Presentation
Paper
Bio
Return to Main Menu

P R E S E N T A T I O N

T15

Thursday, February 15, 2001
1:00PM

USING STATISTICS TO EVALUATE PROCESS IMPROVEMENT

Paul Below
EDS

International Conference On
Software Management & Applications of Software Measurement
February 12-16, 2001
San Diego, CA, USA

Using Statistics to Evaluate Processes or How Do We Know We Improved?

Paul Below

Outline

- Introduction
- Experiments and Quasi-Experiments
- Analytical Methods

Introduction

- Analysis can determine if a perceived difference could be attributed to random variation
- Inferential techniques are commonly used in other fields, we have used them in software engineering for years
- This is an overview, not a training class

Metrics Analysis

The SEPG 2000 Conference had many “level 4” talks, SPC is a hot topic!

- SPC is not new
- SPC for software is not new
- “Too hard”!?
- It takes a long time for a best practice to become widely used

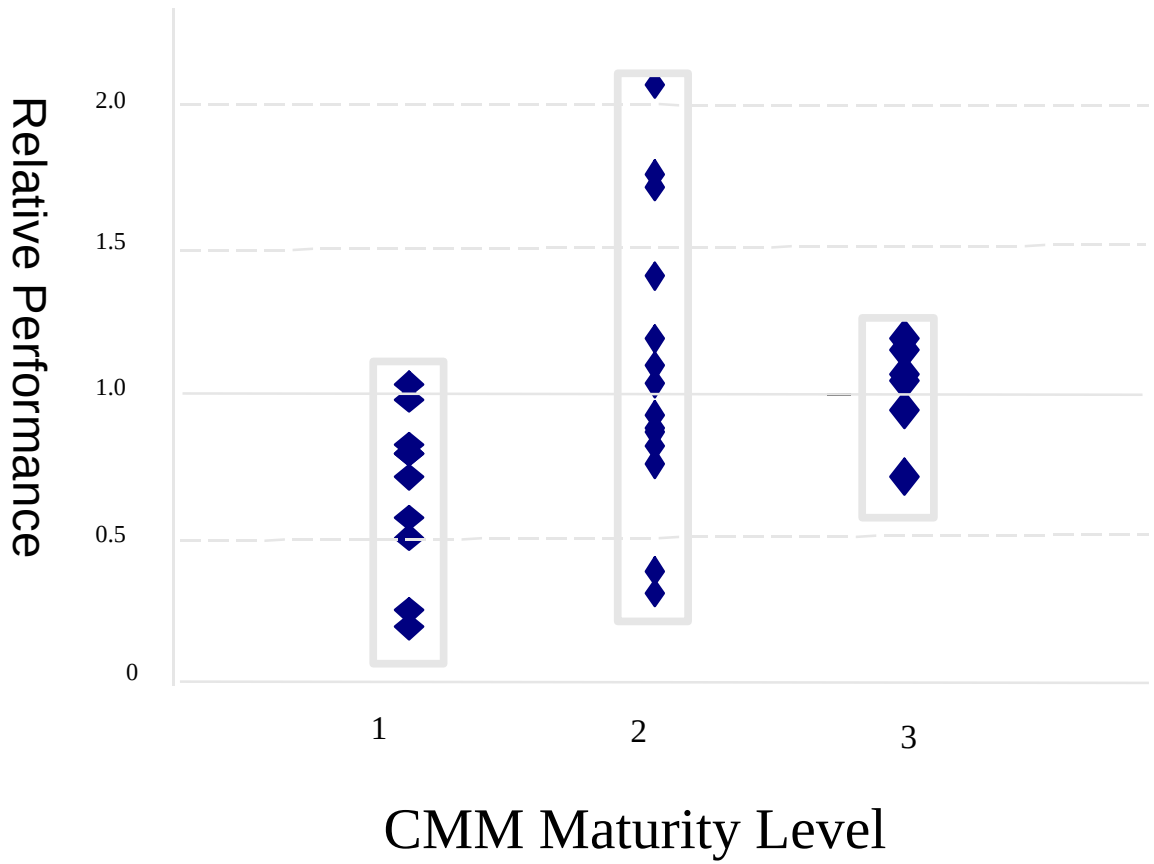
Expand our Set of Techniques

Metrics are used for:

- Benchmarking
- Process improvement
- Prediction and trend analysis
- Business decisions
- …all of which require confidence analysis!

“Is there any way that the data can show improvement when things aren’t improving?” -- Robert Grady

Is This a Meaningful Difference?



Pressure to Product Results

- Why doesn't the data show improvement?
- "Take another sample!"
- Good inference on bad data is no help

"If you torture the data long enough, it will confess." -- Ronald Coase

Experiments, Quasi-Experiments and Studies

“Experiments should be reproducible. They should all fail in the same way.”

Types of Studies

Anecdote è Case Study è Quasi-experimental è Experiment

- Anecdote: "I heard it worked once", cargo cult mentality
- Case Study: some internal validity
- Quasi-Experiment: can demonstrate external validity
- Experiment: can be repeated, need to be carefully designed and controlled

Attributes of Experiments

Subject è Treatment è Reaction

- Random Assignment
- Blocked and Unblocked
- Single Factor and Multi Factor
- Census or Sample
- Double Blind
- When you really have to prove causation (can be expensive)

Limitations of Retrospective Studies

- No pretest, we use previous data from similar past projects
- No random assignment possible
- No control group
- Cannot custom design metrics (have to use what you have)

Quasi-Experimental Designs

- There are many variations
- Common theme is to increase internal validity through reasonable comparisons between groups
- Useful when formal experiment is not possible
- Can address some limitations of retrospective studies

Causation in Absence of Experiment

- Strength and consistency of the association
- Temporal relationship
- Non-spuriousness
- Theoretical adequacy

What Should We Look For?

Are the Conclusions Warranted?

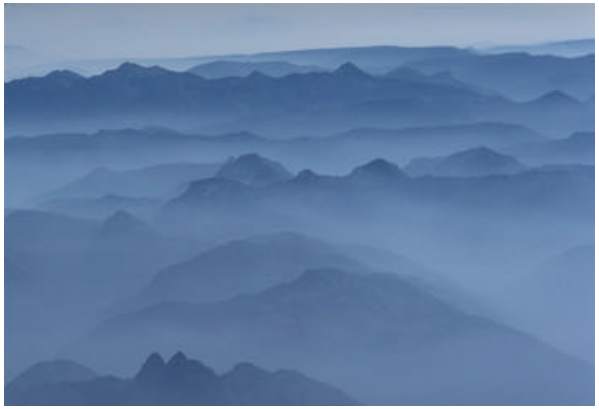
Some information to accompany claims:

- measure of variation
- sample size
- confidence intervals
- data collection methods used
- sources
- analysis methods

Analytical Methods

“There is nothing more deceptive than an obvious fact.” -- Sherlock Holmes

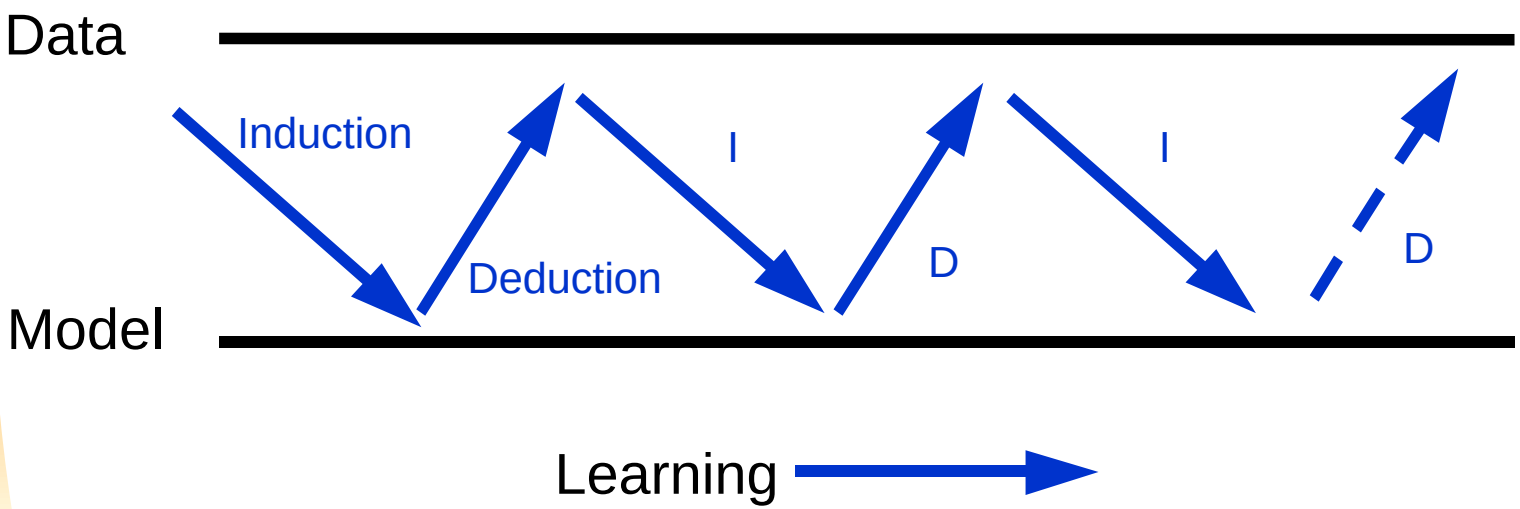
Decision Without Analysis



- Conclusions may be wrong or misleading
- Observed effects tend to be unexplainable
- Statistics allows us to make honest, verifiable conclusions from data

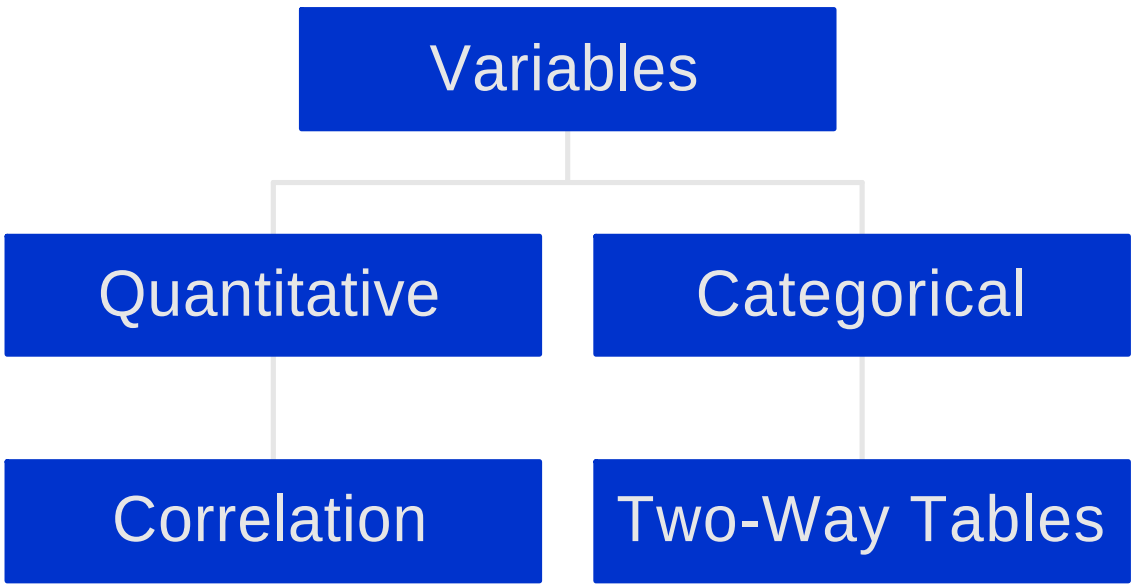
Statistical Thinking

is more important than methods or technology
Analysis is iterative, not one shot



(Modification of Shewhart/Deming cycle by George Box, 2000 Deming lecture, *Statistics for Discovery*)

Types of Confidence Analysis



Two Techniques We Use Frequently

- Inference for difference between two means
  - Works for quantitative variables
  - Compute confidence interval for the difference between the means
- Inference for two-way tables
  - Works for categorical variables
  - Compare actual and expected counts

Expressing the Results “in English”

- “We are 95% certain that the difference in average productivity for these two project types is between 11 and 21 FP/PM.”
- “Some project types have a greater likelihood of cancellation than other types, we would be unlikely to see these results by chance.”

Using Statistics to Evaluate Processes: How Do We Know We Improved?

By Paul Below

Introduction

It is often necessary or advantageous to examine differences between processes (or technologies), for the purpose of making business decisions. Statistical thinking is needed to evaluate the impact of process or other changes on organizational performance. In statistical thinking, past experience is summarized or generalized. Statistical thinking allows us to make predictions and reach conclusions.

The presentation provided a brief explanation of why inferential statistical techniques are useful. Inferential techniques should be used to extend Statistical Process Control (SPC). SPC provides a basis for actions and decisions related to process, typically distinguishing between special and common cause as well as determining root causes. SPC is a merging of techniques (control charts, Pareto charts, etc.) with a type of statistical thinking.

> "Statistical Process Control has always been, first and foremost, a way of thinking which happened to have some techniques attached."[1]

The techniques associated with SPC are very useful, but they are not sufficient alone to provide inferential comparisons. An additional need is the ability to make valid comparisons.

This paper suggests two additional techniques to help evaluate differences: inference for difference between two means (using *t* tests and confidence intervals for the difference between the means), and inference for two way tables (using chi-square tests). These basic statistical techniques should be in our analysis toolbox, along with the traditional SPC tools. This will allow us to make powerful conclusions, such as this example statement:

> "The 95% confidence interval for the difference in the means between projects of type A and B is: 12.3 +/- 5.6 AFP/PM. Therefore, we have reason to believe that there is a real difference in the productivity rates and that we can be confident that the true difference in the means is between 6.7 and 17.9 AFP/PM."

The rest of this paper presents examples of the two techniques.

Comparing Two Means

The following example uses industry data to illustrate comparison of two means. The example question to be answered is to determine if there is a relationship between productivity and project size. This example uses two quantitative variables: project size and

[1] Donald J. Wheeler and David S. Chambers, Understanding Statistical Process Control, second edition. (Knoxville, Tenn: SPC Press, 1992), p. 10.

project productivity. Note that there are other techniques that could be used to help answer this question. For simplicity, this example is limited to one of the two techniques suggested in the presentation.

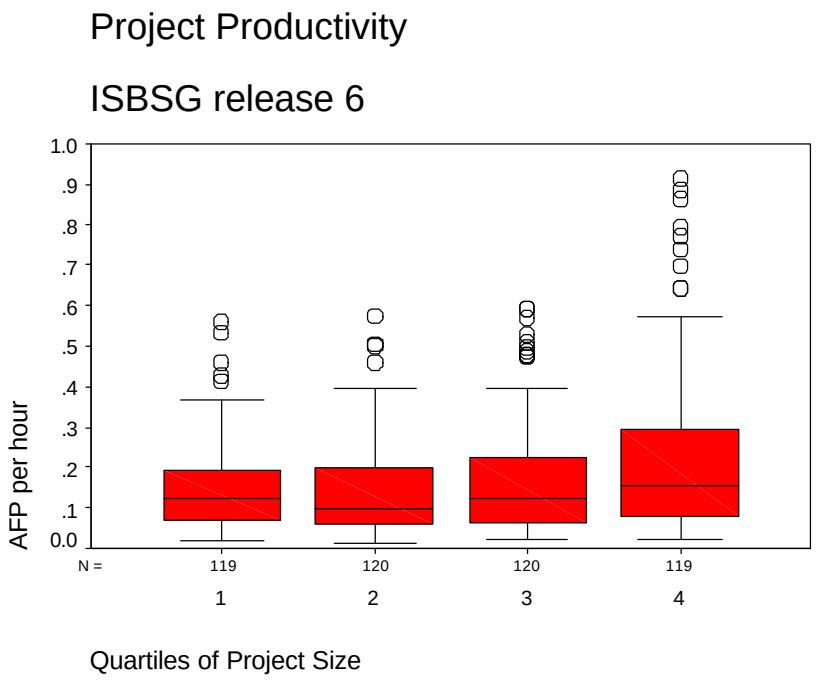
Projects from an industry database[2] were selected based on the following criteria:

- Data quality rating of A or B[3]
- IFPUG approach used for FP count
- Resource level 1 or 2 (project team and support staff) used for effort reporting

The selected projects were divided into quartiles based on project size (in adjusted function points). The following table displays the project sizes at the division points:

Description	Adjusted Function Points
Maximum	17518
75th Percentile	559
50th Percentile	253
25th Percentile	130
Minimum	9

The following box plot illustrates the distribution contained in the data. Outliers are indicated by circles; extreme values are not shown.



[2] International Software Benchmarking Standards Group, ISBSG Data Disk Release 6. http://www.isbsg.org.au/

[3] Descriptions of data elements available from http://www.isbsg.org.au/datadisk.htm

The ISBSG data exhibits no large or obvious difference in productivity rates for different sized projects. However, the interquartile range for the largest projects (larger than 559 AFP) is slightly higher than the boxes for smaller projects. Is this a real difference?

One method for investigating this question is to determine if the difference in mean productivity is significant.

The null hypothesis is that there is no significant difference between the quartiles, and that the projects were drawn from populations that have the same mean; any differences in mean are due to random variation. The alternative hypothesis is that there is a significant difference.

To test the null hypothesis, compute the *t* statistic. Using the *t* distribution, calculate how unusual the observed value is if the null hypothesis is true. The distribution can be obtained from a Student's t table, or from a statistical software program.

The confidence interval of the difference between two means is:

$$\overline{X}_1 - \overline{X}_2 \pm t^* \sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}$$

where $\overline{X}$ is the sample mean, s is the sample standard deviation and n is the sample size.

The test done for this example was at the 95% level, using the statistics package SPSS. The following ANOVA table shows that none of the differences in means between the size quartiles is statistically significant at the 95% level. Read the table as follows:

- Each row in the table corresponds to a comparison of two of the quartiles. The difference in average productivity between the two quartiles is shown in the column labeled "Mean Difference".
- The column labeled "Std. Error" is calculated from the within-group standard deviation and the sample sizes. Ideally, the standard error would be much smaller than the value in the Mean Difference column.
- For our purposes, the key column is "Sig", which shows the observed significance level for the test of the null hypothesis. For a difference to be significant at 95% probability, the value in this column would have to be 0.05 or less.
- The "95% Confidence Interval" for the mean difference gives a range of values that should include the true population difference between the two groups. Typically, intervals for comparisons that are significant will not include the value 0. In other words, if the lower bound is negative and the upper bound is positive, then the result is not highly significant.

Multiple Comparisons

Dependent Variable: AFP per hour
Bonferroni

(I) NTILES of FP	(J) NTILES of FP	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	5.056E-02	4.985E-02	1.000	-8.1515E-02	.1826
	3	4.824E-02	4.985E-02	1.000	-8.3834E-02	.1803
	4	-6.6047E-02	4.996E-02	1.000	-.1984	6.631E-02
2	1	-5.0563E-02	4.985E-02	1.000	-.1826	8.152E-02
	3	-2.3194E-03	4.975E-02	1.000	-.1341	.1295
	4	-.1166	4.985E-02	.118	-.2487	1.547E-02
3	1	-4.8244E-02	4.985E-02	1.000	-.1803	8.383E-02
	2	2.319E-03	4.975E-02	1.000	-.1295	.1341
	4	-.1143	4.985E-02	.134	-.2464	1.779E-02
4	1	6.605E-02	4.996E-02	1.000	-6.6307E-02	.1984
	2	.1166	4.985E-02	.118	-1.5468E-02	.2487
	3	.1143	4.985E-02	.134	-1.7787E-02	.2464

The most significant result is the comparison between quartiles 2 and 4. The mean productivity for these two quartiles differs by 0.1166 FP per Hour. However, the significance is .118 and the confidence interval does include zero. Therefore, there is a chance that the true mean difference is actually the opposite of what it appears! We have not found sufficient evidence to reject the null hypothesis. Even though larger projects have tended to be more productive, this analysis does not support a blanket claim that the project size is a determining factor.

As noted above, the significance number is very useful. In addition, confidence intervals can be important, because statistical significance might not be considered significant in a business or financial sense.

The *t* test for comparing two means could also be used to compare means of quantitative variables, between groups based on a categorical variable (for example, productivity of projects that used a particular process set could be compared to those that used a different process set). The next example covers the situation where we are dealing with only categorical variables.

Inference for Two-Way Tables

Categorical variables (such as project type, programming language used and delivery platform) cannot be analyzed with the test described in the previous section. Two-way tables are used to analyze categorical variables.

For simplicity, this paper uses the term "two-way table" to refer to any crosstabulation of summaries of counts of categorical variables, regardless of the number of dimensions in the

table (two-way tables, three-way tables, etc.). Another name sometimes used for this type of table is contingency table.

The procedure to be used is a chi-square test. First, calculate the statistic by comparing the observed pattern of the observation frequencies to the expected pattern based on a null hypothesis:
$X^2 = \sum [(\text{observed} - \text{expected})^2 / \text{expected}]$

A comparison is made between the computed chi-square statistic and the chi-square distribution to see how unlikely the observed value is if the null hypothesis is true. The distribution can be obtained from a chi-square table or from a statistics software program.

The following example uses data from a single organization. This organization tracked project management effort expended on projects in addition to other effort. The organization also tracked the estimated effort and duration for each project, and the actual effort and duration for completed projects. The organization had a specific goal for the amount of variation between estimated and actual effort and duration.

The question to be answered is whether the amount of effort expended in project management has impacted the project variances (as measured by whether the variation goal was met). The null-hypothesis is that the project management effort did not impact the probability that the project would meet the goal and that any observed differences are due to random variation. The alternative hypothesis is that the observed differences are not due to chance alone.

The projects were divided into three groups according to the percent of effort allocated to project management. Low is 10% or less, medium is between 10 and 20%, high is greater than 20% (these ranges are somewhat arbitrary and may not be appropriate for other organizations, they depend on the project size range as well as the effort collection standards). In addition the projects were categorized by whether they had met the goal for effort variance. The following two-way table displays the counts:

	Project Management		
Effort Variance	Low	Medium	High
Met	3	6	7
Not Met	9	10	9

The resulting chi-square has a p value of roughly 50%. We cannot disprove the null hypothesis; we have not shown that the above distribution varies from some cause other than random chance. In practical business terms, we have failed to uncover any impact of project management on effort variation. It is important to note, however, that we have not proven that there is no impact. There may be an impact, but we have not yet uncovered it. The next step might be to revisit our theories of how project management would be expected to affect project performance, and design further analysis.

The next table is similar to the previous one, except that the variance goal is accuracy of the estimated delivery date rather than estimated effort.

	Project Management		
Date Variance	Low	Medium	High
Met	2	10	13
Not Met	10	6	3

For the above table, the resulting p value is greater than 99.9%. This means that a distribution this extreme would be observed less than one time in a thousand, if the amount of project management did not impact the date variance. Therefore, we can be confident the distribution is not random, and we can reject the null hypothesis.

Note that although we have provided evidence against the null hypothesis, we have not proven causation. It is possible that some additional factor could be at work. If desired, additional categorical variables could be analyzed. The post-hoc analysis above could be strengthened with quasi-experimental techniques or even a formal experiment.

The conclusion that project management impacts delivery date variance may be what we expected. For example:

> "Tracking is a fundamental software management activity. If you don't track a project, you can't manage it. You have no way of knowing whether your plans are being carried out and no way of knowing what you should do next. You have no way of monitoring risks to your project. Effective tracking enables you to detect schedule problems early, while there is still time to do something about them."[4]

Because project managers can take corrective actions in response to monitored risks or observed problems, it seems reasonable that projects with a higher percentage of project management effort would be more likely to meet their estimated dates. The experience of this example organization supports this theory.

Summary

To support a maturing organization, a metrics analyst needs an expanded toolbox of statistical techniques. Methods such as t tests, confidence intervals, and chi-square tests are tools that help us make business decisions.

In addition to the techniques commonly associated with SPC, the analyst can make valuable contributions to the business with statistical methods for inference of quantitative and categorical variables

In closing, consider that our generalizations or predictions based on past experience are sometimes wrong. Statistics can help us learn when our generalizations are correct and when they are not.

[4] Steve McConnell, Rapid Development: Taming Wild Software Schedules. (Redmond: Microsoft Press, 1996), p. 57-58.

"It ain't so much the things we don't know that get us in trouble. It's the things we know that ain't so."[5]

Glossary of Key Terms

Alternative hypothesis. Describes the situation if the null hypothesis is false. Often, describes the situation that would exist if the theory we are testing is true.

ANOVA. Analysis of variance, a procedure for partitioning total variation. It is often used to compare more than two population means.

Boxplot. A graph that displays the median, the interquartile range, and the smallest and largest values for a group. A boxplot is more compact than a histogram but does not show as much detail.

Categorical variable. Data values that represent categories. May have some intrinsic order (ordinal data; for example low medium and high) or no intrinsic order (nominal data; for example project type).

Chi-square. The test statistic used when testing the null hypothesis of independence in a two-way table.

Confidence interval. A defined range of values within which a population parameter (e.g., mean, etc.) may be expected to fall. The width of the range depends on a stated confidence level and the distribution of the population.

Dependent variable. By convention, the vertical axis in a scatter diagram. Also known as the response variable, the value is considered to depend upon or result from the horizontal axis in a scatter diagram, which is known as the independent or explanatory variable.

Extreme value. In a boxplot, values greater than 3 box-lengths from the upper and lower edge of the box.

Interquartile range. In a boxplot, the lower boundary of the box represents the 25th percentile. The upper boundary represents the 75th percentile. The length of the box represents the interquartile range.

Mean. The average of a set of values.

Null hypothesis. The frame of reference against which sample results are to be tested, it describes a single situation. Most of the time, the null hypothesis claims the opposite of what you would like to be true.

Outlier. In a boxplot, values between 1.5 and 3 box-lengths from the upper and lower edge of the box.

P value. The conditional probability that the observed value of a sample statistic could occur by chance, given that a particular claim for the value of the associated population parameter is correct.

Quantitative variable. Numeric data values on an interval or ratio scale.

Standard Deviation. A commonly used measure of variability in a sample or population.

Significance. A measure of the outcome of a hypothesis test. Note that Statistical significance does not necessarily mean the result is significant in business terms.

T test. Used for inference on population mean when the standard deviation of the population is unknown. Will work for any size sample if the population distribution is normal, will also work for skewed distributions, especially if the sample size is greater than about 15.

[5] Artemus Ward, 19th Century American Humorist

Bibliography

Campbell, Donald T. and Julian C. Stanley. Experimental and Quasi-Experimental Designs for Research. Houghton Mifflin College, 1966.

International Software Benchmarking Standards Group, Worldwide Software Development: The Benchmark, Release 5. Clayton, Australia, 1998.

Ishikawa, Kaoru. Guide to Quality Control. Tokyo: Asian Productivity Organization, 1982.

Kazmier, Leonard J. Schaum's Outlines: Business Statistics Third Edition. New York: McGraw-Hill, 1996.

Norusis, Marija J. SPSS 9.0 Guide to Data Analysis. New Jersey: Prentice-Hall, 1999.

Schlotzhauer, Sandra D. and Ramon C. Littell. SAS System for Elementary Statistical Analysis, Second Edition. SAS Publishing, 1997.

Wheeler, Donald J. and David S. Chambers. Understanding Statistical Process Control, Second Edition. Knoxville, Tennessee: SPC Press, 1992.

Further Resources

Numerous measurement papers at the SEI (CMM) website: http://www.sei.cmu.edu/

International Function Point Users Group at http://www.ifpug.org/

International Software Benchmarking Standards Group at http://www.isbsg.org.au/

Refer to any good general statistics text that covers the following concepts related to inference: confidence interval; sampling; significance; null and alternative hypothesis; t procedures; chi-square test; and statistical thinking.

Contact a statistician for help.

The Author

Paul Below has been applying metrics in EDS since 1989. He is project manager of the EDS Information Solutions Delivery Central Metrics Group, coordinating metrics activity for a large organization. The Central Metrics Group integrates data collection, data validation, data analysis, estimating, and metrics consulting activities. Within EDS, he is an instructor and designer for metrics training. Paul taught the graduate-level (Masters in Software Engineering) class in Software Metrics at Seattle University for two years. He has presented at numerous technical conferences. He is a member of ASA and IEEE, and a former IFPUG CFPS.

Paul Below

Paul Below has been applying metrics in EDS since 1989. He is project manager of the EDS Information Solutions Delivery Central Metrics Group, coordinating metrics activity for a large organization. The Central Metrics Group integrates data collection, data validation, data analysis, estimating, and metrics consulting activities. Within EDS, he is an instructor and designer for metrics training.

Paul taught the graduate-level (Masters in Software Engineering) class in Software Metrics at Seattle University for two years. He has presented at numerous technical conferences. He is a member of ASA and IEEE, and a former IFPUG CFPS.