**Probability and Statistics.**

*Note: These topics are specific to the MATLAB Statistics toolbox.*

**Introduction**

To those not familiar with probability and statistics, their first introduction to these topics can be very “unnerving” for because statistics seems to be more related to “magic” than “mathematics”.

As engineers, we are well versed with handling “definite information”. We are comfortable with determing the value of y when

 y = sin2(x)/exp(sin(x)) when x=√2

or determining the value of z when

 

Even though we might need to use a calculator to “solve for y” or use MATLAB to perform the integration to determine the “z” at least we believe there is a right way to “get” the answer.

But when we move into questions involving probability and statistics, it seems as if our “beloved tools” have been taken away and we have the disquieting feeling that the questions asked might not be questions at all!

For example (from my own experience in my undergraduate statistics course I was asked to solve the following problem on an exam:

*You go to the cafeteria to have some fries and you notice that in the “dunk cup” of ketchup you were provided with there are 4 fragments of foreign matter “bug parts” in the one-ounce serving. These might be due to tiny fragments of wings, legs, etc. that were introduced in the spices since “bugs” live in the plants naturally and cannot be economically removed) unless you want to pay $50 for a bottle of ketchup.) To continue, you know there is a specification that the supplier of the ketchup claims is being met (50 parts of foreign matter per 16 oz bottle). Do you have reason to believe that the manufacture is not supplying ketchup meeting the specification?*

I guess I missed the lecture dealing with these statistics concepts and since our professors often asked questions that were either open-ended or had insufficient information provided and in these cases it was “ok” to point out the situation in a humorous fashion. It seemed the question did not have sufficient information for a specific answer so I took the problem to be “a joke.” My exam answer was “there is nothing to be concerned with because bug parts are very nutritious and you’re getting extra protein”. Alas, my sense of humor was lost on the professor (because there IS a specific answer) and I got a “zero” on the problem.

Consider another example, “Suppose that on weekdays the average number of cars passing through Tomer’s Corner at 5:00 AM is 22 cars per 15 minute interval. What fraction of the weekdays in the next month can one expect to see more than 27 cars per 15 minutes at 5:00 AM? Again, how (except by magic) can one possibly hope to provide an answer when so little information is provided?

To those who understand probability and statistics concepts, these are very straightforward and easy problems. To those lacking that knowledge, they hardly seem to be answerable.

In addition, there is a “whole language” that is employed in probability and statistics. This language is “precise” where each term means something very specific whereas the same terms in “everyday usage” may not have precise definitions.

# Basic Definitions and Concepts

Probability is an area of study which involves predicting the relative likelihood of various outcomes. It is a mathematical area which has developed over the past three or four centuries. One of the early uses was to calculate the odds of various gambling games. Its usefulness for describing errors of scientific and engineering measurements was soon realized. Engineers study probability for its many practical uses, ranging from quality control and quality assurance to communication theory in electrical engineering. Engineering measurements are often analyzed using statistics and a good knowledge of probability is needed in order to understand statistics.

Statistics is a word with a variety of meanings. To the man in the street it most often means simply a collection of numbers, such as the number of people living in a country or city, a stock exchange index, or the rate of inflation. These all come under the heading of descriptive statistics, in which items are counted or measured and the results are combined in various ways to give useful results. Descriptive statistics also includes presentations of data in tables and charts for visualization. This type of statistics has many uses in engineering and you are probably already aware of descriptive statistics such as the mean, range, minimum, standard deviation, etc.

Another type of statistics will engage our attention to a much greater extent. That is inferential statistics or statistical inference. For example, it is often not practical to measure all the items produced by a process. Instead, we very frequently take a sample and measure the relevant quantity on each member of the sample. We infer something about all the items of interest from our knowledge of the sample. A particular characteristic of all the items we are interested in constitutes a population. Measurements of the diameter of all possible bolts as they come off a production process would make up a particular population. A sample is a chosen part of the population in question, say the measured diameters of twelve bolts chosen to be representative of all the bolts made under certain

conditions. We need to know how reliable is the information inferred about the population on the basis of our measurements of the sample. Usually samples are selected in a random fashion, an idea that is simple in definition but may be difficult in practice.

Example: Suppose we are interested in the average height of male students at Auburn (our population). We wish to measure a number of students (our sample) to make inferences about the true average height without having to resort to measuring all male students. Suppose you were to randomly select 100 students. Where would you go on campus? How would you select individuals? Are engineering students, students on a basketball scholarship and art students good choices? Should you select a certain number of freshman, sophomore, graduate students, etc.? Should you do it on a Monday morning or a Saturday evening?

Chance (randomness) is a necessary part of any process to be described by probability or statistics. Sometimes that element of chance is due partly or even perhaps entirely to our lack of knowledge of the details of the process. For example, if we had complete knowledge of the composition of every part of the raw materials used to make bolts, and of the physical processes and conditions in their manufacture, in principle we could predict the diameter of each bolt. But in practice we generally lack that complete knowledge, so the diameter of the next bolt to be produced is an unknown quantity described by a random variation. Under these conditions the distribution of diameters can be described by probability and statistics. If we want to improve the quality of those bolts and to make them more uniform, we will have to look into the causes of the variation and make changes in the raw materials or the production process. But even after that, there will very likely be a random variation in diameter that can be described statistically.

Relations which involve chance are called probabilistic or stochastic relations. These are contrasted with deterministic relations, in which there is no element of chance. For example, Ohm’s Law and Newton’s Second Law involve no element of chance, so they are deterministic. However, measurements based on either of these laws do involve elements of chance, so relations between the measured quantities are probabilistic. As mentioned initially, most subject matter discussed in your courses up to this point have been deterministic and not stochastic.

Another term which requires some discussion is randomness. A random action cannot be predicted and so is due to chance. A random sample is one in which every member of the population has an equal likelihood of appearing. Just which items appear in the sample is determined completely by chance. If some items are more likely to appear in the sample than others, then the sample is not random but rather possesses a bias. A random selection is also called an unbiased sample.

# Probability Distribution Function (pdf)

As engineers, we are very familiar with the idea of a function… something that has a dependent value corresponding to an independent value. For example, consider the Fourier Series,

$$y= f\left(x\right)=a\_{0}+\sum\_{n=1}^{\infty }\left(a\_{n}\cos(\frac{nπx}{L})+b\_{n}\sin(\frac{nπx}{L})\right)$$

or the function describing the volume of a pyramid,

$$V=\frac{1}{3}bh$$

These functions can be plotted and have certain characteristics (for example, the volume of a pyramid is a continuous function defined only for x>0.

In processes involving randomness where we are interested in modeling and determining the probability, we also have functions called probability distribution functions. In these cases, the independent variable is the outcome of interest (and the x-axis defines the values for which there are associated probability values which appear on the y-axis).

Any proposed function to model a physical process involving a random outcome must conform to three properties (axioms). Functions that do not adhere to these axioms cannot be probability density functions. *First, we consider the case of discrete processes (those with a countable number of distinct outcomes, such as noting the number of dots showing when two dies (a pair of dice) are rolled or when a coin is flipped or a roulette wheel is spun.*

# Probability Axioms

Axiom 1: 0 < P(A) < 1 for each event A in S (S=A, B, C, D, …)

Axiom 2: P(S) = 1

Axiom 3: P(A U B) = P(A) + P(B)

In words:

Axiom 1: “Every event (outcome) has a probability between 0 (impossible) and 1 (certainty)”

Axiom 2: “The sum of the probabilities of all possible events is 1 (certainty)”

Axiom 3: “If A and B are mutually exclusive events in S, their probabilities are additive”

To demonstrate these axioms: Place 5 blue marbles, 10 red marbles, and 5 white marbles in a bag. Choose 1. What are the probabilities in terms of a probability distribution function P.

P(B) = 5/20 = 25%, P(W) = 5/20 = 25%, P(R) = 10/20 = 50% Each satisfies Axiom 1

P(B | R | W) = 100% Axiom 2

P(B | R) = P(B) + P(R) = 75% Axiom 3. Since the events are mutually exclusive (the marble selected cannot be two colors at once) the probability of seeing a blue marble OR a red marble is the sum of the two probabilities.

So, we know if we find a claimed probability distribution function and it predicts negative probability values, or values greater than 1, it cannot model reality.

In continuous processes, we do not apply Axiom 2 and Axiom 3 to individual values but rather to ranges of values. A continuous process involves sampling spaces with a continuum of outcomes, such as the lengths of metal bars being cut by a milling machine or a sample’s weight or purity or activity, etc.

# Random Variables

A random variable “x” is a variable that can take on any ONE value within its range. THIS IS THE PRINCIPLE DIFFERENCE BETWEEN REGULAR MATH AND STATISTICS. In ordinary math we are used to a variable taking on a single specific value. In statistics, the variables take on an “unknowable” (random) value.

Discrete Random Variable

For example, you roll a die. The possible outcomes (range of values) are 1, 2, 3, 4, 5, or 6 (dots facing up). If we wanted to assign a variable to hold the result of tossing one die, we could call it “x”. The variable “x” can be ANY of the values 1, 2, 3, 4, 5, or 6. It cannot be 0, 8, or -3.

The probability of getting a particular outcome (a specific “x”) is the same for all of the outcomes because they are all equally likely in this case. The obvious value is 1/6.

This is written: x= 1..6 P(x) = 1/6 = 0.1666667

Note that our previous discussion of “plotting functions” (using a line to connect the artificially selected data values) needs to be modified for discrete functions. The plot below on the left is a proper depiction of a discrete functions pdf while the graph on the right is improper since it implies values between 1, 2 , 3, 4, etc. This would imply that there was a 16.667% chance of getting 3.5 dots or 2.244444 dots.

*See code segment 1*

Instructor’s Note: The code used to generate this plot (and other items in this chapter such as tables, figures, etc.) are displayed at the end of the chapter.



# Cumulative Distribution Function (cdf)

Related to probability distribution functions are cumulative distribution functions. In the case of a process having discrete outcomes, the cumulative distribution is the sum of the probabilities associated with seeing the outcome of x or less.

Unlike discrete distributions, the pdf of a continuous distribution at a value is not the probability of observing that value. For continuous distributions the probability of observing any particular value is zero1. To get meaning probabilities you must integrate the pdf over an interval of interest. Therefore, the cumulative distribution function for a continuous function is associated with the area under the probability distribution function for values of x or less. *1 The instructor will elaborate this point!*

**The Rolling of a Pair of Dice**

For a more complicated (discrete) situation, we employ the same basic concept for probability. Simply list the number of different “outcomes” and determine the number of different ways that outcome can be realized. For example, if we roll a pair of die and consider the sum of the dots to be the “random variable”, then the possible outcomes are 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12 dots. The probabilities associated with these outcomes are shown in table below. In the figure immediately following the table is a plot of the **pdf** and **cdf** for this situation.

|  |  |
| --- | --- |
| Sum of Dots Showing (Two Die) | pdf and cdf Determination |
|  | Column 1 is the number of dots showing.Column 2 is the number of ways the outcome can be achieved (of 36 total ways).Column 3 is the pdf.Column 4 is the cdf. |

*See code segment 2*



*See code segment 3*

# Discretization (Continuous Outcome Example)

Suppose that we know that male students at Auburn have an average weight of 150 lbs. with a standard deviation of 15 lbs and female students have an average weight of 120 lbs. with a standard deviation of 10 lbs. At a recent football game, one section of the stadium seated 120 male students and 80 female students. What will be the expected range of total weights of the students in the section (for design purposes)? Stated in more statistical language we wish to know the minimum and maximum weight (load) the section will have to bear a stated fraction of the times the stadium is used (say, 95 percent of the times).

*Let’s start by asking the question, “What is the probability that a male student weighs 146 lbs?”*

Remember that weight is a continuous variable, hence we cannot expect to ask questions like “what is the probability that a student weighs 146 lbs” even though it “sounds ok.” Referring to our previous comments about continuous distributions, the probability is “zero”. Of course, there are male students who will say that they are 6 ft tall who are about 6 ft tall, but if we considered all male students whose true height was within 0.1 in of 6 ft we would find all of them had unique (infinitely precise) heights. No two students alive have the same height (and that does not even take into account that your height is changing with each breath and with the time of day and with your continuing growth). One student’s true height (at a given instant) might be:

S1 = 6.0000000000000000000000000000000000000000010000000002300000000006…

while another student’s might be:

S2 = 6.0000000000000000000000000000000000000000010000000002300000000106…

You can see that there are more decimal values here than people on the planet!

That means, in many situations we must define a convention range of values that we will associate with a particular value.

Definition: A 6.0 ft tall person is between 5’ 11.5” and 6’ 0.5” (arbitrary)

Another definition: A 6.0 ft tall person is between 5.95’ and 6.05’ (arbitrary)

With either one of these definitions we can provide information (a non-zero probability value) about the fraction of individuals whose height is between the minimum and maximum height for this range.

This is important in MATLAB in that we will get misleading information if we forget about this characteristic of continuous distributions.

The following MATLAB code fragment asks “what is the probability of someone having “exactly” the weight of 100 to 200 lbs (some are “hidden” to conserve space). See the results in the table below.

%% pdf of a continuous function

clear; clc;

w=[100:2:114 146:2:154 194:2:200];

pdf=normpdf(w,150,15);

data=[w',pdf'];

>> normpdf(146,150,15)=0.025667. This does NOT mean there is a 2.566% chance that a person will weigh exactly 146 lbs. Instead it refers to a function (pdf) which when integrated over ALL possible weights will give the value 1.0 (100% certainty, everyone has a weight).

|  |
| --- |
| Column1 = wColumn2 = pdf(w,150,15) |
|  |



To correctly answer the question about the probability that someone weighs (about) 146 lbs we must establish an interval centered on 146 to which we will assign the weight 146 lbs. We can perform this with the following (which represents finding the area under the curve from 145.5 to 146.5 lbs:

w=[145.5 146.5];

pp=normcdf(w,150,15)

pp(2)-pp(1)

pp =

 0.38209 0.40775

ans =

 0.025663

In this case BY CONINCIDENCE, we have the same number nearly. But consider the implication of the first (erroneous) technique. Let’s compute the probabilities of weights near 146 (from 145.9 to 146.1 by 0.01).

w=[145.9:0.01:146.1];

pdf=normpdf(w,150,15)

sum(pdf)

pdf =

 Columns 1 through 7

 0.025621 0.025626 0.02563 0.025635 0.02564 0.025644 0.025649

 Columns 8 through 14

 0.025653 0.025658 0.025663 0.025667 0.025672 0.025676 0.025681

 Columns 15 through 21

 0.025685 0.02569 0.025694 0.025699 0.025703 0.025708 0.025712

ans =

 0.53901

So if I add up the probabilities (Axiom 3) of those with weights between 145.9 and 146.1, I get 53.9%, so therefore, more than half of all male students have weights between 145.9 and 146.1 (which is clearly false).

And if we use the same range with an increment of 0.001 lbs then the sum is: 515.91%. Clearly this shows the technique employed is wrong (that is, the values coming from the **normpdf**  function are NOT probabilities but are the values of the probability distribution function (pdf).

**Revised Interpretation of the pdf**

For a discrete function the pdf is the “likelihood” that a random variable takes on a particular value. This IS consistent with our traditional concept of “probability).

For a continuous function the pdf is the “relative likelihood” that a random variable takes on a particular value. The numerical values are useful (for example you can plot the shape of the distribution) and the relative heights of the function can be used to express the “relative probability” of the two values.

For example: compare P(146) against P(120)

p146 = normpdf(146,150,15)

p120 = normpdf(120,150,15)

p146/p120

p146 =

 0.025667

p120 =

 0.0035994

ans =

 7.1309

Hence you are about 7 times more likely (relative) to weigh around 146 lbs than around 120 pound.

Note: We did not actually answer the question which we posed at the beginning of this session. Instead, we have tried to illustrate an important property of discrete and continuous distribution functions.

And to be fair, we didn’t ask an equivalent question about the die outcome either. We should have asked, *“We roll one die. What number came up?”* This is an issue of “simulation” and requires different functions. We will return to answer about simulating the weight of the students seating on the 5 rows in a future lecture.

**Review: Probability Density Functions vs Cumulative Probability Functions**

The probability density function (such as **normpdf**) is used to judge the shape of the distribution function. We speak of a “bell shaped” curve for example. In practice there is very little other use for the pdf (of a continuous function).

The cumulative distribution function (such as **normcdf**), represents the integral of the pdf. The range of the integration traditionally starts at –infinite (if appropriate) and goes to the value “x”. The implication is that the area associated with a continuous cdf(x) represents the probability that the random variable will be in the range of x or less.

For example:

format short e

ww=[100, 150, 194]; % let’s try 3 different weights

normcdf(ww,150,15)

ans =

 4.2906e-004 5.0000e-001 9.9832e-001

Thus, normcdf(100) = 0.00043 (that is, 0.043% of weights are less than 100) and

normcdf(150) = 0.50000 (that is, 50% of weights are less than 150, 50% are greater) and

normcdf(194) = 0.99832, (that is, 99.832% of students weigh less than 194 and 0.168% weigh more than 194).



By taking “differences” in the continuous cdf data, we learn about the probability of being “within” the interval. For example:

normcdf(162) – normcdf(142) = 0.55305 – 0.29690 = 0.25613 (meaning, 25.6% of students weigh between 142 and 162 lbs). Note: We have not discretized this data to simplify the discussion. Generally this will not result in a major difference when intervals are being considered.

#  Discrete Probability Functions

Probability distributions arise from experiments (or phenomena) where the outcome is subject to chance. The nature of the experiment dictates which probability distributions may be appropriate for modeling the resulting random outcomes. The best selection is the one that provides the best agreement with the physical phenomena.

The probabilities associated with different phenomena depend on their “physics” “biology” and “chemistry”. For example, whether a head or tail or something else (landing on edge or disappearing into another dimension) will depend on the coins makeup, the surface its dropped onto or who catches it, the balance of the coin, the fashion in which it was dropped or tossed, etc. We often employ very simple models for this complex behavior out of necessity and in recognition that we usually get satisfactory results.

For example, we may assume that performance on an exam is “normal” (Gaussian) and hence a professor is justified to use the mean and standard deviation to establish the grading curve (this professor does NOT follow that practice).

Similarly we may assume weight data to be normally distributed, or height data (in fact, as mathematics demonstrated, it is impossible for both of those statements to be true simultaneously).

Nevertheless, it will be important to remember that our “predictions” will only be as good as our knowledge of the situation.

The following table shows the available MATLAB functions describing discrete distributions:

| **Name** | [**pdf**](http://www.mathworks.com/help/toolbox/stats/f4218.html#f4232) | [**cdf**](http://www.mathworks.com/help/toolbox/stats/f4218.html#f4244) | [**inv**](http://www.mathworks.com/help/toolbox/stats/f4218.html#f4257) | [**stat**](http://www.mathworks.com/help/toolbox/stats/f4218.html#bqt4w09-1) | [**rnd**](http://www.mathworks.com/help/toolbox/stats/f4218.html#bqtuw3l) |
| --- | --- | --- | --- | --- | --- |
| [Binomial](http://www.mathworks.com/help/toolbox/stats/brn2ivz-9.html) | [binopdf](http://www.mathworks.com/help/toolbox/stats/binopdf.html)  | [binocdf](http://www.mathworks.com/help/toolbox/stats/binocdf.html) | [binoinv](http://www.mathworks.com/help/toolbox/stats/binoinv.html) | [binostat](http://www.mathworks.com/help/toolbox/stats/binostat.html) | [binornd](http://www.mathworks.com/help/toolbox/stats/binornd.html), [random](http://www.mathworks.com/help/toolbox/stats/random.html), [randtool](http://www.mathworks.com/help/toolbox/stats/randtool.html) |
| [Bernoulli](http://www.mathworks.com/help/toolbox/stats/brn2ivz-2.html) |   |   |   |   |   |
| [Geometric](http://www.mathworks.com/help/toolbox/stats/brn2ivz-58.html) | [geopdf](http://www.mathworks.com/help/toolbox/stats/geopdf.html) | [geocdf](http://www.mathworks.com/help/toolbox/stats/geocdf.html) | [geoinv](http://www.mathworks.com/help/toolbox/stats/geoinv.html) | [geostat](http://www.mathworks.com/help/toolbox/stats/geostat.html) | [geornd](http://www.mathworks.com/help/toolbox/stats/geornd.html), [random](http://www.mathworks.com/help/toolbox/stats/random.html), [randtool](http://www.mathworks.com/help/toolbox/stats/randtool.html) |
| [Hypergeometric](http://www.mathworks.com/help/toolbox/stats/brn2ivz-62.html) | [hygepdf](http://www.mathworks.com/help/toolbox/stats/hygepdf.html) | [hygecdf](http://www.mathworks.com/help/toolbox/stats/hygecdf.html) | [hygeinv](http://www.mathworks.com/help/toolbox/stats/hygeinv.html) | [hygestat](http://www.mathworks.com/help/toolbox/stats/hygestat.html) | [hygernd](http://www.mathworks.com/help/toolbox/stats/hygernd.html), [random](http://www.mathworks.com/help/toolbox/stats/random.html) |
| [Multinomial](http://www.mathworks.com/help/toolbox/stats/brn2ivz-84.html) | [mnpdf](http://www.mathworks.com/help/toolbox/stats/mnpdf.html) |   |   |   | [mnrnd](http://www.mathworks.com/help/toolbox/stats/mnrnd.html) |
| [Negative binomial](http://www.mathworks.com/help/toolbox/stats/brn2ivz-101.html) | [nbinpdf](http://www.mathworks.com/help/toolbox/stats/nbinpdf.html) | [nbincdf](http://www.mathworks.com/help/toolbox/stats/nbincdf.html) | [nbininv](http://www.mathworks.com/help/toolbox/stats/nbininv.html) | [nbinstat](http://www.mathworks.com/help/toolbox/stats/nbinstat.html) | [nbinrnd](http://www.mathworks.com/help/toolbox/stats/nbinrnd.html), [random](http://www.mathworks.com/help/toolbox/stats/random.html), [randtool](http://www.mathworks.com/help/toolbox/stats/randtool.html) |
| [Poisson](http://www.mathworks.com/help/toolbox/stats/brn2ivz-127.html) | [poisspdf](http://www.mathworks.com/help/toolbox/stats/poisspdf.html) | [poisscdf](http://www.mathworks.com/help/toolbox/stats/poisscdf.html) | [poissinv](http://www.mathworks.com/help/toolbox/stats/poissinv.html) | [poisstat](http://www.mathworks.com/help/toolbox/stats/poisstat.html) | [poissrnd](http://www.mathworks.com/help/toolbox/stats/poissrnd.html), [random](http://www.mathworks.com/help/toolbox/stats/random.html), [randtool](http://www.mathworks.com/help/toolbox/stats/randtool.html) |
| [Uniform (discrete)](http://www.mathworks.com/help/toolbox/stats/brn2ivz-154.html) | [unidpdf](http://www.mathworks.com/help/toolbox/stats/unidpdf.html) | [unidcdf](http://www.mathworks.com/help/toolbox/stats/unidcdf.html) | [unidinv](http://www.mathworks.com/help/toolbox/stats/unidinv.html) | [unidstat](http://www.mathworks.com/help/toolbox/stats/unidstat.html) | [unidrnd](http://www.mathworks.com/help/toolbox/stats/unidrnd.html), [random](http://www.mathworks.com/help/toolbox/stats/random.html), [randtool](http://www.mathworks.com/help/toolbox/stats/randtool.html) |

We will only be considering four in this course:

* Bernoulli – single success/failure events
* Binomial – series of independent success/failure events
* Poisson – independent “rare” events (usually expressed as a rate)
* Negative Binomial and Geometric

Bernoulli Distribution

The simplest probability function describes events where there are only the outcomes usually called “success” and “failure” or some similar language such as “germinate”, “passed”, “ignited”, “was a dud”, “met spec”, “sold”, etc.

“p” is the probability of success

“q” is the probability of failure where q = 1-p

Hence, if there is a 45% chance a match will ignite on the first “strike” there is a 55% change it will fail to ignite.

The Bernoulli distribution is a special case of the Binomial distribution where n=1.

Let’s write a function called “**bernoulli(p)**” that returns the value “True” p% of the time and False otherwise.

function tf = bernoulli( p )

if rand<p

 tf=true;

else

 tf=false;

end

end

Let’s try the function and generate 30 **bernoulli** events with a probability of 0.2.

p=0.2;

for k=1:30

 bern(k)=bernoulli(p);

end

bern

sum(bern)/30

bern =

 Columns 1 through 11

 1 0 0 0 0 0 0 0 0 0 0

 Columns 12 through 22

 0 0 0 0 0 0 1 0 0 0 1

 Columns 23 through 30

 0 0 0 1 0 1 0 0

ans =

 0.16667

Notice the “expected outcome” (0.2) differs from the “observed outcome” (0.166667). If we repeated this many times, the observed outcome would (should) average to what is expected. Notice that “expected” is used in the sense of statistics (what will be observed in the long-run).

We could have also used the following as our **bernoulli** function:

function tf = bernoulli( p )

tf = binornd(p);

end

Binomial Distribution

The next probability function describes a series of “n” independent Bernoulli events each having probability “p” of success (again, “success” can be interpreted in many ways).

Characteristics of the Binomial Distribution

* The “experiment” (or situation) consists of “n” identical trials
* There are only two possible outcomes on each trial denoted “s” (success) and “f” (failure).
* The probability of “s” remains the same for all trials. The probability is denoted “p” and the probability of “f” is denoted “q” where: p + q = 1
* The binomial random variable (binornd) is the number of “s” in “n”.
* The binomial distribution (binopdf) has two parameters, “p” and “n”.
* The mean, μ , (expected number of successes) of a Binomial random variable is

μ = np

* The variance of a binomial random variable is σ2 = npq

**>> help binornd**

 *BINORND Random arrays from the binomial distribution.*

 *R = BINORND(N,P,MM,NN) returns an array of random numbers chosen from a*

 *binomial distribution with parameters N and P. The size of R is the*

 *common size of N and P if both are arrays. If either parameter is a*

 *scalar, the size of R is the size of the other parameter.*

 *R = BINORND(N,P,MM,NN,...) or R = BINORND(N,P,[MM,NN,...]) returns an*

 *MM-by-NN-by-... array.*

**>> help binopdf**

 *BINOPDF Binomial probability density function.*

 *Y = BINOPDF(X,N,P) returns the binomial probability density*

 *function with parameters N and P at the values in X.*

 *Note that the density function is zero unless X is an integer.*

 *The size of Y is the common size of the input arguments. A scalar input*

 *functions as a constant matrix of the same size as the other inputs.*

This function can be used to answer questions of the following type:

1. **What is chance (probability) of getting 70 seeds (exactly) to germinate when each seed has a probability of 60% and we plant 95 seeds.**

>> binopdf(70, 95, 0.6)= 0.0018498

1. **What is the probability that a coin will be flipped 8 times and come up heads (exactly) 6 times?**

>> binopdf(6, 8, 0.5)= 0.10937

1. **What is the probability that you can put one “blank” bullet in a revolver, rotate the barrel to a random position, squeeze the trigger (exactly) 4 times and not hear any shots?**

>> binopdf(0, 4, 1/6)= 0.48225

Properties of the Binomial Distribution (pdf and cdf)

Let’s look at the properties of the binomial distribution before we consider other problems using this important distribution.

The three charts below show the general shape, symmetry and centering tendency of different “binomial situations”

Case (1) 10 repeated events with a probability of success of 0.5

Case (2) 20 repeated events with a probability of success of 0.5

Case (3) 20 repeated events with a probability of success of 0.1

Notice that the “centering” is occurring at: xcenter = np (the most “likely” outcome).



x = [0:10]';

x\_bino\_pdf = binopdf(x, 10, 0.5);

subplot(1,3,1)

bar(x\_bino\_pdf,0.5)

legend('pdf', 'cdf', 'Location','NorthWest')

x = [0:20]';

x\_bino\_pdf = binopdf(x, 20, 0.5);

subplot(1,3,2)

bar(x\_bino\_pdf,0.5)

x\_bino\_pdf = binopdf(x, 20, 0.1);

subplot(1,3,3)

bar(x\_bino\_pdf,0.5)

legend('pdf', 'cdf', 'Location','NorthWest')

Problems related to the above analysis:

1. **What is the probability of getting at least 7 tails when you flip a coin 10 times?**

Answer: Getting at least 7 tails means getting 7 or 8 or 9 or 10 tails. This is 1.0 - the probability of getting 6 or 5 or 4 or 3 or 2 or 1 or 0 tails, that is:

P(x>=7) = 1.0-P(x<=6)

>> 1.0-binocdf(6,10,0.5) = 0.17188

We could have also solved the problem by considering the “failures” (in this case the heads). Getting 7 or more tails means getting 3 or fewer heads, that is:

P(x<=3)

>> binocdf(3,10,0.5) = 0.17188

1. **What is the probability of getting fewer than 3 heads when you flip a coin 10 times?**

Answer: Getting fewer than 3 heads means getting 2 or 1 or 0 heads. That is:

P(x<=2)

>> binocdf(2,10,0.5) = 0.054687

1. **What is the probability of getting between 12 and 16 heads inclusive when you flip a coin 20 times?**

Answer: Getting between 12 and 16 heads inclusive means getting 12 or 13 or 14 or 15 or 16. Looking at a number line we can see the relationship more easily:

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Call the probability of getting 16 or less P(A) and getting 11 or less P(B)

Then the probability of getting between 12 and 16 is P(A)-P(B)

Hence, the probability we seek is:

P(12≤x≤16) = P(x≤=16) - P(x≤11)

or

>> binocdf(16,20,0.5)- binocdf(11,20,0.5) = 0.25043

1. **What is the probability of seeing more than 5 seeds out of 20 failing to germinate when the germination rate is 90%?**

Answer: Getting more than 5 seeds to fail to germinate means getting 6 or 7 or

8 .. 20 failing to germinate. This is the same as getting 1.0 – the probability of getting 5 or 4 or 3 or 2 or 1 or 0 seeds failing to germinate.

Therefore, we seek:

P(x>5) = 1.0 – P(x≤5)

where p = 0.1 (failing=success)

>> 1-binocdf(5,20,0.1) = 0.011253

**Poisson Distribution**

The Poisson probability distribution function was named for the French mathematician S.D. Poisson and provides a model for the relative frequency of the number of rare events that occur in a unit of time, area, volume, etc (that is, a rate).

For example:

* The number of fatal accidents per month in a manufacturing plant
* The number of visible defeats in a diamond
* The number of people struck by lightning in the month of July in Alabama

Characteristics of the Poisson Distribution

* The “experiment” (or situation) consists of counting the number “y” of times a particular event occurs during a given unit of time or in a given area or volume or weight or distance or etc.
* The probability that an event occurs in a given time (etc.) is the same for all the events
* The number of events that occur in one unit of time is independent of the number that occur in other units
* The Poisson distribution has one parameter λ (lambda) which is the rate (number of events per unit time).
* The mean, μ , (expected number of events) of a Poisson random variable is λ, therefore, μ = λ
* The variance of a Poisson random variable is also λ, therefore, σ2 = λ

**>> help poissrnd**

 *POISSRND Random arrays from the Poisson distribution.*

 *R = POISSRND(LAMBDA) returns an array of random numbers chosen from the*

 *Poisson distribution with parameter LAMBDA. The size of R is the size*

 *of LAMBDA.*

 *R = POISSRND(LAMBDA,M,N,...) or R = POISSRND(LAMBDA,[M,N,...]) returns*

 *an M-by-N-by-... array.*

**>> help poisspdf**

 *POISSPDF Poisson probability density function.*

 *Y = POISSPDF(X,LAMBDA) returns the Poisson probability density*

 *function with parameter LAMBDA at the values in X.*

 *The size of Y is the common size of X and LAMBDA. A scalar input*

 *functions as a constant matrix of the same size as the other input.*

 *Note that the density function is zero unless X is an integer.*



Problems using the Poisson function:

1. **Typically there are 5 attempted murders in Auburn each year. What is the probability that there will be no attempted murders in 2005? How many years would you expect to wait to see a year with no attempted murders?**

Answer:

The probability of seeing 0 murders when they typically occur at a rate of 5 per year is:

P(x=0)

>> poisscdf(0,5) = 0.006738 % that is 0.67%.

How long: If there is a 1% chance something will happen in a year, then we expect to see it happen once each 1/0.01 or 100 years. Therefore, since our probability is 0.67% we expect to see no attempted murders (on average) every 1/0.006738 = 148.4 years

1. **Typically there are 5 attempted murders in Auburn each year. What is the probability that there will be more than 5 attempted murders in 2005?**

Answer:

Seeing more than 5 murders means seeing 6 or 7 or 8 .. ∞. This is 1.0 – probability of seeing 5 or 4 or 3 .. 0. Therefore:

P(x>5) = 1 – P(x≤5)

>> 1.0 – poisscdf(5,5) = 0.384039 % or 38.4%

1. **Typically there are 5 attempted murders in Auburn each year. What is the probability that there will be more than 5 attempted murders in 2005, 2006, and 2007?**

Answer:

We know from (2) that there is a 38.4% chance of seeing the event “more than 5 murders” occur. This is now a Binomial Distribution problem since we want to see the probability of 3 successes in the next three years, therefore,

P(x>5 in 2005, 2006, and 2007)

>> binopdf(3, 3, 0.384) = 0.056623 % or 5.66%

**Negative Binomial Distribution**

The negative binomial distribution describes a situation in which there is a series of identical events each of which is S “success” or F “failure”. We are interested in describing the number of failures “f” that preceed the “rth” success (many authors use “s” in place of “r”), when the constant probability of a success is “p”. This function is similar to the binomial distribution, except that the number of successes is fixed, and the number of trials is variable. Like the binomial, trials are assumed to be independent.

For example, you need to find 10 people with excellent reflexes, and you know the probability that a candidate has these qualifications is 0.3. The negative binomical distribution allows you to calculate the probability that you will interview a certain number of unqualified candidates before finding all 10 qualified candidates.

Characteristics of the Negative Binomial Distribution

1. The “experiment” (or situation) consists of counting the number failures “f” before the “rth” success (r is fixed in advance).
2. The probability that an event occurs is the same for all the events
3. The Negative Binomical distribution has two parameters “p” which is the probability of success on a single trial and “r” the number of successes required.
4. The mean can be determined from μ = r/p
5. The variance can be determined from σ2 = rq/p2

Consider a hypothetical situation (coin flip) where p=0.5 and r = 5 (heads).

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Y | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
| outcome | H | H | T | T | T | T | H | T | T | H | H |
| #successes | 1 | 2 | 2 | 2 | 2 | 2 | 3 | 3 | 3 | 4 | 5 |

Here, f=6 (failures), s=5 (successes), total = 11 events

Typical questions involving the negative binomial distribution:

* What is the probability that I will have to flip a fair coin (p=0.5) 10 times before seeing the 5th head?
* What is the probability that I will have to flip a fair coin (p=0.5) 10 times before seeing the 10th head?
* Two out of every 100 oysters contain a pearl (p=0.02). What is the probability that I will have to open the 100th oysters to find five pearls (that is, 95 failures plus 4 pearls before finding the 5th pearl)?

**>> help nbinrnd**

 *NBINRND Random arrays from the negative binomial distribution.*

 *RND = NBINRND(R,P,M,N) returns an array of random numbers chosen from a*

 *negative binomial distribution with parameters R and P. The size of RND*

 *is the common size of R and P if both are arrays. If either parameter*

 *is a scalar, the size of RND is the size of the other parameter.*

 *RND = NBINRND(R,P,M,N,...) or RND = NBINRND(R,P,[M,N,...]) returns an*

 *M-by-N-by-... array.*

**>> help nbinpdf**

 *NBINPDF Negative binomial probability density function.*

 *Y = NBINPDF(X,R,P) returns the negative binomial probability density*

 *function with parameters R and P at the values in X.*

 *Note that the density function is zero unless X is an integer.*

 *The size of Y is the common size of the input arguments. A scalar input*

 *functions as a constant matrix of the same size as the other inputs.*

Problem using the Negative Binomial Distribution function:

1. **Suppose we are at a rifle range with an old gun that misfires 5 out of 6 times. Define success as the event the gun fires (therefore, p=1/6 = 0.1666667). I want to “fire” the gun 3 times. What is the probability that this will happen on the 10th try (that is the 10th try “fires” and is the third time successfully fired).**

Answer:

>> nbinpdf(7, 3, 1/6) = 0.046514 % or 4.7%

**Geometric Probability Distribution**

The geometric probability distribution is a special case of the negative binomial distribution where r=1, that is, we are looking for the first success after an unbroken string of failures.

Problems using the Geometric Probability Distribution function:

1. **Suppose McDonalds is having a contest where you have a 1 in 4 chance of getting a “free drink”. You have already been to McDonalds three times and not won. What is the probability that you will get a free drink on your next visit?**

Answer:

>> p=0.25 or % or 25%

1. **Suppose McDonalds is having a contest where you have a 1 in 4 chance of getting a “free drink”. You plan to go to the store once daily until you have won. What is the chance that you will have to visit McDonalds four times before getting your free drink?**

Answer:

>> nbinpdf(3, 1, 0.25) = 0.105

**Hypergeometric Distribution Function**

When we are sampling from a infinite or large distribution, we do not need to return the “item” to the “population” before selecting another item. That is, when we flipped a coin and saw a head we didn’t “use up” one of the “head” in the population (there would still be an equal probability of getting a head or a tail on subsequent flips). However, when we deal with finite (small) populations, we “change the odds” when we continue sampling.

We differentiate between these two cases using the nomenclature “with replacement” or “without replacement”. Again, for large populations this makes no difference.

*This description is provided only for your future reference. We will not be looking in detail at this distribution.*

# Continuous Distributions

**Normal and Standard Normal (Gaussian) Distribution**

The standard normal distribution (the so-called z-distribution) is a normal distribution with a mean of 0 and a standard deviation of 1.0. When the mean and standard deviation are not 0 and 1, the distribution is termed “normal” rather than “standard”. The standard normal distribution is the table values normally found in statistics textbooks. Knowing the general properties of the standard normal distribution is essential to understanding all normal distributions.

**NORMPDF Normal probability density function (pdf).**

 *Y = NORMPDF(X,MU,SIGMA) returns the pdf of the normal distribution with*

 *mean MU and standard deviation SIGMA, evaluated at the values in X.*

 *The size of Y is the common size of the input arguments. A scalar*

 *input functions as a constant matrix of the same size as the other*

 *inputs.*

 *Default values for MU and SIGMA are 0 and 1 respectively.*

**>> help norminv**

 *NORMINV Inverse of the normal cumulative distribution function (cdf).*

 *X = NORMINV(P,MU,SIGMA) returns the inverse cdf for the normal*

 *distribution with mean MU and standard deviation SIGMA, evaluated at*

 *the values in P. The size of X is the common size of the input*

 *arguments. A scalar input functions as a constant matrix of the same*

 *size as the other inputs.*

 *Default values for MU and SIGMA are 0 and 1, respectively.*

 *[X,XLO,XUP] = NORMINV(P,MU,SIGMA,PCOV,ALPHA) produces confidence bounds*

 *for X when the input parameters MU and SIGMA are estimates. PCOV is a*

 *2-by-2 matrix containing the covariance matrix of the estimated parameters.*

 *ALPHA has a default value of 0.05, and specifies 100\*(1-ALPHA)% confidence*

 *bounds. XLO and XUP are arrays of the same size as X containing the lower*

 *and upper confidence bounds.*

**>> help normrnd**

 *NORMRND Random arrays from the normal distribution.*

 *R = NORMRND(MU,SIGMA) returns an array of random numbers chosen from a*

 *normal distribution with mean MU and standard deviation SIGMA. The size*

 *of R is the common size of MU and SIGMA if both are arrays. If either*

 *parameter is a scalar, the size of R is the size of the other*

 *parameter.*

 *R = NORMRND(MU,SIGMA,M,N,...) or R = NORMRND(MU,SIGMA,[M,N,...])*

 *returns an M-by-N-by-... array.*

|  |  |  |  |
| --- | --- | --- | --- |
| **num\_sd** | **Item** | **Exact Value** | **TraditionalValue** |
| 1 | middle = normcdf(0,0,1) | middle =0.5 | 0.5 |
| 1 | left\_tail1 = normcdf(-num\_sd,0,1) | left\_tail1 =0.15866 | 1/6 |
| 1 | right1 = normcdf(num\_sd,0,1) | right1 =0.84134 | 5/6 |
| 1 | **area1 = normcdf(num\_sd,0,1)-normcdf(-num\_sd,0,1)** | **area1 =****0.68269** | **2/3** |
| 2 | left\_tail2 = normcdf(-num\_sd,0,1) | left\_tail2 = 0.02275 | 2.5% |
| 2 | right2 = normcdf(num\_sd,0,1) | right2 =0.97725 | 97.5% |
| 2 | **area2 = normcdf(num\_sd,0,1)-normcdf(-num\_sd,0,1)** | **area2 =****0.9545** | **95%** |
| 3 | left\_tail3 = normcdf(-num\_sd,0,1) | left\_tail3 =0.0013499 | 0.1% |
| 3 | right3 = normcdf(num\_sd,0,1) | right3 =0.99865 | 99.9% |
| 3 | **area3 = normcdf(num\_sd,0,1)-normcdf(-num\_sd,0,1)** | **area3 =****0.9973** | **99.8%** |





**Normal Distribution**

Effect of changing population mean (mu) with sigma = 1.



Effect of standard deviation (sigma) holding mean constant (mu=4):



**Problems involving the normal distribution. *The “answers” provided are subject to the notes below!***

**Typically on a fluids exam, students get an average grade of 60 with a standard deviation of 15 points.**

1. **What fraction of a class would be expected to get the grade of 55 (exactly)?**

ans\_a = normpdf(55,60,15)= 0.025159

1. **What fraction of the class would get a grade below 40?**

ans\_b = normcdf(40,60,15)= 0.091211

1. **What fraction of the class would get a grade of between 50 and 80 (inclusive)?**

ans\_c = normcdf(80,60,15)-normcdf(50,60,15)= 0.6563

1. **What fraction of the class would get a grade of between 50 and 80 (exclusive)?**

ans\_d = normcdf(79.5,60,15)-normcdf(50.5,60,15)= 0.63994

1. **What fraction of the class would get a grade of more than 100?**

ans\_e = 1-normcdf(100,60,15)= 0.0038304

**Notes:**

If we understand that grading is being done to the nearest “one” point, then a grade such as 50 should consider grades between 49.5 and 50.5. We should also understand that the concept of “inclusive” and “exclusive” as applied to a discrete distribution is different than when applied to a continuous distribution. (see part (a) for example)

a. This answer is NOT the probability. The probability is 0.0 (continuous distribution). This is the “relative probability”.

b. It would make no difference whether we include the grade of 40 or not. But a correct rendering of this problem would be to numerical grades lower than 40 (that is, 39, 38, etc) which are really 39.5 and lower.

c. The wording “inclusive” is not appropriate for a continuous distribution but can be applied to the grading distribution, hence we should really be considering **normcdf(80.5,60,15)-normcdf(49.5,60,15)**

d. Again, assuming a grading distribution, we really are asking for grades 51 and 79. That would be (with rounding) grades between 79.5 (which would round down to 79) and 50.5 (which would round up to 51).

e. Although grades of greater than 100 are physically impossible (assuming no bonus credit) the distribution nonetheless predicts some students will receive that grade. We might also see negative scores in the case of a distribution such as mean=40 and stdev=20.

For example, **normcdf(0,40,20) = 0.02275** or 2.3% of the class would have negative scores

**Exponential Probability Distribution**

The exponential distribution is a special form of a more general distribution called the gamma density function. *We will not specifically cover that distribution.*

The exponential distribution is employed to model the following situations:

* The length of time between arrivals at a service counter (Geek squad, supermarket checkout, clinic, etc.) when the probability of a customer arrival in any one unit of time is equal to the probability of arrival during any other.
* The time until a radioactive particle decays, or the time between clicks of a Geiger counter
* The time it takes before your next telephone call
* The time until default (on payment to company debt holders) in reduced form credit risk modeling
* Exponential variables can also be used to model situations where certain events occur with a constant probability per unit length, such as the distance between mutations on a DNA strand, or between “roadkills” on a given road

It is also used to model the length of life of industrial equipment or products when the probability that an “old” component will operate at least “t” additional time units, given it is now functioning, is the same as the probability that a “new” component will operate at least “t” time units. Equipment what is subject to periodic maintenance and parts replacement often exhibit this property of “never growing old”. This is the property of “memoryless”.

**Illustration of memoryless.**  The conditional probability that we need to wait, for example, more than another 10 seconds before the first arrival, given that the first arrival has not yet happened after 30 seconds, is equal to the initial probability that we need to wait more than 10 seconds for the first arrival. So, if we waited for 30 seconds and the first arrival didn't happen (T > 30), probability that we'll need to wait another 10 seconds for the first arrival (T > 30 + 10) is the same as the initial probability that we need to wait more than 10 seconds for the first arrival (T > 10). The fact that Pr(T > 40 | T > 30) = Pr(T > 10) does not mean that the events T > 40 and T > 30 are independent.

The exponential distributions and the geometric distributions are the only memoryless probability distributions.

**>> help exppdf**

 *EXPPDF Exponential probability density function.*

 *Y = EXPPDF(X,MU) returns the pdf of the exponential distribution with*

 *mean parameter MU, evaluated at the values in X. The size of Y is*

 *the common size of the input arguments. A scalar input functions as a*

 *constant matrix of the same size as the other input.*

 *The default value for MU is 1.*

**>> help exprnd**

 *EXPRND Random arrays from exponential distribution.*

 *R = EXPRND(MU) returns an array of random numbers chosen from the*

 *exponential distribution with mean parameter MU. The size of R is*

 *the size of MU.*

 *R = EXPRND(MU,M,N,...) or R = EXPRND(MU,[M,N,...]) returns an*

 *M-by-N-by-... array.*

**>> help expcdf**

 *EXPCDF Exponential cumulative distribution function.*

 *P = EXPCDF(X,MU) returns the cdf of the exponential distribution with*

 *mean parameter MU, evaluated at the values in X. The size of P is*

 *the common size of the input arguments. A scalar input functions as a*

 *constant matrix of the same size as the other input.*

 *The default value for MU is 1.*

Problems involving the Exponential distribution function

1. **Breakdowns in equipment in a large industrial plant have been observed to be approximately an exponential distribution with a mean (mu) of 2, that is, 1 every 2 hours. The workday has just started.**
	1. **What is the probability that there will be at least one-hour of operation before the first breakdown?**
	2. **What is the probability that there will be no more than four hours before the first breakdown?**
	3. **What is the probability that the time to the next breakdown will be greater than the average (expected) value? The expected value would be the mean (mu=2 hours).**

ans\_a = 1-expcdf(1,2) = 0.60653

ans\_b = expcdf(4,2) = 0.86466

ans\_c = 1-expcdf(2,2) = 0.36788

1. **The lifetime “y” in hours of the CPU in a certain type of microcomputer is an exponential random variable. The average lifetime of the CPU’s is 1000 hours.**
2. **What is the probability that the CPU will have a lifetime of at least 2000 hours?**
3. **What is the probability that the CPU will have a lifetime of at most 1500 hours?**

ans\_d = 1-expcdf(2000,1000) = 0.13534

ans\_e = expcdf(1500,1000) = 0.77687

**Student’s t Distribution**

In probability and statistics, Student’s t-distribution (or simply the t-distribution) is a continuous probability distribution that arises when estimating the mean of a normally distributed population in situations where the sample size is small and population standard deviation is unknown. It plays a role in a number of widely-used statistical analyses, including the Student’s t-test for assessing the statistical significance of the difference between two sample means, the construction of confidence intervals for the difference between two population means, and in linear regression analysis.

The t-distribution is symmetric and bell-shaped, like the normal distribution, but has heavier tails, meaning that it is more prone to producing values that fall far from its mean.

In the English literature, a derivation of the *t*-distribution was published in 1908 by William Sealy Gossetwhile he worked at the Guinness Brewery in Dublin. One version of the origin of the pseudonym *Student* is that Gosset's employer forbade members of its staff from publishing scientific papers, so he had to hide his identity. Another version is that Guinness did not want their competition to know that they were using the *t*-test to test the quality of raw material. The *t*-test and the associated theory became well-known through the work of R.A. Fisher, who called the distribution "Student's distribution".

****

When to Use the t Distribution

The t distribution can be used with any statistic having a bell-shaped distribution (i.e., approximately normal). The central limit theorem states that the sampling distribution of a statistic will be normal or nearly normal, if any of the following conditions apply.

* The population distribution is normal.
* The sampling distribution is symmetric, unimodal, without outliers, and the sample size is 15 or less.
* The sampling distribution is moderately skewed, unimodal, without outliers, and the sample size is between 16 and 40.
* The sample size is greater than 40, without outliers.

The t distribution should not be used with small samples from populations that are not approximately normal.

**t-Distribution Example Problems**

1. **Acme Corporation manufactures light bulbs. The CEO claims that an average Acme light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?**

We need to compute the t score (t-statistic), based on the following equation:

t = [ x - μ ] / [ s / sqrt( n ) ]
t = ( 290 - 300 ) / [ 50 / sqrt( 15) ] = -10 / 12.909945 = - 0.7745966

where x is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size. The degrees of freedom (DOF) are equal to 15 - 1 = 14.

Now we employ the tcdf to learn the probability.

xbar = 290;

mu = 300;

n = 15;

dof=n-1;

s = 50; % standard deviation of sample

t = (xbar-mu)/(s/sqrt(n)); % t-statistic

ans\_a = tcdf(t,dof)

**t = -0.7746**

**ans\_a = 0.22573**

The cumulative probability is 0.226 or 22.6%. Hence, if the true bulb life were 300 days, there is a 22.6% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

1. **MATLAB t-Distribution from Help. We have a known population mean mu=1 and known population standard deviation of 2. A sample of 100 values is taken from this distribution. We calculate the sample mean and sample standard deviation.**

mu = 1; % Population mean

sigma = 2; % Population standard deviation

n = 100; % Sample size

x = normrnd(mu,sigma,n,1); % Random sample from population

xbar = mean(x); % Sample mean

s = std(x); % Sample standard deviation

**We calculate the t statistic and associated probability.**

t = (xbar-mu)/(s/sqrt(n)) % t-statistic

t =

 0.2489

p = 1-tcdf(t,n-1) % Probability of larger t-statistic

p =

 0.4020

Note: A simpler way to perform this type of calculation is found using the t-test. This probability is the same as the *p* value returned by a *t*-test of the null hypothesis that the sample comes from a normal population with mean *μ*. *The next statistics topic covered hypothesis testing in more detail.*

[h,ptest] = ttest(x,mu,0.05,'right')

h =

 0

ptest =

 0.4020

**Code Segments Used in This Chapter**

Code Segment (1)

%% Code for One Die Roll

sides=6;

x = 1:sides;

p(x) = 1/sides;

figure;

subplot(1,2,1);

bar(x,p,0.5);

axis([0, sides+1, 0, 0.25]);

xlabel('x, # dots')

ylabel('probability, p')

title('pdf For Rolling One Die');

subplot(1,2,2);

plot(x,p,x,p,'o');

axis([0, sides+1, 0, 0.25]);

xlabel('x, # dots')

ylabel('probability, p')

title('pdf For Rolling One Die');

Code Segment (2)

%% Code for Two Die Roll

clear; clc;

x = 1:6;

n = 2;

m = length(x);

X = cell(1, n);

[X{:}] = ndgrid(x);

X = X(end : -1 : 1);

y = cat(n+1, X{:});

y = reshape(y, [m^n, n]);

y(1,n+1) = 0;

for i = 1:n

 y(:,n+1) = y(:,n+1) + y(:,i);

end

%%

f = figure('Position',[0 0 300 700]);

the\_title = ['Die Roll Results'];

title(the\_title);

axis off

cnames = {'Die 1','Die 2','Sum'};

uitable('Parent',f,'Data',int8(y),'ColumnName',cnames, 'Position',[20 0 300 650]);

Code Segment (3)

for i = n\*min(x):(n\*max(x))

 nways(i,1) = i;

 nways(i,2) = sum(y(:,n+1)==i);

end

close all

nways(:,3) = nways(:,2)/sum(nways(:,2))

nways(:,4) = cumsum(nways(:,3))

bar(nways(:,3:4))

axis([1, 13, 0, 1]);

legend('pdf', 'cdf', 'Location','NorthWest')

Code Segment (4)

%% graph of standard normal cdf

clear all; clc;

x = linspace(-3,3);

normal\_cdf = normcdf(x,0,1);

subplot(1,2,1)

plot(x, normal\_cdf)

% graph of standard normal cdf via norminv

p = linspace(0,1);

inv\_norm = norminv(p, 0, 1);

subplot(1,2,2)

plot(inv\_norm, p)

grid on

Code Segment (5)

%% effect of mean

clf

x = linspace(0,8);

n1 = normpdf(x,3,1);

n2 = normpdf(x,4,1);

n3 = normpdf(x,5,1);

hold off

plot(x,n1, '-kd', 'LineWidth',2, ...

 'MarkerFaceColor','b')

hold on; grid on

plot(x,n2, '-ks', 'LineWidth',2, ...

 'MarkerFaceColor','m')

plot(x,n3, '-k^', 'LineWidth',2, ...

 'MarkerFaceColor','y')

legend('\mu = 3','\mu = 4','\mu = 5')

Code Segment (6)

%% effect of sigma

clf

x = linspace(0,8);

n1 = normpdf(x,4,1);

n2 = normpdf(x,4,2);

n3 = normpdf(x,4,0.5);

hold off

plot(x,n1, '-kd', 'LineWidth',2, ...

 'MarkerFaceColor','b')

hold on; grid on

plot(x,n2, '-ks', 'LineWidth',2, ...

 'MarkerFaceColor','m')

plot(x,n3, '-k^', 'LineWidth',2, ...

 'MarkerFaceColor','y')

legend('\sigma = 1','\sigma = 2','\sigma = 0.5')