The Effectiveness Use of Statistics

George Isaac[1], Ragy Isaac[2]

Keywords: Statistics, Data Analysis, Independence, Sampling, Variation

Abstract: Statistical independence is one of the most important statistical ideas and is required in order to generalize a conclusion from a sample to a population. The estimated differences between averages for phenomena of interest take center stage and much less attention is devoted to variation, however. The study of variation is a key aspect to generalizing results and making predictions. This paper addresses concepts of variation in statistics and attempts to simplify various statistical ideas. Two cases will be presented for the need of statistical understanding in day-to-day life garnered by an enhanced understanding of statistical independence and variation. An industrial example and a lithographic example are used to demonstrate the steps needed to analyze seemingly innocuous dataset and how the application of these concepts is crucial.

Introduction

Frequently conclusions are easily drawn from a well-designed experiment, even when rather elementary methods of analysis are employed. Conversely, even the most sophisticated statistical analysis cannot salvage a badly designed experiment. - George Box [1]

Even more important than learning about statistical techniques is the development of what might be called a capability for statistical thinking. - George Box [1]

This paper was written for those who collect and try to make sense of data and was designed to be understood by data analysts, industrial experimenters, scientists, and engineers. Every effort was made to avoid mathematical equations. The goal of this writing is to educate those who have not had much statistical experience while also correcting and updating the knowledge of those who may have incorrectly used statistics in the past.

Applied statistics is an integral part of the information gathering and learning process for any industry. Statistics is the science of describing past events and predicting

[1]Continuous Process Improvement, Dow Jones
[2]Quality, Goss International

those that may come in the future. By gathering and analyzing data with a set of tools, predictions about the future may be made, allowing for informed decisions. Furthermore, statistics is a science that attempts to differentiate experimental error in a system from the change caused by a phenomenon. A statistical thinker always embraces errors as the focus of an analysis to generate a concrete answer from a dataset riddled with random error or variation.

A statistical mentality is quite applicable to day-to-day life: familiarity with the language of statistics allows us to critically appraise and interpret the world around us. Statistical terminology is encountered frequently in the media. Estimates, significance, projections, averages, median, margin of error, scientific poll, non-scientific poll etc. are all statistical concepts and terms. For example, statistics can help us understand the meaning of non-scientific poll. This refers to a poll in which the respondent selection is biased; only people who read or watch a specific media provide their opinion. It should be concluded that the story is propaganda and bad science. The poll does not reflect the opinions of the general public, and the results should be discounted. Similarly, the same concept applies to product sampling during testing as will be explained in the next sections.

Statistical Independence

The critical principal that has been violated in a non-scientific poll is that of statistical independence. Statistical independence means that the opinion of one respondent does not affect the opinion of any other respondent. A pollster is attempting to represent an entire population. Say the pollster has access to and surveys only a specific special interest group at a conference. In this case, every member of the sample group will have similar voting patterns for that particular topic, especially if they are free to discuss their opinions with each other after the question is asked. Their opinions will be similar because they are now statistically dependent. This is common sense; the opinion of a group that has the same interests will be similar. Votes derived only from a special interest group do not represent the unbiased opinion of the whole population.

The purpose of an opinion poll is to ask questions of a small sample of a population of people who can represent the attitudes, opinions, or projected behavior of an entire population (Figure 1). In effect, the fundamental goal of an opinion poll is to come up with the same results that would have been obtained had every member of a population been interviewed. The key to reaching this objective is equal probability of selection. Every member of a population must have an equal chance or probability of being selected in a sample. The sample will then be representative of the population, and statisticians call it statistical independence or independent sampling.

Figure 1: *Small independent sample can represent the opinions of an entire population*

Testing a product or a new procedure and making an inference on future performance is similar to polling. The conclusion from a relatively small sample can predict performance if the sampling is performed correctly. The conclusion reached would be biased without statistical independence.

For an opinion poll or any product testing to be successful, every respondent must have an equal chance of being included in the sample; this is the key. Polling science is highly successful in predicting the outcome despite the fact that members of a polling sample are diverse and that there is always natural variation in any sample.

Variation

Variation is an inevitable fixture of life. Those that use statistics must constantly be aware of how variation affects all decision-making. While variation will always be present in any system, it can be reduced. This reduction can improve the quality of a product, lower its customer complaints, and reduce production waste.

Figure 2: *Variation makes it difficult to draw the right conclusion*

Statistical thinking always embraces the concept of variation as an observable reality, present everywhere and in everything. In order to determine whether a change observed in the result of a process can be explained by a modification deliberately applied or not must take variation into account (Figure 2). This will be clearly shown in the industrial experiment below where a plant manager wants to increase the production yield by switching from his standard process A to an assumed better process B. Variation is the reason why sophisticated statistical methods are developed to filter out the noise from the variable of interest. The success or failure of any statistical analysis is hinged on the correct determination of the magnitude of noise since noise is used as the scaling factor.

The occurrence of random variation or noise can be demonstrated with an adaptation of the EP Box paper helicopter [2] experiment. In this experiment, the paper helicopter is dropped ten times to the floor under as similar conditions as possible in an attempt to hit an arbitrary target such as coin on the floor (Figure 3). The distance between the bottom of the helicopter at its falling spot on the floor and the coin represents the **unexplained** variation or **random variation**. The point where the helicopter hits the floor is schematically represented by the blue dots. The exact same process is being carried out, but the random variation inherent in the process causes the airplane to land in different spots each time. When a partner with a notebook or fan creates a wind draft next to the flight path of the helicopter, it creates a different kind of purposeful variation. This is demonstrated by the red dot. This purposeful variation is analogous to an action to improve a process.

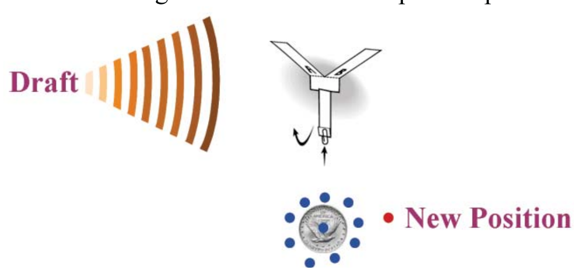

Figure 3: *George Box helicopter experiment*

Variability Decomposition

In order to obtain more precise results, data should be analyzed by decomposing variation into its different sources. Because every process has inherent variation, it can be difficult to differentiate between the normal variation of a system and variation due to a change applied. In order to do this, the variance should be compared to what chance alone would yield. It can be visualized that interference with the process, such as fanning the helicopter 3, produces different kinds of variation. These are called **explained**, **controlled**, or **effect** variation. Data analysts generally attempt to find causes for this explained variation, referred to as the **signal**. Uncontrolled variation, or **noise**, is what is leftover once all the patterns or signal have been removed. Noise, by definition, is the variation when no patterns or signal can be found. The day-to-day data is always a combination of signal in the presence of noise. The key to a successful analysis is to separate the signal from the noise while being careful not to conclude a pattern where none exists.

Figure 4: *The snake is invisible if its skin pattern matches its surrounding*

Mathematicians and data analysts exploit the idea of signal versus noise (or variation) in many industrial applications such as artificial intelligence, machine learning, and pattern recognition. To demonstrate, imagine that a snake (Figure 4) is on a rock. The skin of the snake has a pattern. If this pattern matches that of the rock upon which the snake is situated, the snake will be invisible. The same applies to data. If the signal that is being searched for matches the noise of the system, the signal will not be seen. Applied to industry, if a new raw material is to be tested and the process is highly variable, it might not be possible to conclude concretely if the new material is better or worse. If the noise is high, the signal might get lost.

Most of data analysis and statistical procedure is focused on understanding variation induced by special causes, such as fanning the paper helicopter or testing new materials or procedures. This can be called **treatment**, since it was purposefully imposed on the system. This produces extra variability that is different from the variability caused by the noise (Figure 5). The treatment could have been trialing a better design for paper helicopter to resist the wind! The existence of the random variation is a nuisance that disguises the treatment. In order to determine the effect of the treatment, the noise must be investigated.

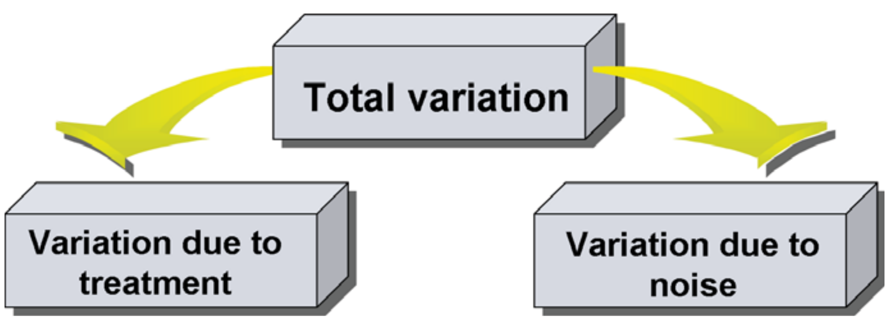

Figure 5: *Separate noise from treatment; random variation is a nuisance that disguises the treatment*

Industrial Experiment

Throughout this paper, a practical industrial experiment, given by George Box in his book *Statistics for the Experimenters*, is used to demonstrate critical statistical thinking. It is a simple example with far reaching applications frequently encountered in many industrial situations. The industrial question is as follows: a plant manager wants to increase production yield. Process A represents his standard production process and process B is assumed to be the improved process. Does process B produce a higher yield than process A?

The plant manger performs the following comparative experiment: ten consecutive batches were manufactured using process A, produc- tion was changed afterwards to process B, and ten more consecutive batches were produced.

Analysis Using External Dataset

Typical data analysis is performed on groups, not individual data points. Groups are treated as the observations in the analysis. The test data and historical data of ten individual batch yields are aggregated into group means. The follow represents the sampling that the plant manager accomplished:

- Two ten consecutive batch experiment
- Groups of the two ten batches are averaged to obtain the groups sample means
- Groups sample means were subtracted from each other to produce what is termed **experimental difference**
- 210 historical yields
- Divided into groups of ten consecutive batches to be similar to the two ten-batch experiment
- Groups of ten consecutive batches are averaged to obtain the group means
- Groups means were subtracted from each other to produce what is termed as historical reference

Descriptive and Inferential Statistics

In any data evaluation, analysis should begin with a picture showing the collected raw data using appropriate, descriptive graphs, depending on the data type. These pictures are referred to as **descriptive statistics**. In fact, statistics as a science can be divided into two branches: descriptive and **inferential statistics**. Inferential statistics are concerned with making predictions, or inferences, about a population from observations, such as predicting if process B has higher yield than process A. Simply put, this branch of statistics evaluates a sample and attempts to generalize the results to the much larger population allegedly represented by the sample. For descriptive statistics, if a pattern over time is being investigated, scatter-plots (Figure 6) and (Figure 7) are used. These plots have time plotted on the x-axis and the data value (yield) plotted on the y-axis. For patterns unrelated to time, frequency plots are used. Examples of these are histograms (Figure 8), stem-and-leaf plots (Figure 9), and box-and-whiskers plots (Figure 7) superimposed on a scatter-plot.

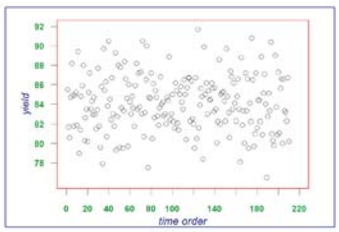

Figure 6: *Plot of 210 yields from past batches*

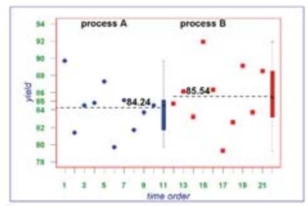

Figure 7: *Yield values in time order for comparative experiment*

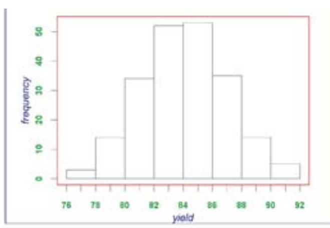

Figure 8: *Histogram of 210 obs. from industrial process*

Figure 9: *Stem-and-leaf plot of 210 yields from past batches*

Experimental Difference in Means

The experimental difference between two processes is calculated by subtracting the ten-batch average yield of process A from the ten- batch average yield of process B and is found to be equal to 1.3 units. In the case of the plant manager, higher yield is considered desirable.

After determining an experimental difference of 1.3 units, it is necessary to count how many times a 1.3 unit difference appears in the 210-batch historical reference. This will indicate whether process B does have a positive impact on the production yield. If a difference of 1.3 units rarely occurs in the historical reference, then process B is probably better than process A. If the difference is more prevalent, the plant manager may conclude that the difference is nothing more than the expected variation of the typical production process. In other words if the experimental difference is a common occurrence in the historical reference then process A and process B are practically the same. The conclusion would be that investing to change from process A to process B is a waste of money.

The historical reference shows that only 4.7% of the data are larger than or equal to 1.3 units. This is rare, and therefore it should be suspected that process B produces higher yield than process A. The historical reference dataset provided all the information needed to make a judgment on how the difference between process A and process B should be viewed. It shows the probability that the results of the analysis could have occurred only by chance. In other words, the historical reference provided a reference dataset or distribution to evaluate how significant the average between means should be considered.

Cost of Formal Statistical Procedures

In the above analysis, no formal statistical procedures were used, only simple addition and division. The use of statistical procedures comes with restriction on the collected data, such as statistical independence, that must be obeyed. Statistical independence in the production example requires that the yield of any batch does not affect the yield of any other batch. It caries the same above mentioned idea of the need for independence in polling. If the historical yield values are independent of each other, the occurrence of one yield does not affect the occurrence of the other. The 210-entry industrial historical data cannot assume independence of yields; remnant product in the vat assures those consecutive batches are dependent. Likewise, the experimental data A and B cannot claim independence.

Similar thinking can be expanded to many industrial applications. In the printing production world, serial dependence is everywhere. Consecutive printed samples, collected from a given press-folder, are guaranteed to be serially dependent. Formal statistical procedures that depend on independence cannot be used on this type of data without further manipulation and/or precautions to induce independence. Violation of these restrictions can have a detrimental effect on the validity of the results and sometimes cause the wrong conclusion to be reached.

One frequently used formal statistical procedure, the t-test, puts considerable restriction on the collected data. The t-test is not appropriate to make a judgment on how significant the experimental differences in the plant experiment are since the collected data are not independent. If the t-test is assumed, incorrectly, to be appropriate in the above industrial example, it would demonstrate whether the experimental difference is unlikely to have occurred because of random variation without the need for historical data. This same exact question was answered by the use of the historical reference.

In many industrial situations, it is not always possible to have relevant and reliable historical data to judge the significance of the experimental difference. In the absence of relevant and reliable historical data, experimenters typically use the t-test to test the significance of the experimental difference and assert that the tested samples represent future performance. The t-test uses only the samples collected during the test. This type of data is called an **internal dataset** since it is collected strictly from the sample data.

If the plant manager uses the t-test, without the knowledge that the data is not independent, the result of analysis will indicate that the **experimental difference** occurs 19.5% of the time during typical production. It shows that the difference is more common and conflicts with the prior conclusions of 4.7% using the historical reference data. This is at the core of determining the appropriate analysis for a test to reach the correct conclusion. The preceding case violates the independence assumption of the t-distribution. The two ten- batch samples from process A and process B were treated as independent random samples, but they certainly are not. They are serially correlated samples. Consequently, the wrong conclusion is reached.

Absence of Historical Data

The plant experiment analysis so far relied exclusively on the historical reference data to assess the significance of the experimental difference. If the historical reference data does not exist and if the plant manager followed the same procedure to perform the experiment, no correct conclusion can be reached. There will be no clear direction of what process the plant be following to improve yield.

If the plant manager knows that he does not have historical data and the t-test requires independent data then he must change his testing procedure to induce independence. He would assign the labels ’process A’ and ’process B’ randomly to the test to be performed and chooses which process to test accordingly. The t-test can then be used to test the significance of the difference between processes A and B. The toning example below, in the Lithographic Experiment, demonstrates the proper sampling to induce independence. With independent sampling there is no need for historical data. The two statistical analysis tools: Descriptive and inferential analyses are also demonstrated as well.

Lithographic Experiment

Modern offset lithography utilizes a flexible, grained, and anodized aluminum image carrier covered with a photosensitive polymer layer. The image to be reproduced is created with an appropriate laser to harden the image location. The un-imaged soft photosensitive polymer is removed with a developer, exposing the grained and anodized aluminum layer. This aluminum image carrier is mounted on a lithographic press to reproduce the desired image on paper or film. During the printing process, two fluids are needed for the reproduction: a lithographic **ink** and an aqueous dampening solution known as **water**. The water acts as a barrier, disallowing the ink from adhering to any location that does not have an image, known as **non-imaged area**.

Figure 10: *Examples of toning - Heavy toning on the left, light toning in the middle, and no toning on the right*

One of the issues encountered during lithographic printing is known as **toning** (Figure 10). This occurs when undesirable faint image appears on the non-image areas. The following practical example shows the proper use of statistics to compare the magnitude of the undesirable toning associated with two different inks, A and B.

Methodology

In the absence of historical data, randomness must be utilized for the t-test to be valid in comparing ink A and ink B. Systematic, random, multi-stage sampling were used since the production of printed papers is sequential. Randomized paired sampling for ink A and ink B is used to minimize the effect of unforeseen variables. Ink A and ink B were simultaneously tested on four press units with the same dampening solution on actual press production for five days and paired samples were collected. It must be noted that sampling of ink A and ink B are obviously sequential. Independence was induced by partitioning each press run into mutually exclusive intervals spanning 5000 printed copies. A random number generator determined when the four sets of samples were to be pulled from each interval. Samples from ink A and ink B were collected at the sample press-speed point. This minimizes the effect of time-dependent errors due to extraneous variables such as humidity, press temperature, and others. Often precision can be increased and bias be reduced by comparing matched pairs of samples while randomization ensures the validity of the test.

The ink density of plain paper in the non-imaged area was measured directly next to the sample spot. The toning value was defined as the difference between blank imaged and adjacent non-imaged area. Any values larger than 0.02 optical density units were considered a defect.

Within the collected samples, four sheets, two with the ink A and two with the ink B, were selected and the optical densities were measured at eight different positions across the sheet, one in every other inking column, using handheld densitometers.

The optical densities of the paper at each position were then subtracted from the toned density to determine toning defect.

Results

As mentioned, the analysis should always begin by looking at the descriptive statistics. The histogram in (Figure 11) shows that ink A and ink B came from completely two different populations with no overlap. The same can be concluded from the boxplots in (Figure 12). Inferential statistics also conclude that the toned density of ink A is higher than the toned density of ink B with 100% certainty.

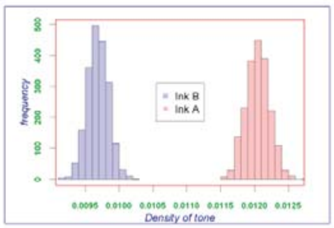

Figure 11: *Histograms of toning: ink A and ink B*

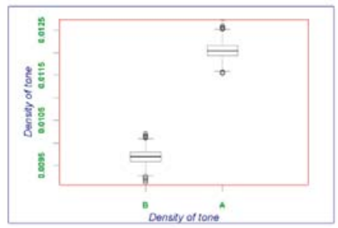

Figure 12: *Boxplot of tone-densities of ink A and ink B*

Miscellaneous Raw Data Validation and Manipulation

The data should follow a standard format with variables represented as columns and cases as rows in a spreadsheet. Only one data table should be placed in each sheet, preferably in a comma separated variable (.csv) format. It is very helpful if every column has only one header that starts at the first row.

When approaching data for analysis, it is crucial that the data is validated or checked for accuracy. All data outside the main pattern (outliers) must be investigated. If the outliers are proven to be erroneous, they must be removed from the data. Outlier detection can be easily done by reviewing the datas scatter-plots.

	Raw Data		Data - Mean		(Data - Mean)/StDev	
	Celsius	Fahrenheit	Celsius	Fahrenheit	Celsius	Fahrenheit
	0.00	32.00	-20.00	-36.00	-0.51	-0.51
	2.00	35.60	-18.00	-32.40	-0.46	-0.46
	4.00	39.20	-16.00	-28.80	-0.41	-0.41
	6.00	42.80	-14.00	-25.20	-0.36	-0.36
	8.00	46.40	-12.00	-21.60	-0.31	-0.31
	100.00	212.00	80.00	144.00	2.04	2.04
Mean	20.00	68.00	0.00	0.00	0.00	0.00
StDev	39.29	70.73	39.29	70.73	1.00	1.00

Figure 13: *Standardizing shows that two identical datasets, with different scales, are identical*

Quite often, variables that are measured with different scales need to be compared. This can be helpful, for example, if a manager attempts to evaluate the performance of multiple departments with varying outputs. Variables measured at different scales or units of measurement (Fahrenheit, Celsius, pounds, feet, gallons, etc.) should be standardized before any analysis is carried out to equalize the range and data variability (Figure 13). This is accomplished by subtracting the data mean and dividing by its standard deviation. To demonstrate this concept, start with two identical temperature data sets presented in the Fahrenheit and Celsius scales. The first two columns show the raw data. The third and fourth columns represent the mean subtracted from the respective data point. It can still be seen that columns three and four appear different. Columns five and six represent the mean subtracted from the respective data point then divided by its standard deviation. Columns five and six appear identical. Subtracting the mean and dividing by the standard deviance, known as data standardizing, revealed the truth about the two data sets; it clearly demonstrated that they are identical. Data standardizing equalizes the data range and variability, a very important, simple, and powerful manipulation.

Conclusions

- The ultimate focus is future performance. A test by itself is useless and a waste of time if it does not correctly predict future performance. In the same way, an opinion poll is only successful if it predicts the future accurately. When applied to the printing world, the ability to predict future performance will allow for the best possible copies to leave the press-folder during every run.

- In any analysis, the sample averages are compared, not their individual members. The test analysis compares the difference in means and evaluates if this difference can be due to natural variability or due to the modification

- The average difference, or the treatment, is always compared to the natural variation or noise

- Simple analysis comparing the difference between the averages requires long sequence of relevant previous records or indepen- dence through random sampling [1]

- Independence should never be assumed, it requires hard work and planning

- Important testing precautions must be taken to ensure independence if no historical dataset is available

- Although the above mentioned plant experiment had relevant 210 consecutive batches of process A, ten new consecutive batches of process A were tested, immediately followed by ten consecutive batches of process B. This ensures that both processes are tested under almost the same conditions and minimizes the effect of unforeseen lurking variables

- The historical data of the standard process were manipulated to be similar to the test data of two ten batches. Only then the historical data can be useful when comparing the standard and the new processes

- Validate and evaluate the health of the data before performing data analysis. Outliers can skew averages and estimated variation

- Data standardizing equalizes the data range and variability. It is a very important, simple, and powerful manipulation to allow proper comparison

Acknowledgements

We would like to acknowledge and express our appreciation to Paul Cousineau for his input, knowledge, and constructive critiques. Additionally, we would like to thank Sarah Isaac and Christina Andre for their technical knowledge and editing support.

References

[1] Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building. John Wiley & Sons, Inc., 1978.

[2] George Box. Teaching engineers experimental design with a paper helicopter. The Center for Quality and Productivity Improvement, UNIVERSITY OF WICONSIN, 1991.