AC 2009-848: NONPARAMETRIC, COMPUTER-INTENSIVE STATISTICS: A PRIMER

Trent McDonald, West Inc.

Dr. Trent McDonald is a Consulting Statistician and Senior Manager at Western EcoSystems Technology, Inc. in Laramie, Wyoming.

David Mukai, University of Wyoming

Dr. David Mukai is an associate professor of civil engineering at the University of Wyoming in Laramie, Wyoming.

Non-Parametric, Computer-Intensive Statistics: A Primer

Abstract

The authors have developed a first course in statistics for engineers based on non-parametric, computer-intensive (NPCI) statistical methods. These methods do not rely on calculus or knowledge of statistical distribution theory, and as such can be taught earlier in a curriculum, are more intuitive, are less-recipe driven, and can be retained longer than traditional parametric statistics. In this paper, we provide a primer on NPCI methods. Basic NPCI concepts of bootstrapping and permutation are described. These concepts are then applied to confidence interval construction and hypothesis testing. Several examples taken from the course are worked to elucidate the methods.

Introduction

The authors have developed a new type of entry level statistics course focused on non-parametric computer-intensive (NPCI) statistics. NPCI methods do not rely on calculus because they do not depend on assumed distribution functions (thus non-parametric), instead their theory relies heavily on simple sampling concepts and their implementation utilizes computer re-sampling (thus computer-intensive). As a first course in statistics, NPCI methods are more useful for many students than traditional statistics because the basic theory posits that sampling from a sample of observed data mimics sampling from a conceptual (or real) population.

The potential benefits of a NPCI course are threefold. First, the course can be taught earlier in a curriculum than traditional statistics. Second, the methods are more intuitive and therefore stay with students longer. Finally, more sophisticated statistical procedures can be taught and used. These benefits mean that students are better equipped to solve statistical problems later in their careers. The benefits of NPCI are being investigated and results are presented elsewhere. This paper focuses on the concepts, methods, and applications of NPCI statistics.

NPCI Concepts

The theory behind many NPCI methods is not new. Many of the basic concepts have been in the statistics literature since the 1940's. However, NPCI methods did not see widespread application until the early 1980's because the necessary computing power was not available. With the advent of cheap and easy-to-use computers, computer intensive methods for realistic data sets became possible. The modern emergence of NPCI methods generally began in the late 1970s[1,2].

The basic concept of NPCI statistics is that variation of a statistic over repeated re-sampling of the sample will mimic variation of the statistic over repeated sampling of the population, if that were possible. In other words, resampling the original data is statistically equivalent to going back in time and conducting the same experiment again.

For the purposes of this paper, we introduce two main NPCI techniques; bootstrapping and permutation. Bootstrapping is primarily used to construct confidence intervals, while permutation is used to conduct hypothesis tests. Bootstrapping relies on sampling with replacement, while permutation relies on sampling without replacement.

To illustrate bootstrapping, consider the following example. Assume an experiment has n experimental units, and that each unit produces one number (statistic, e.g., hardness, cycles until breakage, etc.). Bootstrapping produces a pseudo sample by sampling n units and their associated number from the original sample. Bootstrap sampling is done "with replacement", meaning that the pseudo sample units are always drawn from the full sample and may contain duplicate units from the original sample. In this way, multiple different pseudo samples can be drawn. In most cases, these multiple pseudo samples mimic replication of the experiment, and thus a statistic computed on pseudo samples will have the same variation as the statistic computed on multiple replications of the experiment, if that were possible.

Permutation methods permute units between two or more samples. For example, consider sample A with n_A units and sample B with n_B units. Permutation combines the two samples and randomly assigns the units into new pseudo A and pseudo B samples. These pseudo A and B samples have exactly the same statistical properties, including identical means. These pseudo samples mimic a null hypothesis situation in which the populations have identical means. Variation of any statistic that measures the distance between means (such as a t statistic) over repeated permutation should mimic variation of the statistic applied to the original populations.

Applications

In this section, we expand discussion of the two applications mentioned above, confidence interval construction and hypothesis testing, and give some examples.

Confidence Intervals

Confidence intervals of any statistic can be computed by generating a large number of pseudo samples by bootstrapping. The statistic in question should be computed on all the pseudo samples, plus the original sample. A confidence interval is then the appropriate percentiles of the list of computed statistics. For example, to construct a 95% interval for the mean of a population, we generate 999 pseudo samples by bootstrapping. Including the original sample, we have 1000 means, the variance of which should reflect the variance of the original mean. After sorting, the 25th mean (2.5 percentile) and 975th mean (97.5 percentile) of this list are pulled to give the 95% confidence interval.

This confidence interval estimation method works regardless of the underlying distribution of the statistic. This means, for example, that the exact same procedure can be used to construct confidence intervals for complicated statistics that are not normally distributed. For example, it is easy to compute the 95% confidence interval of the standard deviation or the median. Instead of computing 1000 means, we compute 1000 standard deviations or medians. The 25th and 975th values in these sorted lists then form 95% confidence intervals for the true value of their respective statistic in the population. A more complex application is determining a 95% confidence interval for a regression coefficient. In this case, 1000 sets of data are generated (typically by bootstrapping residuals) and 1,000 regressions are carried out. The pseudo coefficients are then sorted and the middle 95% are taken as the 95% confidence interval. Numerical examples of confidence interval construction for means and for regression coefficients are given later.

Hypothesis Testing

Based on the principle of permutation, NPCI statistics can also be used to test hypotheses. If two samples were drawn from the same population (as opposed to being drawn from two distinct populations), the units in each sample could be interchanged with no change in the statistical distance between the two samples. For example, if we have two samples, Sample A and Sample B, then under the null hypothesis of no difference in the underlying means we could exchange units from A and B and the t-statistic of the two samples will change only by chance (or at random).

The strategy of a two-sample permutation test is to construct a large number of permuted samples. To do this, the units from A and B are randomly distributed to pseudo samples A and B with the constraint that the number of units in A and B are the same as the original samples. In this technique, unlike bootstrapping, there is no possibility for units to be used more than once. Once a pseudo pair of A and B samples are obtained, a regular t-statistic is computed. This procedure is then repeated a large number of times and the original t-statistic is compared to the distribution of t-statistics computed on the pseudo sample pairs. If Samples A and B were from the same population (no difference), then the original t-statistic should be indistinguishable from the pseudo t-statistics. However, if Samples A and B were from different populations (statistically different means), then the original t-statistic should be large in magnitude when compared to all the pseudo t-statistics. In fact, the percentile of the original t-statistic when combined with the pseudo t-statistics is the p-value. For example, if we generate 999 pseudo A and B samples and the original t-statistic is the 4th largest in magnitude from the list of 1000 t-statistics, then the p-value is 0.004. Numerical examples of one and two sided hypothesis tests on means of date are presented later.

Implementation and Examples

The basic concepts and applications of boot-strapping and permutation methods are relatively straight-forward and intuitive. The main difficulty with the method lies in implementation. For boot-strapping methods, practitioners must be able to accomplish the following tasks:

- Construct a large number of samples by sampling with replacement
- Apply a statistic to a large number of samples
- Sort a large number of statistics according to some criteria
- Select a given percentile of the statistic

For two-sample permutation testing, practitioners must be able to accomplish the following tasks:

- Construct a large number of sample pairs with permutation
- Compute the t-statistic for a large number of sample pairs
- Sort a large number of t-statistics

- Find a percentile of the original t-statistic

Because of the large number of pseudo-samples or pseudo-sample-pairs, the above tasks are usually done on a computer. For illustration purposes, it is possible to do a few sample iterations by hand, but hand computation is not practical for real world applications. Being able to do these tasks on a computer requires basic competence in computer programming or spreadsheet use. To aid application, the authors have implemented bootstrap and permutation routines in Excel, Mathcad, MATLAB, and R.

Example 1: Confidence Interval for the Mean and Standard Deviation

This example comes from a Junior-level Civil and Architectural Engineering experimental laboratory course. The students collected 10 hardness readings[3] and were asked to compute the mean, mode, median, standard deviation, 95% confidence interval for the mean, and a 95% confidence interval for the standard deviation of the data. A typical set of readings is shown in Table 1.

Table 1. Rockwell hardness readings (HRB).

91	93	93	93	93	93	94	94	93	94

The students constructed the 95% confidence interval of the mean and standard deviation using bootstrapping techniques in R. The instructor provided an example in R to calculate the 95% confidence interval for the mean of a data set. The students used this script to compute the 95% confidence interval of the mean then altered the script to compute the 95% confidence interval of the standard deviation. The original sample and five pseudo samples are shown in Table 2.

Table 2. Original hardness data with five pseudo samples.

Sample	Data										*Mean*	*Standard Deviation*
Original Sample	91	93	93	93	93	93	94	94	93	94	*93.1*	*0.876*
Pseudo Sample 1	93	93	94	93	94	94	93	94	93	91	*93.2*	*0.919*
Pseudo Sample 2	93	94	93	93	93	94	93	93	93	94	*93.3*	*0.483*
Pseudo Sample 3	94	94	93	93	94	93	93	93	93	93	*93.3*	*0.483*
Pseudo Sample 4	94	93	93	93	94	93	94	93	91	93	*93.1*	*0.876*
Pseudo Sample 5	93	94	93	91	93	94	94	94	93	94	*93.3*	*0.949*

This process was repeated for 999 pseudo samples for a total of 1000 samples. The students then took the 25th and 975th ordered mean and standard deviations to construct the 95% confidence intervals. The 95% confidence interval for the mean was 83.6 to 92.5 HRB, and the 95% confidence interval of the standard deviation was 0.316 to 1.252 HRB. In this case, computing a traditional 95% confidence interval for the mean is relatively simple for anyone with at least 1 course in statistics; however, computing a traditional parametric 95% confidence interval for the standard deviation is difficult even for graduate students in statistics. The students in this lab easily modified the script to do this because the jump from confidence intervals of means to confidence interval of standard deviations was based on basic transparent sampling principals.

Example 2: Confidence Interval of Non-Linear Regression Coefficients

A Civil Engineering M.S. student studied concrete creep over a period of one year. Creep is the phenomenon where concrete continues to deform over time under constant load[4]. Upon loading, concrete experiences an instantaneous strain called elastic strain. Over time, the strain increases, even with no additional load. This additional strain due to load is called creep strain. The ratio of creep strain to instantaneous strain is called the creep coefficient. Thus, a creep coefficient of 2 means that an additional deformation of double the initial deformation has taken place; that is, the total deformation has tripled (the original deformation plus double the initial deformation).

The M.S. student obtained 5 concrete cylinders, placed the same load on all cylinders, and measured deformation daily for a 1 week, weekly for 1 month, and monthly for 1 year. The measured creep coefficients for all 5 units are plotted versus age of concrete in Figure 1 along with the best fitting creep equation (below).

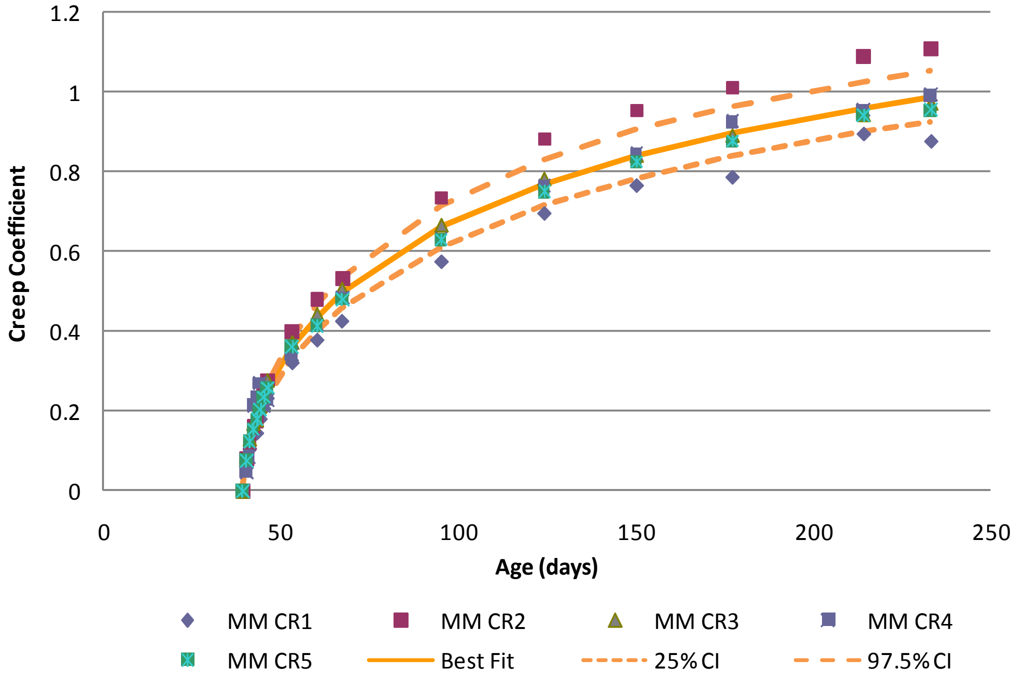

Figure 1. Creep coefficient vs. concrete age for 5 cylinders under constant load. Solid line is the best fitting standard equation for creep. Confidence intervals (dashed line) were determined by the bootstrap method described in the text.

The standard equation for modeling creep is[5]:

$$V_t = \frac{t^{\psi}}{D + t^{\psi}} V_u \qquad \text{Equation 1}$$

where:

$V_t =$ creep coefficient after t days under load

$V_u =$ ultimate creep coefficient (estimated)

$\psi =$ empirical constant (estimated)

$D =$ constant based on the duration of curing (estimated)

$t =$ time under load (in days)

The best fit line in Figure 1 was constructed by minimizing the sum of the squared residuals (method of least squares) of the data from all five samples by varying D, ψ, and V_u.

In this case, residuals around the fitted line are not independent of time under load or sample unit. Thus, true sample size is not total number of measurements, but total number of cylinders (i.e., 5), and individual residuals cannot be bootstrapped. Instead, all data from the five cylinders were bootstrapped. That is, a random sample with replacement was taken from the set (1,2,3,4,5) and data from the corresponding units was chosen for the pseudo sample. For example, if the random sample was (1,3,1,4,5), data from cylinder one was taken twice, and data from cylinders 3, 4, and 5 were taken once. The coefficients D, ψ, and V_u were then re-computed on the pseudo sample using the least-squares method in Excel (via Solver). The original sample and five pseudo samples are shown in Table 3

Table 3. Original creep sample and five pseudo samples.

Sample	Units					D	ψ	V_u
Original Sample 1	1	2	3	4	5	19.8	0.640	1.66
Pseudo Sample 1	1	3	1	4	5	18.9	0.644	1.53
Pseudo Sample 2	4	2	3	5	4	19.8	0.629	1.73
Pseudo Sample 3	5	3	4	2	2	19.9	0.650	1.70
Pseudo Sample 4	5	3	1	1	1	19.9	0.641	1.55
Pseudo Sample 5	4	3	3	5	5	18.5	0.630	1.62

Re-sampling cylinders and re-fitting was repeated 999 times to produce a set of 1000 values of D, ψ, and V_u. To compute a confidence interval for a particular age (say, $t = 250$), the set of 1000 coefficients were used to calculate 1000 creep coefficients for that age and these values were sorted. The creep values corresponding to the 25th and 975th values from this list of 1000 were then used as the lower and upper endpoints of the 95% confidence interval for that age. This process was repeated for all observed ages, and the lower and upper endpoints were connected in the graph.

Primary interest in this problem was eventual creep in the distant future. From the functional form of the standard creep equation, eventual creep is estimated by the value of V_u. The lower and upper endpoints of a 95% confidence interval for the real eventual creep were constructed as the 25th and 975th value of V_u from the list of 1000 values.

Computing equivalent confidence intervals for this problem using traditional parametric statistical methods is challenging, even for graduate students (and some professors) in statistics, due to the non-linear form of the equation and dependencies in the residuals. Although challenging via parametric methods, accurate confidence intervals for the coefficients of the creep equation were accessible to undergraduates and others who understand how the experiment was run, the basics of sampling, and how to program a computer to do the replications.

Example 3: Hypothesis Testing

The students in a Junior-level engineering laboratory course were instructed to determine if the number of cycles to break a paper clip when bent 90 degrees was different than the number of cycles until breakage when the clip was bent 135 degrees. The students had complete freedom in choosing how many units they wanted to test, developing a hypothesis, and testing the hypothesis. Some students tested their hypothesis with a parametric t-test and others with a 2-sample permutation test. The data from a student who opted for the permutation test is shown in Table 4. Data are given for angles other than 90 and 135 degrees because the students were also required to develop an expression relating cycles until failure to angle of bend.

Table 4. Paper clip experiment data.

Specimen ID	Cycles to Failure			
	45 Degree	90 Degree	135 Degree	180 Degree
1	86	10	8	4
2	55	8	6	4
3	61	19	14	4
4	82	16	9	3
5	68	16	8	4
6	71	25	6	3
7	56	17	10	3
8	89	15	10	3
9	91	17	12	6
10	107	16	6	6

In this example, the student broke 10 paper clips in Sample A (90 degree bend) and 10 paper clips in Sample B (135 degree bend). The original t-statistic for the data was 4.14. The student then made 999 pseudo pairs of samples by permuting the 20 combined values among the samples and computed 999 additional t-statistics. The original t was the largest of the 1000 t's, so the p-value was 0.001. Thus the student rejected the hypothesis that the two angles had the same number of cycles to failure. Many other students opted to do a parametric two-sample t-test because Excel has a built-in function for this test. The parametric t's p-value of 0.0009 was similar to the permutation t's p-value in this case. The parametric two-sample t-test, however, is not technically correct because the responses (cycles until failure) were counts and they violated the assumption of normality. Had this student only tested 1, 2 or 3 paper clips in each sample the parametric and permutation p-values would likely have lead to different conclusions. The permutation test was the correct test in this case.

Conclusion

In this paper, we presented two concepts for non-parametric computer-intensive statistics – bootstrapping and permutation. The concepts are used for two applications: confidence intervals and hypothesis testing. Three examples using NPCI statistics were presented: constructing a confidence intervals for the mean and for the standard deviation of a data set, constructing confidence intervals for a non-linear regression, and a hypothesis test.

These methods require knowledge of experimental methods, basic sampling theory, and rudimentary programming skills to conduct. They do not require calculus or knowledge of any statistical distribution theory for random variables. As such, these methods can be taught earlier in a curriculum. Furthermore, some of these analyses are difficult or impossible for undergraduates (and others) using traditional parametric methods. The computer intensive methods are intuitive and a course based on them is less recipe-oriented than traditional first courses in statistics. Because such a course is less recipe-oriented, the methods will likely be retained longer outside the first statistical course. If the methods are retained and internalized by students, they will be equipped to tackle a wide range of real-world problems in their careers.

Bibliography

1. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics , 7*, 1-26.

2. Efron, B. (1981). Nonparametric standard errors and confidence intervals. *Canadian journal of Statistics , 9*, 139-172.

3. Mamlouk, M. a. (2006). Materials for Civil and Construction Engineers. In M. a. Mamlouk, *Materials for Civil and Construction Engineers* (pp. 122-124). Upper Saddle River, NJ: Prentice Hall.

4. Mindess, S., Youg, J. F., & Darwin, D. (2003). Concrete. In S. Mindess, J. F. Youg, & D. Darwin, *Concrete* (pp. 440-453). Upper Saddle River, NJ: Prentice Hall

5. ACI Committee 209. (2001). *Prediction of Creep, Shrinkage and Termperature Effects, ACI 209R-92.* Farmington Hills, MI: American Concrete Instituted.