SSD Global Solutions

IMPORTANT NOTE FROM SSD GLOBAL SOLUTIONS

Please read this entire intro before proceeding. As you know, to enter the LSS Black Belt program there is an assumption that you already understand basic math and statistics. However, we realize that many students have not been exposed to these disciplines recently. We also know that we have both stat savvy students and students who have taken only Stat 101.

Statistics is a large body of knowledge. What is covered here are the statistical concepts commonly used in Six Sigma and occasionally used in Lean Six Sigma when there is enough data. However, these concepts are on most industry standard tests.

Many statistical books are difficult or tedious to read and for that reason *Leaner* decided to select videos from industry leaders and former students to discuss the basic concepts. By design, we chose videos that offered follow up material. A great deal of the videos we selected offer complete statistical courses on-line at very reasonable prices.

Most presenters will cover all the suggested topics and you may elect to watch their series instead of our suggested one because you relate more to that speaker. This is perfectly acceptable since we are all covering the same body of knowledge. The reading material and questions related to the reading material will apply regardless of the video you watch. In many cases the video will not be necessary.

It is important to note that many of the statistical information in this text usually requires statistical software. Statistical software is so sophisticated, there may be specific instructions related to the particular stat program on how to get the desired results. For example, some statistical programs will use information from a simple spreadsheet to create a hypothesis test or design of experiment whereas other packages may have you enter information similar to a data entry form with field. Fortunately, all graphs, charts and calculation output look the same with the understanding that some packages offer a more cosmetic appeal.

Students not familiar with statistical software, but who are planning to use software in their LSS Black Belt projects, should allow ample time to learn the software. Software such as QI Macros and MiniTab have built in tutors but certain packages require taking an additional course or self-study to master. Use of statistical software is optional, in this program, and should be reserved for students who are using mega data. Most the more basic statistical function can be handled in MS-Excel.

Reviewing this material, reading and watching the videos, in order, will build on the knowledge from the previous material. Things that might not be clear will be clearer as you go through the material. The repetition is by design.

You may notice an over-abundance of links and videos to supplement the reading. Naturally, these are optional. Before starting this document, you may want to take a moment to review the

Q&A document. You may remember more about statistics than you think!

When you have completed Q&A portion please email Alex.Ramsey@SSDGlobal.net. The text material begins now. The Q&A document is a separate document. Please report any broken links to Kayli Bartlett at bartlettcontentsolutions@gmail.com.

TABLE OF CONTENTS

Introduction 4
Basic Statistics 4
MEASURES of CENTRAL TENDENCY 4
PROBABILITY 5
DISTRIBUTION MODELS 6
HYPOTHESIS TESTING 8
Confidence Intervals 8
CHI-SQUARED TEST 9
Standard Deviation 9
MONTE CARLO METHOD 10
T-TEST 10
ANOVA 10
DESIGN of EXPERIMENT 11

Introduction

Basic Statistics

The definition of what is meant by *statistics* and *statistical analysis* has changed considerably over the last few decades. The treatment of topics in this review are handled informally and should provide a fun overview of areas you should be familiar with to be a successful Lean Six Sigma practitioner. Our intent is to provide a wide-ranging and comprehensive, though not exhaustive, coverage of statistical concepts and methods used in Lean Six Sigma but more often in Six Sigma.

Please note that there are a vast number of books on statistics - Amazon alone lists 10,000 "professional and technical" works with statistics in their title. There is no single book or website on statistics that meets the need of all levels and requirements of readers.

Most of the videos and reading in this review are short. However, this first recommended video sets the stage as to what the statistical method is all about and is designed with accelerated learning in mind regarding “getting the big picture”.

A Quick and Basic Explanation that Covers Many Topics in This Material is a video by Doctoral Journal: (40 Minutes): https://www.youtube.com/watch?v=HpyRybBEDQ0

Other shorter suggested videos that provide a short introduction include:

Introduction to Statistics Anywhere Math (10 minutes):
https://www.youtube.com/watch?v=LMSyiAJm99g

Introduction to Statistics (5 minutes) SimpliPro: https://www.youtube.com/watch?v=MXaJ7sa7q-8

MEASURES of CENTRAL TENDENCY

The most common measures of central tendency are the arithmetic mean, the median and the mode. A central tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution.

Mean = Average
Mode = Number Most of Often in the data set
Median=After the data set is arranged sequentially it is the middle number
Range = in a data set subtract the highest number from the lowest number this is your range

VIDEO:

Your Teacher.com (2 minutes) https://www.youtube.com/watch?v=81zcjULlh58

Math.Help (3 minutes): https://www.youtube.com/watch?v=hX-BsyXGyb8

PROBABILITY

Probability theory is the branch of mathematics concerned with **probability**. ... As a mathematical foundation for statistics, **probability theory** is essential to many human activities that involve quantitative analysis of data.

In class you may remember we said the index for Sigma was 1-6; the index for CpK was 1-2; Correlation was -1 to +1. Probability is 0 to 1. A p-value is a number between 0 and 1.

When asked about the probability of a coin landing on heads, you would probably answer that the chance is ½ or 50%.

Imagine that you toss that same coin 20 times. How many times would you expect it to land on heads? You might say, 50% of the time, or half of the 20 times. So you would expect it to land on heads 10 times. This is the theoretical probability.

The theoretical probability is what you expect to happen, but it isn't always what actually happens. The table below shows the results after Sunil tossed the coin 20 times.

Outcomes	Frequency
Heads	13
Tails	7
Total	20

This shows the experimental probability. You can think of it as the probability determined from the results of an experiment. It is what actually happens instead of what you were expecting to happen.

The experimental probability of landing on heads is $\frac{13}{20} = \frac{65}{100} = 0.65 = 65\%$.

It actually landed on heads more times than we expected.

Now, Sunil continues to toss the same coin for 50 total tosses. The results are shown below.

Outcomes	Frequency
Heads	26
Tails	24
Total	50

Now the experimental probability of landing on heads is $\frac{26}{50} = \frac{52}{100} = 0.52 = 52\%$

The probability is still slightly higher than expected, but as more trials were conducted, the experimental probability became closer to the theoretical probability.

VIDEO:

Introduction to Probability Theory (don't forget to skip the commercial for the company unless you are interested) Tutor Guy (20 Minutes): https://www.youtube.com/watch?v=-8eSOmTPUbk

Introduction Probability Theory: Siraj Raval Fun Speaker (9 minutes)
https://www.youtube.com/watch?v=PrkiRVcrxOs

DISTRIBUTION MODELS

Probability distributions are fundamental to statistics, just like data structures are to computer science.

Things happen all the time: dice are rolled, it rains, and buses arrive. After the fact, the specific outcomes are certain: the dice came up 3 and 4; there was half an inch of rain today, the bus took 3 minutes to arrive. Before, we can only talk about how likely the outcomes are. Probability distributions describe what we think the probability of each outcome is which is sometimes more interesting to know than simply which single outcome is most likely. They come in many shapes, but in only one size: probabilities in a distribution always add up to 1.

For example, as demonstrated, flipping a coin has two outcomes: it lands heads or tails. Before the flip, we believe there's a 1 in 2 chance, or 0.5 probability, of heads. The same is true for tails.

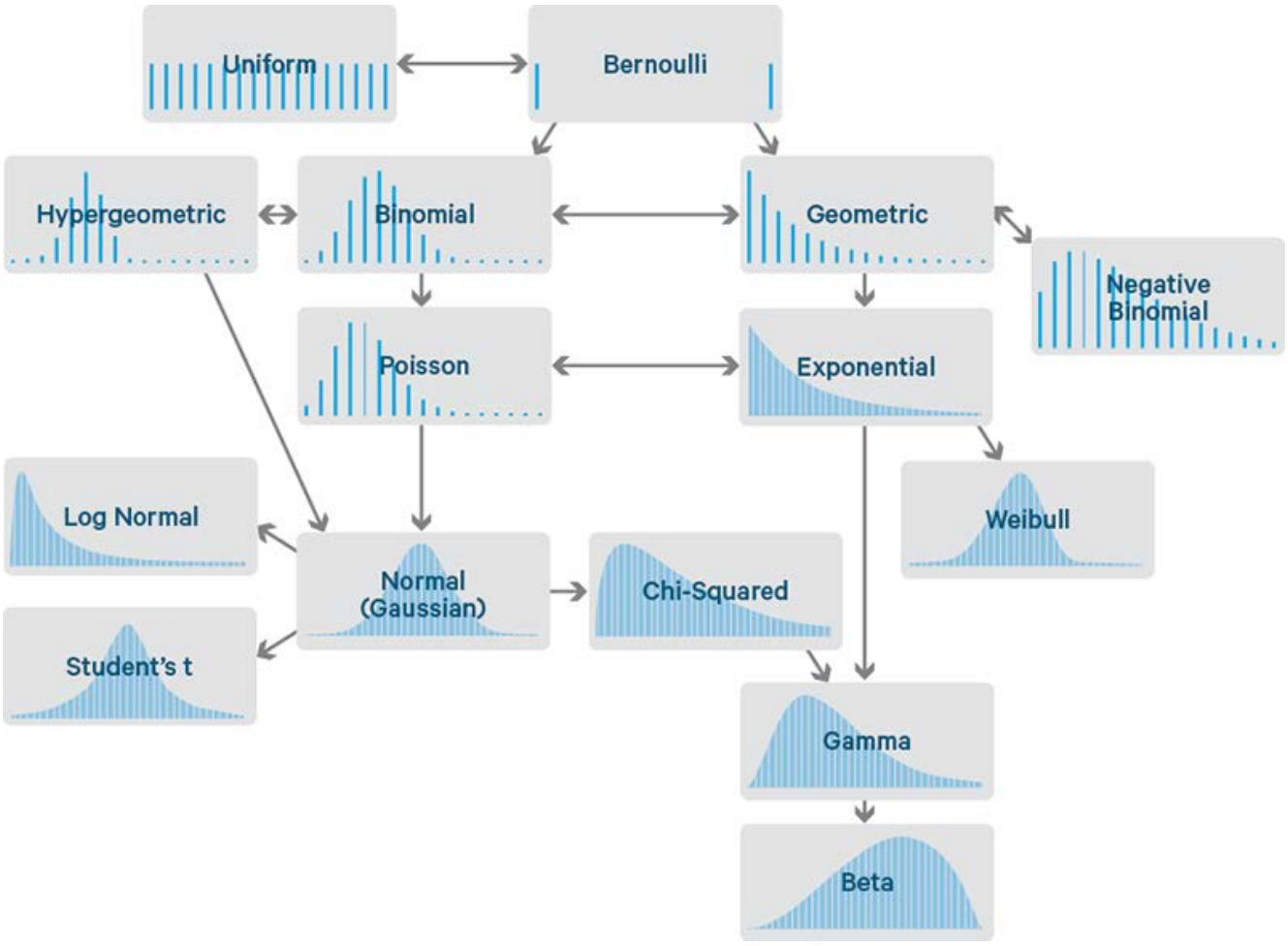
Videos:

Normal Distribution: Explain Simply How2Stats: (5 minutes)
https://www.youtube.com/watch?v=xgQhefFOXrM

Binomial made Easy (5 minutes) Randall Heyman:
https://www.youtube.com/watch?v=Bb0wZwjhEWY

Bernoulli (6 minutes) Kahn Institute:
https://www.youtube.com/watch?v=O8vB1eInP_8

Despite exotic names, the common distributions relate to each other in intuitive and interesting ways that make them easy to recall. The ones typically addressed in LSS is Normal, Bernoulli and Binomial



SAMPLING

In statistics, quality assurance, and survey methodology, sampling is concerned with the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population.

The following sampling methods are examples of probability sampling:

- Simple Random Sampling (SRS)
- Stratified Sampling.
- Cluster Sampling.

- Systematic Sampling.
- Multistage Sampling (in which some of the methods above are combined in stages)

Video

Sampling – the most popular types (3 minutes) StatHelp:
https://www.youtube.com/watch?v=be9e-Q-jC-0

HYPOTHESIS TESTING

A statistical hypothesis, sometimes called confirmatory data analysis, is a hypothesis that is testable on the basis of observing a process that is modeled via a set of random variables.

A statistical hypothesis test is a method of statistical inference. Commonly, two statistical data sets are compared, or a data set obtained by sampling is compared against a synthetic data set from an idealized model. A hypothesis is proposed for the statistical relationship between the two data sets, and this is compared as an alternative to an idealized null hypothesis that proposes no relationship between two data sets.

Videos

Introduction to Hypothesis Testing JB Stats (10 minutes):
https://www.youtube.com/watch?v=tTeMYuS87oU

Simple Hypothesis Testing (10 minutes) Kahn Academy:
https://www.youtube.com/watch?v=5D1gV37bKXY

Type I and Type II Errors in Hypothesis Testing

In statistical hypothesis testing, a type I error is the incorrect rejection of a true null hypothesis (a "false positive"), while a type II error is incorrectly retaining a false null hypothesis (a "false negative").More simply stated, a type I error is the (false) detection of an effect that is not present, while a type II error is the failure to detect an effect that is present.

Video

Statslecture (5 minutes): https://www.youtube.com/watch?v=FHT6e_mdGoU

Bionic Turtle (10 minutes): https://www.youtube.com/watch?v=taEmnrTxuzo

Confidence Intervals

In statistics, a confidence interval (CI) is a type of interval estimate (of a population parameter) that is computed from the observed data. The confidence level is the frequency (i.e., the proportion) of possible confidence intervals that contain the true value of their corresponding parameter. In other words, if

confidence intervals are constructed using a given confidence level in an infinite number of independent experiments, the proportion of those intervals that contain the true value of the parameter will match the confidence level.

So, if your significance level is 0.05, the corresponding confidence level is 95%. If the P value is less than your significance (alpha) level, the hypothesis test is statistically significant. If the confidence interval does not contain the null hypothesis value, the results are statistically significant.

Video

Understanding Confidence Intervals: Statistics Help (4 minutes):
https://www.youtube.com/watch?v=tFWsuO9f74o

Statistics Made Simple – Confidence Intervals (5 minutes):
https://www.youtube.com/watch?v=Hn6C21GC0vA

CHI-SQUARED TEST

A chi-square test, also written as test, is any statistical hypothesis test wherein the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test.

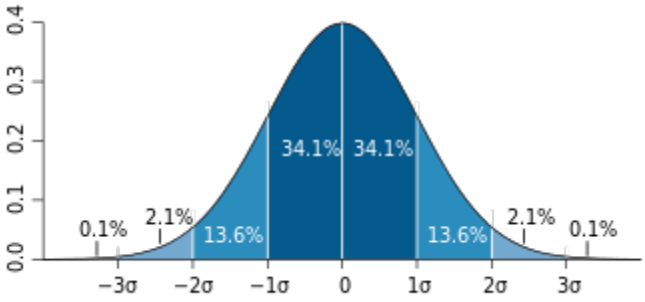
Easy to understand video, Kahn Academy: (10 minutes)
https://www.youtube.com/watch?v=2QeDRsxSF9M

Even a simpler Chi-Squared, J Eisenberg (5 minutes)
https://www.youtube.com/watch?v=VskmMgXmkMQ

Standard Deviation

In statistics, the standard deviation (SD, also represented by the Greek letter sigma σ or the Latin letter s) is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

A low standard deviation indicates that the data points tend to be close to the mean (also called the expected value) of the set, while a high standard deviation indicates that the data points are spread out over a wider range of values.



Deviation just means how far from the normal -- The Standard Deviation is a measure of how spread out

numbers are. It is measurement of spread.

Video: Standard Deviation Made Super Easy, (4 minutes)
https://www.youtube.com/watch?v=HV2HtOBSqjo

Example of SD, Jeremy Jones: (2 minutes): https://www.youtube.com/watch?v=MRqtXL2WX2M

MONTE CARLO METHOD

Monte Carlo methods (or Monte Carlo experiments) are a broad class of computational algorithms that rely on repeated random sampling to obtain numerical results. Their essential idea is using randomness to solve problems that might be deterministic in principle.

We are sticking with coins just to bring home the point:

Video

Monte Carlo Simulation of a Coin Toss: (3 minutes):
https://www.youtube.com/watch?v=NSLfoHT4bR8

How it Works: (5 minutes) Risk Management: https://www.youtube.com/watch?v=NxgVBeTfAio

If you are interested this is a Step-by-Step: (12 minutes) Trading.com:
https://www.youtube.com/watch?v=3gcLRU24-w0
(this is a common application in stocks and trading).

T-TEST

A t-test is an analysis of two populations means through the use of statistical examination; a t-test with two samples is commonly used with small sample sizes, testing the difference between the samples when the variances of two normal distributions are not known.

In statistics, the number of degrees of freedom is the number of values in the final calculation of a statistic that are free to vary. The number of independent ways by which a dynamic system can move, without violating any constraint imposed on it, is called number of degrees of freedom.

InfoL, (2 minutes) What is it? https://www.youtube.com/watch?v=DQaR-5vV4TY
Example: Tess St. John (4 minutes) https://www.youtube.com/watch?v=ZY5XxJ2aJNc

An F-Test is comparing more than two columns of data.

Degrees of Freedom, Studyable, (6 minutes): https://www.youtube.com/watch?v=VIlVWeUQ0vs
Why are Degrees of Freedom used in SD? Statistics Fun (7 minutes):

ANOVA

Analysis of variance (ANOVA) is a collection of statistical models used to analyze the differences among group means and their associated procedures (such as "variation" among and between groups), developed

by statistician and evolutionary biologist Ronald Fisher.

Introduction to ANOVA – No Math (7 minutes) Jackson University:
https://www.youtube.com/watch?v=qV-WoquC4dA

Example of using an ANOVA: (5 Minutes) J. Eisenberg:
https://www.youtube.com/watch?v=ITf4vHhyGpc

DESIGN of EXPERIMENT

Design of experiments (DOE) is a systematic method to determine the relationship between factors affecting a process and the output of that process. In other words, it is used to find cause-and-effect relationships. This information is needed to manage process inputs in order to optimize the output.

If I only had two things to compare you might use a hypothesis test.

The term experimental design refers to a plan for assigning experimental units to treatment conditions. A good experimental design serves three purposes. Causation. It allows the experimenter to make causal inferences about the relationship between independent variables and a dependent variable.

An award-winning article can be found here: https://www.isixsigma.com/tools-templates/design-of-experiments-doe/design-experiments-%E2%90%93-primer/

This ASQ link provides explanation and has a hands on exercise: http://asq.org/learn-about-quality/data-collection-analysis-tools/overview/design-of-experiments.html

Video

Experiments Reminder, Brainstem, (8 minutes): https://www.youtube.com/watch?v=tK2mBsSb3uw

8th Grade Example of Using a Design for Experiment: Michell Mitchell (10:45)
https://www.youtube.com/watch?v=Wd1Ul3bg188

Kahn Academy Example - DOE (10 minutes): https://www.youtube.com/watch?v=DaBq0naj0YY