****

[**www.SBAnalysis.com**](http://www.SBAnalysis.com)

**The Journal of Applied Business Statistics**

 **Sample Issue (Article Excerpts)**

**The Journal of Applied Business Statistics**

**Sample Issue - Contents**

**General Applied Business Statistics – Article Excerpts**

Data and Statistics 1

Descriptive Statistics: Properties of Numerical Data – Part One 3

Introduction to Probability 8

Continuous Probability – Uniform and Exponential Distributions 14

**Lean Six Sigma – Article Excerpts**

Value-Stream Mapping 18

Histogram 18

Trend Chart 19

Box Plot 19

**A word of preface to this Sample Excerpt Issue**

If you are new to math, or haven’t used it much since your high school days, please do not be alarmed by this sample. There may be terms, Greek letters or definitions that look foreign to you. With your subscription to the Journal of Applied Business Statistics, we explain in detail each concept prior to moving on to the next. Furthermore, we include a Quick Start Guide as a light math refresher so you can *hit the ground running* with journal issue number one.

With your entrepreneurial thinking wide open, we trust that after reading this sample, you will see how invaluable data analysis can be to the growth of your business/career. **Enjoy!**

**General Applied Business Statistics – Article Excerpts**

**Data and Statistics**

Sample Article Excerpt Taken from Journal Issue No.01

Michael J. Anderson

**1. WHAT IS STATISTICS**

Statistics is the science of collecting, organizing, analyzing, interpreting, and presenting of data. In a nutshell, it is a way of getting information from data. The entirety of this Journal is to assist business owners, and managers at all levels, in utilizing statistical methods to make more informed, and one might argue, more successful business decisions.

It would be hard to imagine a day going by without coming across statistics, whether in a magazine or newspaper article, over the internet or even when checking out at a coffee shop (press here to add a 20% gratuity).

* Batting averages in baseball.
* Percent chance that it will rain tomorrow.
* The percentage change of the Consumer Price Index.
* Is there a correlation between cat ownership and blood pressure?
* The probability of picking the winning lottery numbers.

Examples of statistics in business might be:

* The number of injuries at plant B decreased by 5% last year.
* Last month’s ad campaign yielded 100 new customers.
* The average time spent on the firm’s website was 5.5 minutes per visitor.

The problem is that these are simply numbers on a page, and without statistical analysis, they do little in helping a company make sound decisions. The science of statistics provides a method to analyze and interpret the data – are the data correlated with another variable? Is the data point an outlier or indicative of a pattern? Statistical analysis gives us the tools to spot trends and potential pitfalls. A business cannot escape from data, and applied statistics gives it the methods to work with data effectively.

**1.1 Descriptive Statistics**

Descriptive statistics are the methods used to summarize data, to make it easy for the reader to understand. One form of descriptive statistics uses graphical techniques to present the data in a way that make it easier to extract useful information. Another form uses numerical techniques to summarize the data set. For example, suppose a staff member was analyzing the data of BTU usage for a firm’s manufacturing plants in 4 different regions, see Table 1.

**Table 1:** Total BTU Usage Based on Region – Partial List of 5,684 Data Points

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Data Point** | **Region** | **Total BTUs** |  | **Data Point** | **Region** | **Total BTUs** |
| 1 | 4 | 48974.570 |  | ….5678 | 1 | 105937.379 |
| 2 | 3 | 41534.000 |  | 5679 | 2 | 88123.914 |
| 3 | 3 | 96548.160 |  | 5680 | 4 | 66380.621 |
| 4 | 2 | 89284.345 |  | 5681 | 3 | 79258.454 |
| 5 | 1 | 36582.146 |  | 5682 | 2 | 28839.497 |
| 6 | 1 | 30294.147 |  | 5683 | 3 | 22424.616 |
| 7…. | 3 | 52661.720 |  | 5684 | 2 | 12323.219 |

Management would cringe if handed 5000 plus data points as in Table 1. Conversely, descriptive statistics as in Figure 1 makes it quite easy to see what is going on with the data.

|  |  |
| --- | --- |
| **Region** | **Average BTUs** |
| 1 | 97663.269 |
| 2 | 94244.458 |
| 3 | 71244.117 |
| 4 | 61816.500 |

**Figure 1:** Descriptive Statistical Summary of the Data in Table 1.

Descriptive statistics is what is typically contained in a magazine or newspaper article, and it is what management, or a customer is given, as a clear-to-follow summary, of potentially vast amounts of data.

**1.2 Inferential Statistics**

Inferential statistics is the method used to draw conclusions or inferences about a population based on a sample.

 **Population**: The set of all data points in a particular study.

 **Sample**: A subset of the population.

Many situations consist of a very large group of data points, and to analyze the entire population is often cost prohibitive. It is usually faster, easier and less expensive to draw conclusions or make estimates about a population based on a sample. For example, say we had a swimming pool filled with small colored marbles – red, blue, black and white – and we wanted to know how many there were of each color. The statistician would take a sample, and from that be able to calculate the quantity of each colored marble in the pool. Similarly, a parts manufacturer would take a sample from all the parts in order to calculate the expected number of defects in the entire population. We will cover sampling in detail in a future journal article.

**Descriptive Statistics: Properties of Numerical Data – Part One**

Sample Article Excerpt Taken from Journal Issue No.01

Michael J. Anderson

**1. SUMMARY MEASURES**

In the previous article, recall that we defined a Frequency Distribution as a summary of the data. We will continue that theme by developing numerical summary measures for data sets.

Typically, the parameters associated with a population are unknown. Therefore, we select a random sample from the population and use summary measures to estimate the unknown population parameters.

 A **Statistic** is a numerical value calculated from the observations in the sample.

If the summarizing value is for the entire population, it is called a parameter, if for a sample, it is called a statistic. We will now investigate these summary measures, grouping them into *measures of location*, and *measures of variability* as well as a few additional topics.

**1.1 Measures of Location**

We will begin by looking at measures of location, especially measures of central location, described crudely as *averages*, as they are indicative of the “center” or “middle” of a data set.

**Mean**

By far the most widely used measure of central location is what the layperson calls the “average” but in the parlance of the statistical world, it is called the mean, aka arithmetic mean. If we are dealing with the entire population, the mean is denoted by m, and if dealing with a sample from the population, we use x̄ (called x bar in the math world), and it is defined in equation [1], where *n* is the number of observations (sample size) and *xi* is the *i*th observation.

$Sample Mean: \overbar{x}= \frac{Sum of the values}{Number of values}= \frac{\sum\_{}^{}x\_{i}}{n} where:i=1, 2, 3, …, n$ [1]

 Note: the Greek letter S (sigma) represents summation.

As an illustration, let us consider an online retailer who has a small sample of receipt totals, and it wants to calculate the mean per receipt. Given the following data:

$11.10, $54.08, $12.60, $12.60, $30.00, $9.90, $25.50, $14.85, $19.90. In the form of Equation [1]:

|  |  |
| --- | --- |
| X1 =  | $11.10 |
| X2 =  | $54.08 |
| X3 =  | $12.60 |
| X4 =  | $12.60 |
| X5 =  | $30.00 |
| X6 =  | $9.90 |
| X7 =  | $25.50 |
| X8 =  | $14.85 |
| X9 =  | $19.90 |

 $\overbar{x}$ = (x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9) / 9

 = ($11.10 + $54.08 + $12.60 + $12.60 + $30.00 + $9.90 + $25.50 + $14.85 + $19.90) / 9

= $21.17

In the above example we used a small sample size of nine items, (for illustration only). However, we would almost never use such a small sample size without advanced statistical techniques. In the business world, sample size is determined prior to data collection, to ensure that the confidence interval is narrow enough for sound business decisions. We will discuss sampling in an upcoming journal article.

**Population Mean**

The population mean differs from the sample mean in that it is the mean of an entire population, e.g., the mean age of everyone in the United States. In the example above, all the online sales receipts would represent the entire population. The sample mean is considered a point estimator of the population mean.

For the population mean, represented by m, N represents the number of observations in the entire population, and it is defined in equation [2].

$Population Mean: µ = \frac{\sum\_{}^{}x\_{i}}{N} where:i=1, 2, 3, …, N$ [2]

**Weighted Mean**

Another way, and arguably a better way, of writing the formula for the sample/population mean is as follows:

 $Sample Mean: \overbar{x}= \frac{1}{n} \sum\_{}^{}x\_{i}=\frac{1}{n}\left(x\_{1}+x\_{2}+…+x\_{n}\right)=\frac{1}{n}x\_{1}+\frac{1}{n}x\_{2}+…+\frac{1}{n}x\_{n}$

This form of the mean is more versatile and intuitive in that the 1/n term is actually a scaling or weighting factor.

In a standard mean calculation, each observation is given the same weight, but what if that wasn’t the case. Probably the example that most of us are familiar with is when a university calculates the GPA of its students – it gives a weight of 4 for an “A” grade, 3 for a “B” grade, 2 for a “C” grade, 1 for a “D” grade and 0 for an “F.” Likewise, weights are very important in business, and to calculate a weighted mean we use equation [3].

 $Weighted Mean: \overbar{x}= \frac{\sum\_{}^{}w\_{i}x\_{i}}{\sum\_{}^{}w\_{i}} where: w\_{i}=the weight for data point i$ [3]

As an example, say a company wanted to calculate the average cost of labor to produce each of its products given the data in Table 1.

**Table 1:** Labor Hours & Rates to Produce Products

|  |  |  |
| --- | --- | --- |
|  |  | **Labor Hours perUnit of output** |
|  |  |
| **Labor Grade** | **Pay Rate ($/Hour)** | **Product 1** | **Product 2** |
| Unskilled | $15.00 | 2 | 4 |
| Semiskilled | 20.00 | 3 | 6 |
| Highly Skilled | 30.00 | 7 | 6 |
|  |  Totals: | **12** | **16** |

It would be incorrect to use our standard mean calculation of: ($15 + $20 + $30)/3 = $21.67/hour.

For this example, we must use a weighted mean as follows:

 **Product 1:** $Weighted Mean: \overbar{x}= \frac{\left(2\*\$15\right)+\left(3\*\$20\right)+(7\*\$30)}{(2+3+7)}= \frac{\$300}{12}=\$25.00 per Hour$

 **Product 2:** $Weighted Mean: \overbar{x}= \frac{\left(4\*\$15\right)+\left(6\*\$20\right)+(6\*\$30)}{(4+6+6)}= \frac{\$360}{16}=\$22.50 per Hour$

There are countless of other examples in business where you would want to use a weighted mean. For example, say a milk producer had dairy farms in Michigan, Wisconsin and California, and the production in thousands of gallons per cow was 35.0, 52.0 and 48.6 respectively. We couldn’t say that the average milk production per state is (35.0 + 52.0 + 48.6)/3 = 45.2, because there are not equal numbers of cows at each of the dairy farms – the company would apply a weighted mean here.

**Geometric Mean**

Although very powerful, the arithmetic mean and median are not appropriate when the variable is a growth rate or rate of change, such as an investment over periods of time, or the change in production at a manufacturing facility. We need to use equation [4] to solve.

$Geometric Mean: \overbar{x}\_{g}= \sqrt[n]{(x\_{1})\left(x\_{2}\right)\left(x\_{3}\right)…(x\_{n})} = [\left(x\_{1}\right)\left(x\_{2}\right)\left(x\_{3}\right)…\left(x\_{n}\right)]^{^{1}/\_{n}}$ [4]

As way of example, say a firm assembled a product consisting of several parts. It needed to project its inventory needs in July based on the quantity demanded from its customers – see Table 2.

**Table 2:** Month-Over-Month Change in Demand

|  |  |
| --- | --- |
|  | Product A |
| January | 2.00% |
| February | 3.10% |
| March | 1.75% |
| April | 6.00% |
| May | 4.30% |
| June | 3.90% |

$$\overbar{x}\_{g}= \sqrt[6]{(0.02)(0.031)(0.0175)(0.06)(0.043)(0.039)} = 3.21\%$$

With this result, the production manager can project that it will need 3.21% more parts to fulfill the projected sales for July. However, the firm would want a much larger sample size for better accuracy, and to discover any seasonal changes – the greater the sample size, the more accurate the predictions become. We will address the appropriate sample size in an upcoming article of this journal.

Had the company used the standard mean calculation, they would have gotten: $\overbar{x}$ = (0.02 + 0.031 + 0.0175 + 0.06 + 0.043 + 0.039)/6 = 3.51% which would be incorrect. Even though it’s only off by 0.3%, that could be critical depending on the application.

**Median**

Another measure of central location is the median. Regarding the mean, there is a shortcoming when the data contains outliers – a value(s) that is abnormally higher or lower than the bulk of the overall group of data. For example, if one is calculating the mean of a group of home values, such as:

$700,000, $425,000, $550,000, $625,000, $495,000, $525,000, and $37,000,000.

We can see the outlier, namely $37M, would affect the mean quite dramatically. Outliers can occur for various reasons such as human error during data entry, gross errors in measuring the data, gross calculation errors, malfunctioning equipment, or many other reasons. The median is a method to avoid this issue. To write the formula for the median is far too cumbersome and outside the scope of this journal, therefore we simply state the rule as follows:

**Median**

 Arrange the data in ascending order (smallest to largest value).

 1. If the data set has an odd number of values, the median is the middle value.

2. If the data has an even number of values, the median is the average of the two middle values.

As an example, suppose we use the home value data above. If we were to calculate the mean, we would get a value of $5,760,000, which as we can see is far from the bulk of the data. Using the median…

 Arrange the data in ascending order: $425,000

 $495,000

 $525,000

 $550,000

 $625,000

 $700,000

 $37,000,000

 Since we have 7 values, i.e., an odd number, the median is the middle value, namely $550,000.

As an example of an even number of values. Say a company conducted seminars each month and the following was the number of attendees in each class: 55, 47, 44, 70, 62, 66, 59, 49, 9, 40, 76 and 61.

Arrange the data in ascending order (easily done in Excel with the Sort command):

9

40

44

47

49

**55**

**59**

61

62

66

70

76

Since we have 12 values, i.e., an even number, we take the average of the two middle values, namely 55 and 59. Therefore the median = (55 + 59)/2 = 57

In this data set, the mean of all the values, including the outlier value of 9, is 53.2. Again, we see that the median does a far better job of representing the bulk of the data.

**1.2 Measures of Variability**

In addition to measures of location, sometimes (often), measures of variability are necessary to get a true picture of the data. As an example, say a manufacturer of motorcycles outsourced its tires from two different suppliers and it wanted to analyze which had better delivery performance. They gathered a group of data of their past 100 shipments from Supplier A and Supplier B. What they discovered was that the mean delivery time for each supplier was 12.5 days, but does that tell whether the two suppliers are equally reliable? The histogram in Table 2 should tell you in an instant that they are not.

**Table 2:** Delivery Time of a Firm’s Two Suppliers

The above example is an illustration of variability and how critical it can be in the decision-making process. In manufacturing, and many other industries, it is vital to know when supplies will be delivered for the assembly line to not experience delays. It can also help with cash flow by eliminating the need to place orders weeks in advance to insure they will arrive when needed [**See associated Excel file**].

There are other important business measures of variability, however they are Not included in this Sample Issue.

**Introduction to Probability**

Sample Article Excerpt Taken from Journal Issue No.02

Michael J. Anderson

Business leaders and managers often base their decisions on uncertainties, for example:,

* What is the likelihood that sales will decrease if we increase prices?
* What are the chances that each customer that comes into the store will make a purchase?
* What is the probability that a work-related injury will not occur in the 1st quarter of the year?

So what is *probability*?

 **Probability** is the chance, likelihood, or possibility that a particular event will occur.

In all instances, the probability is a proportion (a fraction), whose value ranges between 0 and 1, inclusive. An event that has no chance of occurring, has a probability of 0. Conversely, an event that is sure to occur has a probability of 1. Unfortunately, in business zeros and ones are not very common – that would make our jobs far too easy, and frankly, quite boring. In the simplest case, where each outcome is equally likely, we define probability in equation [1].

$Probability of Occurence=P= \frac{Number of ways in which the event occurs}{Total number of possible outcomes}$ [1]

Ubiquitous to probability textbooks, let’s look at the six-sided die as an example. Each face has either 1, 2, 3, 4, 5 or 6 dots. Therefore, the number of ways that each number can occur is one, and the total number of possible outcomes is six (six-sided die). The probability of rolling any number, e.g., 6 is:

Probability of rolling a 6 = P(6) = 1/6 = 0.167 = 16.7%

Another requirement for assigning probabilities is that the sum of all the probabilities for all experimental outcomes must equal 1.0. In the die problem we have:

P(1 dot) + P(2 dots) + P(3) + P(4) + P(5) + P(6) = 1/6 +1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 1.0

1. **PROBABILITY DEFINITIONS, DETERMINING PROBABILITIES AND COUNTING RULES**

The fundamental definitions are for how outcomes of one or more random circumstances may be related to each other.

 **Event**: Each possible outcome of a variable is referred to as an event.

 A **simple event** is a unique possible outcome of a random circumstance.

For example, for the six-sided die, there are six simple events, i.e., each number is a unique outcome. An event can be any one of these simple events, a subset of them, or a set of all of them. For example, the event of an even number of dots consists of 3 simple events, namely two, four, or six dots.

 The collection of all possible events is called the **sample space**.

One event is the **compliment** of another event if the two events do not contain any of the same simple events and together they cover the entire sample space. We use the notation Ac to represent the compliment of event A.

An event either happens or it doesn’t. A person on your website will either buy something or they won’t. A piece of equipment will either fail or it will not. Symbolically, P(A) + P(Ac) = 1.

You can use the compliment to solve some simple probability questions. Say your data shows that one of your employees has a 0.95 probability that the parts they make are within the defined specifications. Therefore, using the compliment you can conclude that there is a 0.05 probability that he will produce a part outside of the specs (1 – 0.95 = 0.05).

It business it is common to have two events that do not contain any of the same simple events, and together do not cover the entire sample space – unlike complimentary events which do cover the entire sample space.

The *union* (∪) of A and B is the event containing all sample points that belong to *either* event A *or* event B or *both* is written as P(A∪B).

The *intersection* (∩) of events A and B are the sample points belonging to *both* A *and* B and is written as P(A∩B).

**Example 1:** Say the shipping department of a company analyzes it’s past 225 deliveries – the goal is to have the shipment arrive on-time, and without any damage to their product. At the conclusion of the analysis, the manager determined that 38 of the 225 shipments arrived late, 21 of the deliveries arrived damaged, and 9 of the shipments arrived late *and* were damaged upon arrival.

For the analysis, the manager defined:

L = the event in which the shipment arrived late

D = the event that the shipment arrived damaged

P(L) = 38/225 = 0.169

P(D) = 21/225 = 0.093

P(L∩D) = 9/225 = 0.040

With these probabilities in hand, the manager can now calculate what is the probability that a shipment will arrive either late *or* damaged, namely P(L∪D).

P(L∪D) = P(L) + P(D) - P(L∩D)

P(L∪D) = 0.169 + 0.093 – 0.040

P(L∪D) = 0.222

Thus, there is a 22.2% probability that a shipment will arrive either late or damaged.

**1.2 Rules for Counting, Combinations, and Permutations**

Recall from Equation [1] that to determine the probability of an event, we must know the total number of possible outcomes. Sometime this can be a bit of a challenge to determine, so here we will show you some formal methods.

Let’s begin with an example of tossing a two-sided coin twice, head (H) or tail (T). We want to determine the total number of possible outcomes of such an experiment.



**Figure 1:** Tree Diagram for Possible Outcomes for a Coin Tossing Experiment

As Figure 1 shows, there are four possible outcomes namely, (H H), (H T), (T H) or (T T), and we call this the sample space (S) of the experiment:

S = {(H H), (H T), (T H), (T T)}

For a more business-like example… Suppose a shop foreman hired a new staff member, and he wanted to analyze that person’s on time arrival to work each day during their first 14 days. Variables represented by: On Time (OT) and Late (L). The same tree diagram would apply with a new branch for each day. You are probably thinking, “the tree diagram will become **HUGE**,” and you would be exactly right. Therefore, we typically use equation [10] to solve such problems.

**Counting Rule No.1**

If any one of *k* different mutually exclusive and collectively exhaustive events on each of *n* trials, then the total number of outcomes is:

$Total number of outcomes= k^{n}$ [10]

If we apply [10] to the coin flipping example, k = 2 (Head or Tail), and n = 2 (we tossed the coin twice). Therefore, total possible outcomes = kn = 22 = 2\*2 = 4

In the second example above, k = 2 (On Time or Late), and n = 14 (14 day evaluation period). Therefore, total possible outcomes = kn = 214 = 2\*2\*2\*2\*…\*214 = 16,384. Quite a large tree diagram indeed.

**Note:** Other counting methods and when to use other techniques not included in this Sample Issue.

1. **BAYES’ THEOREM**

Thomas Bayes (1702-1761), an English minister and another giant in the field of statistics, so much so that we now have an entire branch of statistics called Bayesian statistics. **Note:** In other publications you will see Bayes’ Theorem called equivalently Bayes’ Law or Bayes’ Rule.

Bayes’ theorem is another **conditional probability**, i.e. the probability of A, given that B has already occurred. The derivation is done utilizing Equations [3], [4], [8] and [9]. After some algebraic manipulation and substitution, we arrive at the final theorem shown in Equation [14].

$P\left(B\right)= \frac{P\left(A\right)\*P(B|A)}{P(B)}$ [14]

Again, through mathematical manipulation, we can also solve the probability of B given A as in [15].

$P\left(A\right)= \frac{P\left(B\right)\*P(A|B)}{P(A)}$ [15]

The power of this interchangeability is that in the real-world of business, sometimes it is easy to find A given B, and other times it will be easier to find B given A – see example 4.

Let’s now put a few of the tools we have learned in this article, including Bayes’ theorem, to work for us with some examples. First let’s look at a non-business example.

**Example 4:** Say a married couple was going to have two children. What is the probability of having two girls *given* that they have at least one girl.

First, set your variables

G = Girl

B = Boy

1G = One Girl

2G = Two Girls

Next, we find the total number of outcomes

Using [10], with *k* = 2 (the couple is having two children), and *n* = 2 (each birth can produce either a Boy or a Girl).

kn = 22 = 4

The four combinations are: (GG), (GB), (BG), (BB)

Now let’s use [14] and rewrite it using our specific variables:

$$P\left(B\right)= \frac{P\left(A\right)\*P(B|A)}{P(B)}$$

$$P\left(at least 1G\right)= \frac{P\left(2G\right)\*P(1G|2G)}{P(1G)}$$

Using our combinations above, we solve the probability of having two girls - P(2G).

Of the 4 combinations above, we see that there is only 1 that fits, namely (GG). Therefore:

P(2G) = 1/4 = 0.25

As mentioned above, we us either use [14] or [15] based on which is easier to solve…

Since we are trying to solve the probability of the couple having one girl, given the fact that they have two girls. **Don’t over think it.** Simple logic says if they have 2 girls, then the probability of them having 1 is 100% (equivalently 1.0).

P(1G|2G) = 100% = 1.0

Next, we solve the denominator – the probability of having 1 girl

Again, we use our combinations and we see that there are 3 of the 4 that contain a girl. Therefore:

P(1G) = 3/4 = 0.75

Finally, we will plug all these values we solved into our rewritten equation

$$P\left(at least 1G\right)= \frac{P\left(2G\right)\*P(1G|2G)}{P(1G)}$$

$$P\left(at least 1G\right)= \frac{0.25\*1.0}{0.75}$$

P(2G|at least 1G) = 0.25/0.75 = 0.33

**Conclusion:** The probability of the couple having two girls given that they have one girl is 0.33 or 33%.

**Note:** This is only an example – we don’t suggest using it to plan your family, for there are countless variables involved in what sex a child will be.

Now let’s look at Bayes’ theorem with a more business-like example.

**Business Leader:** We hope the following example will spawn your thinking on how to apply Bayes’ method to your specific business application.

**Example 5:** Say a distributor sells a product that it purchases from two different suppliers - Supplier 1 and Supplier 2. Since this firm is very efficient in terms of analyzing its data, they have already determined that, when they get the product from Supplier 1, there will be 4.3% (0.043) of them that will be defective, and when from Supplier 2, 6.9% (0.069) will be defective. Written Probabilistically as:

P(Supplier 1’s part are defective) = 0.043

P(Supplier 2’s part are defective) = 0.069

The distributer is ready to package the product, so it has two buckets with 100 parts in each – Bucket A came from Supplier 1, and Bucket B, from Supplier 2. The packager randomly selects a part from the two buckets, until all 200 parts are used up.

**The question is**, if the end user of the product receives a defective part, what is the probability that it came from Bucket 1 (Supplier 1), and what is the probability it came from Bucket 2 (Supplier 2)?

First, set your variables

A = Select a defective part

B1 = The event that the part was chosen from Bucket 1

B2 = The event that the part was chosen from Bucket 2

Next, write out what we already know

We know the probability that if we have selected a defective part, it came from Bucket 1:

P(select a defective part | came from Bucket 1) = P(A|B1) = 4.3/100 = 0.043

We also know that the probability that if we have selected a defective part, it came from Bucket 2:

P(select a defective part | came from Bucket 2) = P(A|B2) = 6.9/100 = 0.069

Further, since we are choosing the parts at random from two buckets, we know that the part is equally likely to be drawn from Bucket 1 as it is from Bucket 2. Therefore:

P(B1) = P(B2) = 1/2 = 0.50

Now let’s start solving our unknown variables

What is the overall probability of picking a defective part:

For this we will use the fact that we have mutually exclusive events, picking a defective part or not,

we can use our formula for marginal probability above and add the intersected events.

Since we are looking at disjoint unions, the probability of disjoint unions, is the sum of those

Individual probabilities – written as:

P(selecting a defective part) = P(A) = P(A∩B1) + P(A∩B2)

To solve P(A) we use equation [3], rewritten as:

P(A∩B) = P(B)\*P(A|B)

Therefore:

P(A) = P(A∩B1) + P(A∩B2)

= P(B1)\*P(A|B1) + P(B2)\*P(A|B2)

To solve for P(A) we plug all the values we solved into our rewritten equation

P(A) = P(B1)\*P(A|B1) + P(B2)\*P(A|B2)

= 0.50\*0.043 + 0.50\*0.069

= 0.0215 + 0.0345

= 0.056

Now we can use Bayes’ theorem to find the probability that the defective part was from Supplier 1 or Supplier 2. Again, utilizing [14] we have:

$$P\left(A\right)= \frac{P\left(B\right)\*P(A|B)}{P(A)}$$

$$P\left(Part is Defective\right)=P(B1|A)= \frac{P\left(B1\right)\*P(A|B1)}{P(A)}$$

We now plug all of or previously computed values into this equation to get:

$$P\left(A\right)= \frac{0.5\*0.043}{0.056} = 0.384 = 38.4\%$$

Now we need to find the probability that the defective part came from Supplier 2, and there is an easy way and a more difficult way – FYI, mathematicians like the easy way (work smarter not harder).

The more difficult way is to use Bayes and replace the data for B1 with the B2 values and solve.

The easier way is to use the **compliment**, and as stated above, “you can use this method in countless business applications,” and here is one of them. Since the total probability must equal 1.0, we can use equation [7] as follows:

$$Probability that an event \left(A\right) does not occur=P\left(A^{c}\right)=1-p(A) $$

$$Probability that the defective part was not from Supplier 2, \left(B1\right)=P\left(B1^{c}\right)=1-p(B1)$$

$$P\left(B1^{c}\right)=1-0.384 = 0.616 = 61.6\%$$

**Conclusion:** If the end user receives a defective part, we discovered which supplier it came from:

P(Supplier 1) = P(B1) = 38.4%

P(Supplier 2) = P(B2) = 61.6%

In statistics, and mathematics in general, you want to ponder your answer to see if it makes sense, and our result does since Supplier 2 shipped more defects from the outset.

From a business perspective, we see a substantial difference between the two suppliers. Even though the difference in defects between the two was minimal (only 2.6%), the difference to the end user is 23.2%. This tells the firm that it needs to hold its suppliers to a certain threshold in order to minimize the number of defective parts that the end user will receive.

**Continuous Probability – Uniform and Exponential Distributions**

Sample Article Excerpt Taken from Journal Issue No.03

Michael J. Anderson

1. **UNIFORM PROBABILITY DISTRIBUTION**

Obviously the simplest of all the probability distributions – but don’t let that fool you in thinking you can’t get good results when using it.

To describe it, consider a random variable *x* that represents the driving time of a company freight truck from Los Angeles to Portland, Oregon. We know that the delivery time can be any value between 15 and 21 hours – since *x* can assume any value in that interval, we know the random variable is continuous and not discrete. The firm has sufficient delivery data that it can conclude that the probability of the delivery time of any 1-hour interval are the same as any other 1-hour interval. Since every 1-hour period is equally likely to occur, we say that *x* has a uniform probability distribution - see equation 1.

**Uniform Probability Density Function**

$$f\left(x\right)= \frac{1}{b-a} where a\leq x\leq b [1]$$

For the drive time above, *f(x)* = 1/(21-15) = 1/6 = 0.167, depicted in Figure 1.



**Figure 1:** Uniform Probability Distribution for Driving Time

As is common, the probability is simply the area under the graph, and in this case, since it’s a rectangle, the area is just the base multiplied by the height. Now the firm can answer questions such as, what is the probability that the delivery will take more than 19.5 hours?

**Solution:**

P(x) = Base(Height)

= (21-19.5)0.167

= 1.5(0.167)

= 0.251

Therefore, there is a 25.1% chance that the delivery will take more than 19.5 hours.

This simple illustration should show you that a firm can get solid answers, that will help with decision making, when its data follows a uniform distribution.

Two other useful formulas for the uniform distribution are the mean and standard deviation as shown in equations [2] and [3].

$$μ= \frac{a+b}{2} [2]$$

$$σ= \frac{b-a}{\sqrt{12}} [3]$$

1. **EXPONENTIAL PROBABILITY DISTRIBUTION**

The exponential distribution has been used extensively in evaluating arrival times, which is has a variety of business applications such as time between phone orders, hits on a website, and the time it takes to load a delivery truck. Recall that the normal distribution had two parameters, the mean (µ) and standard deviation (σ). The exponential distribution is a one-parameter distribution, namely lambda (𝜆) which represents the mean number of arrivals, and 1/𝜆 represents the average time between arrivals. For example, say a website averaged 25 hits per hour. Therefore, the average time between hits would be:

1/𝜆 = 1/25 = 0.04, i.e., every 0.04 hours (2.4 minutes), the firm’s website would get a hit.

For the exponential random variable, the mean (µ) and standard deviation (σ) are equal, i.e., µ = σ = 1/𝜆. If you plot a histogram or line chart of your data, and it resemble Figure 2, then your data is exponentially distributed.

**Figure 2:** Shape of an exponential distribution

the **exponential probability density function (pdf)** can be seen in equation [4].

**Exponential Probability Density Function**

$$f\left(x\right)= \frac{1}{λ}e^{-\frac{x}{λ}} [4]$$

 Where: 𝜆 = mean number of arrivals

 *e* = 2.71828

 *x* = any value of the continuous variable such that

 0 < x < ∞

As promised, no integral calculus here, however with it, we derive the following:

**Probability Associated with an Exponential Random Variable**

 If *X* is an exponential random variable, then:

$$P(X >x)= e^{-λx}$$

$P\left(X <x\right)=1- e^{-λx}$

$$P\left(x\_{1} <X <x\_{2}\right)=P\left(X<x\_{2}\right)-P\left(X<x\_{1}\right)= e^{-λx\_{1}}- e^{-λx\_{2}}$$

Let’s illustrate with an example:

**Example 1 [see associated Excel file]:** Say an auto parts manufacturer had a piece of equipment and its lifetime was exponentially distributed with a mean of 12 years. The equipment was in its 7th year, and for capital budgeting purposes the firm wanted to determine the following probabilities:

1. The likelihood that the equipment would last more than 12 years: P(X > *x*) = P(X > 12)
2. The likelihood that the equipment would last less than 12 years: P(X < *x*) = P(X < 12)
3. The likelihood that the equipment would last between 9 and 11 years: P(*x1* < X < *x2*) = P(9 < X < 11)

**Solution:** From above, the mean of the equipment is 12 years, therefore we need to solve for 𝜆:

1/𝜆 = 12 years

1 = 12𝜆

Finally, 𝜆 = 1/12 = 0.083 years

1. For question *a*, we use: P(X > 12) = $e^{-λx}$

= *e*-(0.083)12

 = 2.71828-(0.083)12

 = 0.3679

 = 36.79%

 So the equipment will have a 36.94% probability of lasting longer than 12 years.

1. For question *b*, we use: P(X < 12) = 1 - $e^{-λx}$

= 1 - *e*-(0.083)12

 = 1 - 2.71828-(0.083)12

 = 1 - 0.3679

 = 63.21%

 So the equipment will have a 63.06% probability of lasting less than 12 years.

1. For question *a*, we use: P(9 < X < 11) = $e^{-λx\_{1}}- e^{-λx\_{2}}$

= *e*-(0.083)9 - *e*-(0.083)11

 = 2.71828-(0.083)9 - 2.71828-(0.083)11

 = 0.0725

 = 7.25%

 So the equipment will have a 7.25% probability of lasting between 9 and 11 years.

With these results, the firm now has solid data to make a decision that is much more likely to result in success.

**Lean Six Sigma – Article Excerpts**

****

One of the two product offerings included in The Journal of Applies Business Statistics is the entire Lean Six Sigma DMAIC Training Course, and Certification if desired.

The material included in this training course is *all-inclusive*, and can easily be applied and streamlined to address any problem-solving issue. In addition, there is an abundant amount of *added material* included in this offer that is guaranteed to be *unmatched* by others (Automated Tools, Reference Material, Fast Track DMAIC’s, and *Support*).

In this article I have included three sample condensed excerpts from 2 of the 12 Journals being offered, along with their associated automated excel tool.

**Define Phase Journal #2 (Excel Value Stream Map included)**

**VALUE-STREAM MAPPING** [**See Associated Excel File Workbook**]

Value-stream mapping (VSM) is a lean processing technique used to analyze, design, and manage the flow of services, materials and information in your work process. The primary purpose of a value stream map is to displaying every vital step of your workflow, and evaluate if it brings value to your customer, both value-added and non-value-added steps. Those that do not add value to the customer should be eliminated.

Value-stream mapping allows you to analyze your process in-depth and provide insight into where changes can be made to improve the efficiency of your work flow.



**Measure Phase Journal #4 (Plug & Play tool included)**

**HISTOGRAM** **(Plug & Play tool included)** **[See Associated Excel Workbook File]**

A histogram is a common graphical tool that is used to display the distribution of data, it is a great tool because it brings your data to life, Histograms display the shape of your distribution, the central tendency of your distribution, the spread of values in your sample data set, and any outliers in your distribution.

The histogram is constructed by taking the difference between the min and max observation and dividing it up into evenly spaced intervals, or ranges of data, which are represented by vertical bars, similar in appearance to a bar graph. The number of observations in each interval are then counted and their frequency plotted as the height of each bar.

The x-axis represents the data ranges, and the y-axis represents the numerical count or percentage of occurrences in the data set. The higher the bar, the greater the frequency of the data values.



**Measure Phase Journal #4 (Plug & Play tool included)**

**TREND CHART [See Associated Excel Workbook File]**

Trend Charts, also known as Run Charts, or Time Series Charts, are usedto monitor and identify changes in your process with respect to time (trends going up or down, unusual patterns, and unusual occurrences.

**Analyze Phase Journal #6 (Plug & Play tool included)**

**BOX PLOT – WISKERS DIAGRAM [See Associated Excel Workbook File]**

A box plot, also known as a box and whisker diagram, is a basic graphing tool that displays the centering, spread, and distribution of a continuous data set. It is made up of a box and whiskers (and occasional outliers) that correspond to each fourth, or quartile, of the data set. The box represents the second and third quartiles of data. The line that bisects the box is the median of the entire data set-50% of the data points fall below this line and 50% fall above it. The first and fourth quartiles are represented by "whiskers," or lines that extend from both ends of the box. Boxplots use the concept of placing the data into quartiles (each quartile is 25% of the data).

Boxplots provide an instant picture of variation and some insight into search strategies for finding what caused the variation.



This Sample Journal were shortened excerpts of **over 100 topics** covered in the 12 journals, along with **over 50 automated Plug & Play Excel Tools** and loads of reference material.

A small group of the other topics and tools that will be covered throughout the 12 journals: Regression Analysis, Sampling, Analysis of Variance, Project Charter, Process Mapping, Control Charts, Process Capability, Cycle Time, Root Cause Analysis, Solution Prioritization, Detailed Kaizen Application, House of Quality, Risk Analysis, Control Plan, and many more.