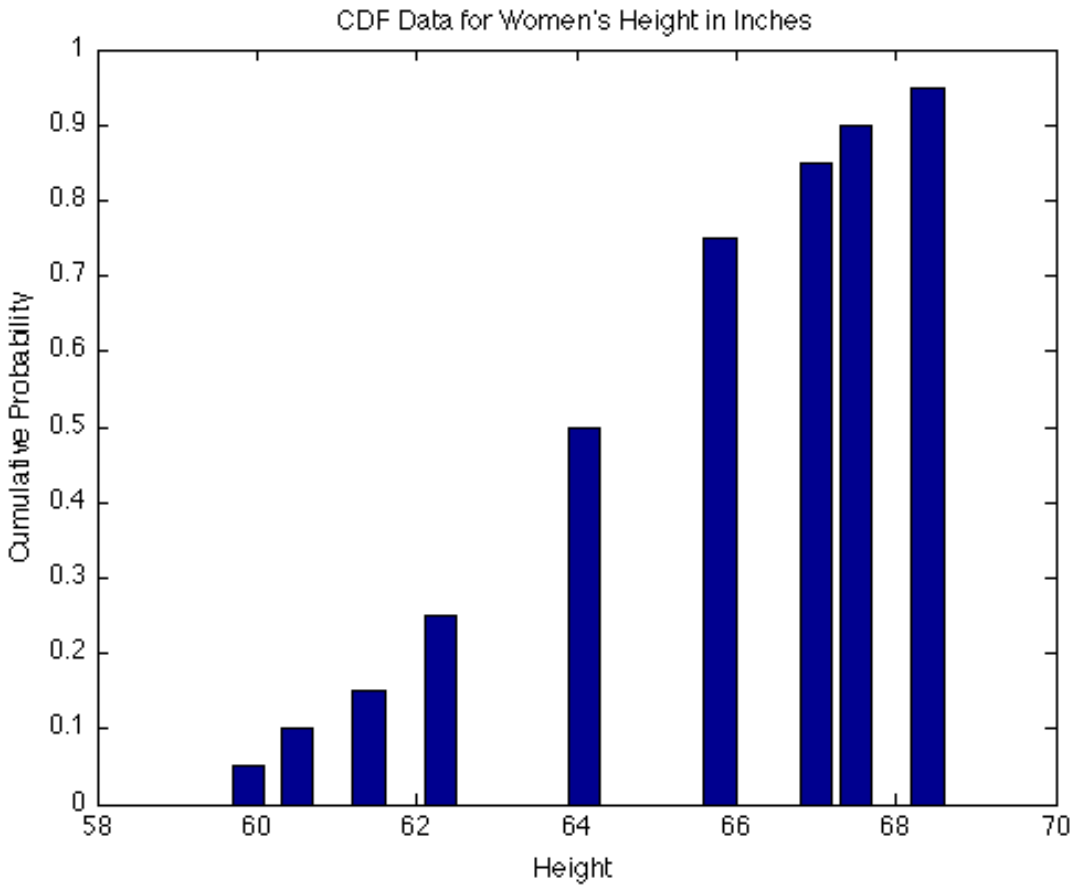
Since we are used to dealing with probabilities, each percentile value **C** should simply be divided by 100.

VirginiaTech

Fitting Distributions to Data

Remember that **H** is our independent variable for this CDF!

bar (h, c)



VirginiaTech

Fitting Distributions to Data

The data in this form has very limited usefulness. We'd like to be able to come up with a model for the data, a formula that would summarize and perhaps explain what is going on, and one that would let us make plots and so forth.

We will start out by assuming that this data can be described by a normal curve. If that were exactly true, then we would only have to guess the right values of μ and σ and we would have our formula.

But we will be satisfied if we can find a normal curve that fits the data well.

VirginiaTech

Fitting Distributions to Data

MATLAB has a function called **lsqnonlin** which allows us to describe a set of **(x,y)** data, and a function **y = f(x)** which includes parameters (say **a** and **b**).

Then **lsqnonlin** tries to find values of **a** and **b** for which the resulting function is closest to the data. (It does this by minimizing the sum of the squares of the errors.)

You may be familiar with this concept when you tried to fit (x,y) data by a straight line **y=ax+b**. And now we are going to doing something similar, except we will be fitting a normal curve, so our **a** and **b** will be the mean and standard deviation.

VirginiaTech

Fitting Distributions to Data

Unfortunately, we can't fit the PDF function to our data. We don't have probabilities, we have CDF data. That means that we are looking for values μ and σ so that our data is matched as closely as possible by the CDF formula:

$$c(h) = \frac{1}{2}(1 + \mathrm{erf}(\frac{h - \mu}{\sigma\sqrt{2}}))$$

This is the function we are trying to fit. But if we can do the fitting, then we find out the values of μ and σ and we automatically get a formula for the PDF, for instance.

VirginiaTech

Fitting: The Difference Routine

```
function diff = women_fit ( param )
%
%  The parameter values are input in PARAM.
%
  mu = param(1);
  sigma = param(2);
%
%  Here is the data we are trying to match.
%
  cdata = [ 0.05, 0.10, 0.15, 0.25, 0.50, 0.75, 0.85, 0.90, 0.95];
  hdata = [ 59.9, 60.5, 61.4, 62.3, 64.1, 65.8, 67.0, 67.5, 68.4];
%
%  We evaluate our formula at the HDATA values.
%
  cformula = 0.5 + 0.5 * erf ( ( hdata - mu ) / sigma / sqrt ( 2 ) );
%
%  We compare our CDATA values to the output of the formula.
%
  diff = cdata - cformula;

  return
end
```

VirginiaTech

Fitting: The Main Routine

```
%
%  Initial parameter estimate.
%
  param(1:2) = [ 55, 1 ];

  param_new = lsqnonlin ( @women_fit, param );

  mu = param_new(1)
  sigma = param_new(2)

  cdata = [ 0.05, 0.10, 0.15, 0.25, 0.50, 0.75, 0.85, 0.90, 0.95];
  hdata = [ 59.9, 60.5, 61.4, 62.3, 64.1, 65.8, 67.0, 67.5, 68.4];

  h = 58 : 0.25 : 70;
  c = 0.5 + 0.5 * erf ( ( h - mu ) / sigma / sqrt ( 2 ) );

  plot ( h, c, 'LineWidth', 2 );
  hold on
  bar ( hdata, cdata );

  title ( 'Compare CDF data and fitted formula' )
  xlabel ( 'Height' )
  ylabel ( 'CDF' )

  hold off

  p = exp ( - ( h - mu ).^2 / 2 / sigma^2 ) / sqrt ( 2 * pi * sigma^2 );
  plot ( h, p, 'LineWidth', 2 );
  title ( 'PDF using fitted MU and SIGMA' );
  xlabel ( 'Height' )
  ylabel ( 'PDF' )
```

VirginiaTech

Fitting: The Results

Our little program returns quickly with the computed answers:

param(1) $= \mu =$ mean $= 64.0925$
param(2) $= \sigma =$ std $= 2.6609$

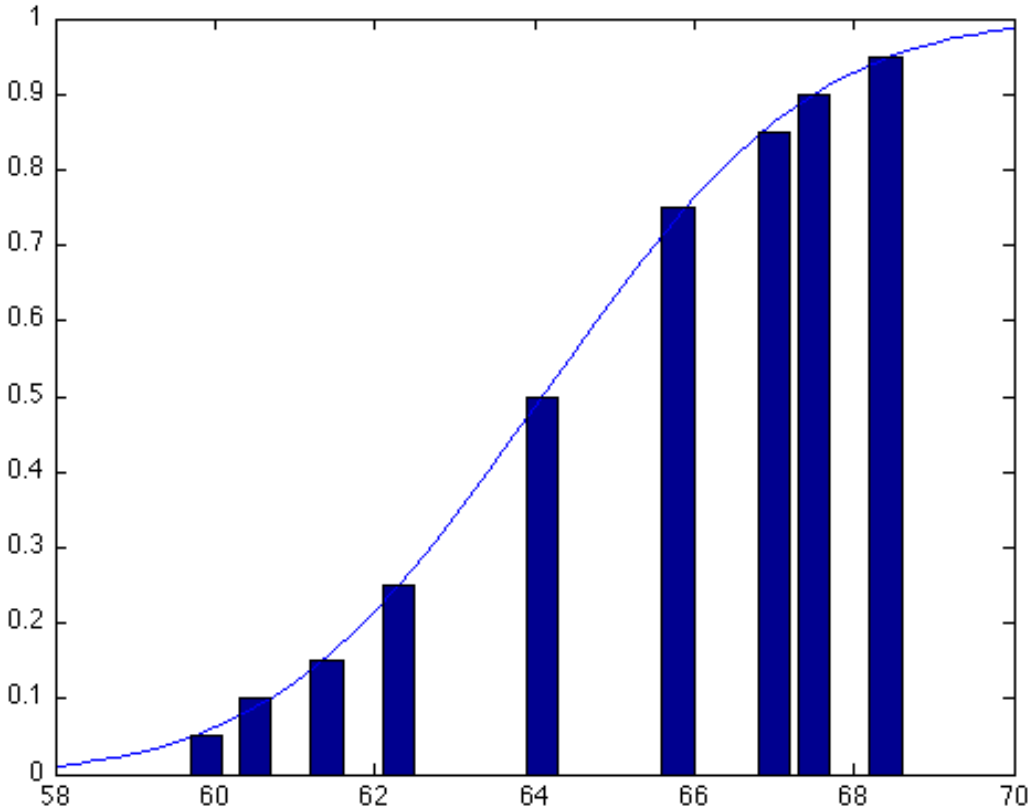
It's no surprise that the mean is 64.0925; after all, our original data actually included the 50th percentile value of 64.1. The 50th percentile is only the median value, but if we assuming heights are roughly symmetric about that point, then it is also the mean.

The standard deviation tells us that, if this is close to a normal distribution, then roughly 68% of the data will lie within one standard deviation of the mean.

VirginiaTech

Fitting Distributions to Data

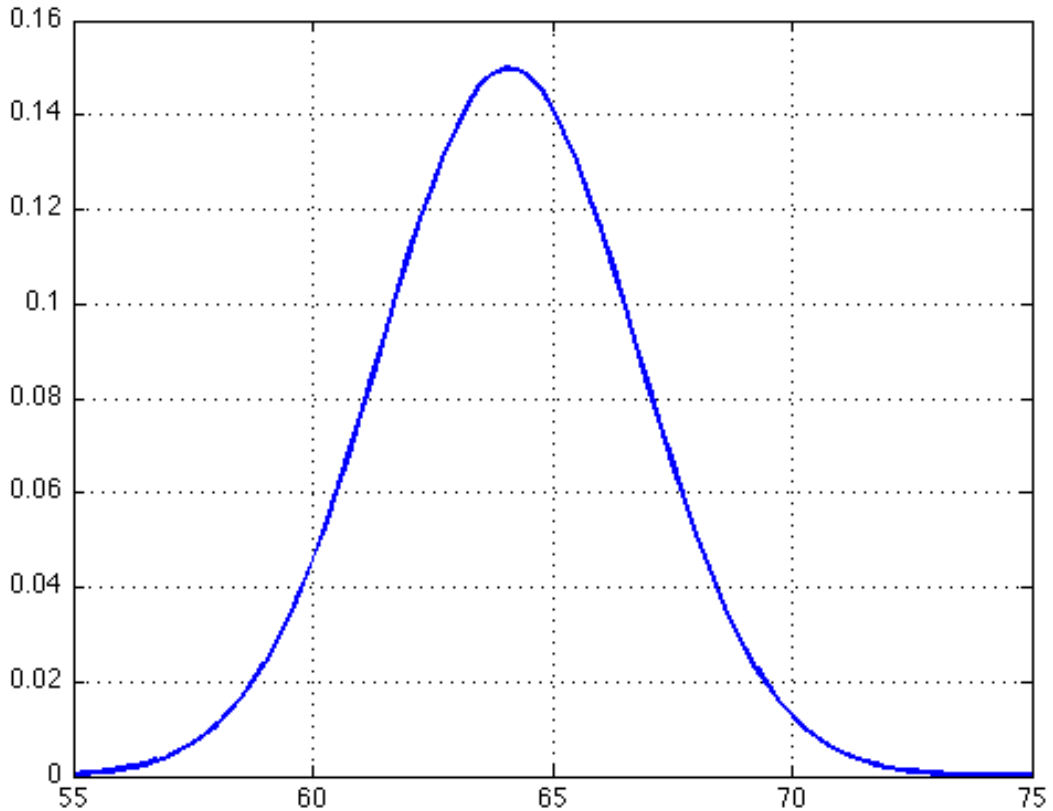
Here we compare our initial data to the CDF curve:



VirginiaTech

Fitting Distributions to Data

We can use the computed μ and σ to plot the PDF:



VirginiaTech

Fitting Distributions to Data: NOT an Assignment

Here is data on height measurements **H** in inches for men between the ages of 20 and 29. The data has been summarized as *percentiles* labeled **C**.

C	5	10	15	25	50	75	85	90	95
H	64.1	65.7	66.5	67.5	69.6	71.4	72.6	73.5	74.6

Can you remember the steps that were involved in going from this raw data to an estimate for the mean and standard deviation of the normal PDF that approximates this data?

What values of μ and σ do you get?

VirginiaTech