Statistical Confidence

Table of Contents

Introduction

This START sheet discusses the practical aspects of statistical confidence, i.e., the confidence that we place (from a probabilistic point of view) on a statement about some random phenomenon of interest. Statistical confidence gives us the probability that the statement given is true. For example, we can state, "we are 95% confident that the product defect rate is between 5% and 10%." Here, our statistical confidence is 95%, i.e., 19 out of 20 times this statement will be true.

In general, confidence is the internal conviction that some situation or statement holds true. For example, we want some confidence that a certain stock we seek to buy is solid and will not lose its value in a short time. On the other hand, such confidence stems from our long and hard study of the stock market and the company in question, or, equivalently on the sound judgment and experience of our stockbroker. Sound judgment and experience is the foundation of confidence.

Confidence is based on some knowledge of the situation - not just on good will. The more we know about our stock and about the market, for example, the more confident we can be about its future behavior. In statistics, the issue is no different: the more we know about the random process we are interested in, the more we can trust (have confidence in) a probabilistic statement we make. Let's analyze this assertion by parts.

Intuitive Background & Concept

A quantitative random process may reflect, for example, the quality (or reliability) of a product batch. We can measure this quality by counting the number of defects per thousand parts produced (or by the number of failures per million

hours). But, as we know, product characteristics vary from batch to batch, because the production process is random.

Hence, we must observe the process for an extensive number of hours, to obtain a point estimate, e.g., the average number of defective items per batch or average number of failures per unit time considered. And we must also obtain a measure of the variation of these outcomes. These two values allow us to state what the true average number of defectives per batch will be**,** with a given probability (i.e., confidence interval).

As an example, a graphical representation of the number of defective items found in 5,000 samples of 1,000-unit batches is presented in Figure 1. Results reflect a defect rate of 8%.

In the graph, the height (value on the y axis) represents the frequency with which a given number of defective items (value on the x axis) occur in 5,000 samples of batches of 1,000 items. For example, there were 300 occurrences of 70 defects among the 5,000 samples of batches of 1,000 items each.
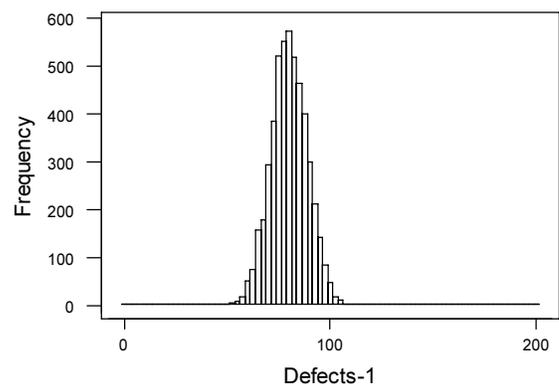

Figure 1. Example 1: Number of defective items in 5,000 1,000-unit batches (Defects-1)

We can see that observations of, say, 70 to 90 defects per 1,000 items are much more frequent than, say, less than 60 or more than 100 defects per 1,000 items. However, these observations can and do occur. Furthermore, if by chance they fall in our sample, they can bias our point and interval estimates, especially if our sample is small. That is why we need to draw large and random samples or to observe (or sample) the process for a long time - in the same way we

need to monitor the stock market for a long time - to acquire a high confidence in our estimate.

The second aspect of this problem is process variation, measured through standard deviations, ranges, or other measures of dispersion. Two random production processes may have the same batch averages and still have totally different variation patterns. We present in Figure 2 another process that also exhibits 8% defectives per 1,000 items. But now the process has a much larger variation than before.
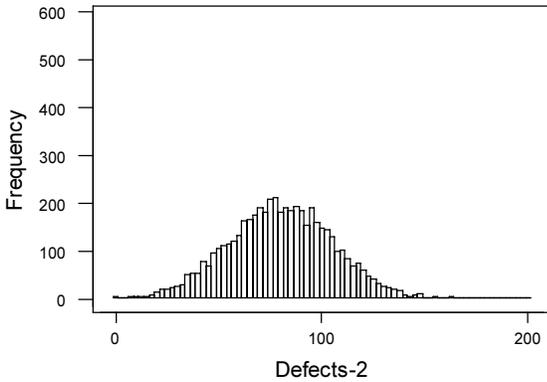

Figure 2. Example 2: Number of defective items in 5,000 1,000-unit batches (Defects-2)

We now compare the processes in Figures 1 and 2 and present the descriptive statistics for samples of 5,000. The results of the comparison are summarized in Table 1

Table 1. Comparison of Processes from Figures 1 and 2

Measure of Comparison	Variable	
	Defects-1	Defects-2
N	5,000	5,000
Mean	79.918	80.134
Median	80.00	80.55
Standard Deviation	8.727	25.055
Minimum	51.0	0.00
Maximum	114.0	199.7
Q1	74.0	63.07
Q3	86.0	96.6

Notice how the means and medians (measures of central tendency) are still practically the same. But the standard deviations and inter-quartile ranges (Q3-Q1) (measures of variation) for the second process (Defects-2) are larger than those of the first.

This characteristic is also reflected in the corresponding graphs. Observe how values below 50 or above 120 were practically non-existent in the first process (Defects-1). But such values are very plausible in the second. These values affect the Maximum and Minimum statistics.

The degree of uncertainty of our statement - or equivalently the confidence it yields - is a direct result of the sample size (or observation time) as well as of the natural variation of the phenomenon under study, expressed in terms of its Standard Deviation (σ).

We can say the same about the stock market. If a stock goes up and down continuously or we have followed it for a very short time, our confidence that this stock will go up cannot be as high as that which we place in a stock that has consistently increased its value during the past 50 weeks of observation.

Numerical Examples

To illustrate these issues, consider Table 2. It shows decreasing sample sizes from two populations that have the same 8% defect rate. But the standard deviation of the number of defectives, per batch of 1000, for the first population is 8.6. And that of the second population is 25 (three times larger). We want to show the consequences of decreasing the sample size (or observation time) and increasing variability in the phenomenon under study (represented here by a larger standard deviation).

Table 2. Measures for Populations from Figures 1 and 2 for Different Sample Sizes

Sample Size	Population 1 (See Figure 1)			Population 2 (See Figure 2)		
	Mean	Std. Dev.	Min.	Mean	Std. Dev.	Min.
Theoretical	80.00	8.60	0.00	80.00	25.00	0.00
5,000	79.92	8.73	51.00	80.13	25.06	0.00
1,000	79.97	8.95	55.00	79.57	25.53	0.00
500	80.23	8.84	56.00	79.15	24.28	0.00
100	80.08	8.93	61.00	79.01	24.64	23.98
50	80.10	7.66	62.00	83.18	27.09	34.81
30	81.03	10.47	61.00	77.76	23.71	45.72
20	79.65	6.26	72.00	77.04	23.21	30.67
10	83.30	9.23	68.00	80.32	18.16	47.66
5	84.20	12.15	69.00	97.90	45.30	35.70

First notice how, for very large samples (n $\geq$ 100), the theoretical and estimated Means and Standard Deviations are relatively close. For smaller sample sizes (n $\leq$ 20), the Standard Deviations, Maximums, Minimums, Quartiles and even the Means vary rapidly. This is one of the problems with small sample sizes. They drive down our "confidence" in the estimation of the population mean number of batch defectives, which constitutes our "statement of interest" (the statement could have also been the reliability of the product, its mean life, or any other population parameter).

Our statement of interest consists in providing a confidence interval (CI) that includes the population batch mean defectives, with a high probability (e.g., 95% of the time we obtain such an interval). The confidence statement is, precisely, that the CI will cover the true average batch mean defectives at least 95% of the

time. This raises two issues: the length of the CI and its probability. These two issues are related by the following equation:

$$\bar{x} \pm z_{\alpha/2}\, \frac{\sigma}{\sqrt{n}} = \bar{x} \pm H \quad \text{where} \quad H = z_{\alpha/2}\, \frac{\sigma}{\sqrt{n}}$$

The formula has four elements: length of the interval 2H, or half-length H, which is added and subtracted from the sample mean ($\bar{x}$), confidence level (1 - α), population standard deviation (σ) and sample size (n). Z is the Normal Standard percentile.

Since σ is a fixed (population) characteristic of the process, we can only control two of the remaining three factors: length (or half-length) of the interval, confidence level, and sample size. For example, let confidence level 1 - α = 0.95 and sample size n = 30. Then, from Table 2 and data, and using z(0.975) = 1.96 we obtain the two CIs for the respective populations:

First Population:

$\quad$ 81.03 $\pm$ H = 81.03 $\pm$ 1.96*(8.6/$\sqrt{30}$) = (81.03 $\pm$ 3.08)

Second Population:

$\quad$ 77.76 $\pm$ H = 77.76 $\pm$ 1.96*(25.0/$\sqrt{30}$) = (77.76 $\pm$ 8.95)

The results have the following meaning: the CI (77.95,84.11) for the first population, and CI (68.81,86.71) for the second, include 95% of the time the true mean number of defectives (80) per batch of 1000 items produced. We recall that the second group has a standard deviation three times larger than that of the first. It could very well be that (once in 20 times) the sample includes, by chance, say several very low numbers and produces a CI that does not contain the true batch mean defectives. Also, for the same confidence level (95%) and sample size (n = 30) the length (2H) of both intervals is very different (6.16 vs. 17.9) due to the different variations (σ = 8.6 or 25) of the two processes.

Discussion

We can always state, without hesitation (or the need to draw any sample), that the batch mean number of defectives always lies between zero and the batch size. And we can have absolute (i.e., 100%) confidence in this statement. However, this result is of little use, for our interval is so large that it has no practical value. Hence, a high confidence by itself is not enough.

The objective of deriving a CI for an (unknown) parameter is to estimate its value. Let's compare it to throwing a hat over a coin sitting on a table. The CI is the hat and the coin is the parameter. If the hat (CI) is too small, it will not cover the coin (parameter) very often. But if it does, the area where we know the coin lies is small. If instead we use a large Mexican sombrero (large CI) it will likely cover the coin (parameter) often. But the sombrero (CI) will be so large that the coin (parameter) may be as lost under the hat as before - and we will have gained nothing

from deriving such CI. This is the crucial trade-off problem in a confidence interval derivation: to find one, with a large enough probability of coverage (1 - α), and a small enough CI length (2H) as to be of practical use.

Therefore, we must strike a balance between the usefulness of the statement (e.g., the CI length 2H) and the assurance it instills (e.g., confidence level 1 - α). Such a balance depends highly on the variation (standard deviation σ) of the phenomenon under study and is achieved by using an adequate sample size (n). As illustration, reconsider the previous example, but now pre-establishing a FIXED CI half-length (H) of $\pm$ <u>two units</u>, for a confidence level of 95%. This means that the population and sample means will be, 95% of the times, at most two units apart.

For this case, the sample size (n) required to achieve such half-length (H = 2), with population σ = 8.6, confidence level 1 - α = 0.95 (that implies $z_{\alpha/2}$ = 1.96) is obtained by using the equation:

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{H^2} = \frac{1.96^2 * 8.6^2}{2^2} = 71.03$$

That is, we need to draw a sample of size n = 71 batches of 1,000 items, and average their respective number of defects per batch. Then, 95% of the time, this average will be, at most, two units away from the true population mean (80) of batch defects.

The restrictions that must be imposed in a statistical analysis to achieve a given confidence level are similar to those imposed to achieve high confidence for investors in the stock market. They include restrictions about stockbroker expertise and experience as well as about the length of stock market observation. In the statistical context, these factors are now replaced by restrictions on sample size, confidence level, interval length, etc.

Summary and Conclusions

We have seen how the statistical confidence of probabilistic statements is derived from the problem variability (standard deviation, σ) characteristics, as well as from the analyst's choices of sample size (n) and half-length of the interval (H). It is necessary to strike a balance between the confidence imposed (probability of it being true) and the statement (e.g., the length of the CI). If both are imposed, then the price that must be paid is a larger sample size or longer observation period, all other factors being equal.

It is also important to have reasonable expectations on statistical statements and their confidences. These expectations can be derived from similar or past experiences in the problems under study.

For Further Study

1. Coppola, A. Practical Statistical Tools for Reliability Engineers. RAC, 1999.
2. Sadlon, R. Mechanical Applications in Reliability Engineering. RAC, 1993.
3. Romeu, J.L. and C. Grethlein. Statistical Analysis of Materials Data. AMPTIAC, 1999.
4. Walpole, R., R. Myers, and S. Myers. Probability and Statistics for Engineers and Scientists. Prentice Hall. NJ, 1998.
5. Dixon, W.J. and F.J. Massey. Introduction to Statistical Analysis (3rd Ed). McGraw-Hill. NY, 1969.

About the Author

Dr. Jorge Luis Romeu has over thirty years of statistical and operations research experience in consulting, research, and teaching. He was a consultant for the petrochemical, construction, and agricultural industries. Dr. Romeu has also worked in statistical and simulation modeling and in data analysis of software and hardware reliability, software engineering, and ecological problems.

Dr. Romeu has taught undergraduate and graduate statistics, operations research, and computer science in several American and foreign universities. He teaches short, intensive professional training courses. He is currently an Adjunct Professor of Statistics and Operations Research for Syracuse University and a Practicing Faculty of that school's Institute for Manufacturing Enterprises.

For his work in education and research and for his publications and presentations, Dr. Romeu has been elected Chartered Statistician Fellow of the Royal Statistical Society, Full Member of the Operations Research Society of America, and Fellow of the Institute of Statisticians.

Romeu has received several international grants and awards, including a Fulbright Senior Lectureship and a Speaker Specialist Grant from the Department of State, in Mexico. He has extensive experience in international assignments in Spain and Latin America and is fluent in Spanish, English, and French.

Romeu is a senior technical advisor for reliability and advanced information technology research with IIT Research Institute (IITRI). Since joining IITRI in 1998, Romeu has provided consulting for several statistical and operations research projects. He has written a State of the Art Report on Statistical Analysis of Materials Data, designed and taught a three-day intensive statistics course for practicing engineers, and written a series of articles on statistics and data analysis for the AMPTIAC Newsletter and RAC Journal.

Other START Sheets Available

Many Selected Topics in Assurance Related Technologies (START) sheets have been published on subjects of interest in reliability, maintainability, quality, and supportability. START sheets are available on-line in their entirety at <http://rac.iitri.org/DATA/START>.

For further information on RAC START Sheets contact the:

Reliability Analysis Center
201 Mill Street
Rome, NY 13440-6916
Toll Free: (888) RAC-USER
Fax: (315) 337-9932

or visit our web site at:

<http://rac.iitri.org>

About the Reliability Analysis Center

The Reliability Analysis Center is a Department of Defense Information Analysis Center (IAC). RAC serves as a government and industry focal point for efforts to improve the reliability, maintainability, supportability and quality of manufactured components and systems. To this end, RAC collects, analyzes, archives in computerized databases, and publishes data concerning the quality and reliability of equipments and systems, as well as the microcircuit, discrete semiconductor, and electromechanical and mechanical components that comprise them. RAC also evaluates and publishes information on engineering techniques and methods. Information is distributed through data compilations, application guides, data products and programs on computer media, public and private training courses, and consulting services. Located in Rome, NY, the Reliability Analysis Center is sponsored by the Defense Technical Information Center (DTIC). Since its inception in 1968, the RAC has been operated by IIT Research Institute (IITRI). Technical management of the RAC is provided by the U.S. Air Force's Research Laboratory Information Directorate (formerly Rome Laboratory).