*Proceedings of the 1988 Winter Simulation Conference*
M. Abrams, P. Haigh, and J. Comfort (eds.)

# Bootstrap and correlated data

David Alan Grier
George Washington University
Department of Statistics/
Computer and Information Systems
Washington, DC 20052

## I. A Simplified Example

Eleven black bank tellers work at a business that employs 120 whites†. After one of the black employees is denied promotion four times, she sues the bank for racial discrimination. One of the statistics that is used in the trial is the number of black employees that are paid less than the average white salary (9 out of the 11). This statistic is noted by one of the judges on the trial as seeming to affirm the plaintiff's case, but little is done to analyze it.

A naive approach to analyze the statistic would be to note that the number of blacks earning less than the number of whites is simply a sign statistic§, with a binomial distribution. We will denote this statistic $S_{11}$. In this case, we are testing the null hypothesis that the distribution of the black salaries is equal to that of the white salaries. In this circumstance, we assume that a black is equally likely to be above or below the mean, and hence the probability density of $S_{11}$ is:

$$\Pr\{S_{11} = k\} = \binom{11}{k} 0.5^{11},$$

where $\binom{N}{K}$ is simply the binomial coefficient. In the case given, the one sided p value is

$$\Pr\{S_{11} \geq 9\} = 0.037,$$

which, at a 0.05 level, would be strong enough evidence to reject the null hypothesis and to accept the claim that blacks were paid less than whites.

There is a problem with this naive approach, because the mean of the white salaries is a random quantity. The business had a fairly constant turnover of staff and the average salary of the white staff depended on the experience and seniority of the white employees. This introduces an unwanted correlation structure into our problem. Before, going further, let us introduce some notation.

Let the random variables $W_i, i = 1, \ldots, N_W$ be the salaries of the white employees in our sample, let us further assume that these random variables are independent, having mean $\mu_W$ and variance $\sigma_W^2$.

Also, let the random variables $B_i$ $i = 1, \ldots, N_B$ be the salaries of the black employees in our sample. Again, we assume that they are independent, having mean $\mu_W$ and variance $\sigma_B^2$. Further more, we assume that $B_i$ is independent of $W_j$ $1 \leq i \leq N_B$ and $1 \leq j \leq N_W$.

Our sign statistic is then based on the quantities

$$R_i = B_i - \frac{\sum_{j=1}^{N_W} W_j}{N_W} \qquad i = 1, \ldots, N_B$$

and these quantities are no longer independent but have a positive correlation. The covariance between the $R_i$ is

$$\begin{aligned}
\mathrm{Cov}(R_i, R_j) &= \mathrm{Cov}(B_i - \frac{\sum_{j=1}^{N_W} W_j}{N_W}, B_j - \frac{\sum_{j=1}^{N_W} W_j}{N_W}) \\
&= \mathrm{Cov}(B_i, B_j) - \mathrm{Cov}(\frac{\sum_{j=1}^{N_W} W_j}{N_W}, B_j) \\
&\quad - \mathrm{Cov}(B_i, \frac{\sum_{j=1}^{N_W} W_j}{N_W}) \\
&\quad + \mathrm{Cov}(\frac{\sum_{j=1}^{N_W} W_j}{N_W}, \frac{\sum_{j=1}^{N_W} W_j}{N_W}) \\
&= 0 - 0 - 0 + \frac{\mathrm{Var}(W_1)}{N_W} \\
&= \frac{\sigma_W^2}{N_W} \qquad 1 \leq i < j \leq N_B.
\end{aligned}$$

An hence, the correlation is

$$\mathrm{Corr}(R_i, R_j) = \frac{\frac{\sigma_W^2}{N_W}}{\sigma_B^2 + \frac{\sigma_W^2}{N_W}} \qquad 1 \leq i < j \leq N_B$$

which is always positive, and is constant for all pairs of $R_i$.

The effect of positive correlation on the sign statistic is to spread the distribution. As the correlation approaches 1, the probability that either all the $R_i$ will be positive or all of them will be negative also approaches 1. The naive approach of just using the binomial distribution will be too liberal, rejecting the null hypothesis more frequently than should be the case. If we assume that the data are normal and have the same variance, we can compute the distribution of the sign statistic (Gastwirth and Grier, 1988). That distribution is:

$$\mathrm{F}_{S_{11}}(k) =$$

$$\binom{n}{k} \int_{-\infty}^{+\infty} \Phi(-\rho^{\frac{1}{2}} x(1-\rho^{\frac{1}{2}}))^k (1-\Phi(-\rho^{\frac{1}{2}} x(1-\rho^{\frac{1}{2}})))^{n-k} d\Phi(x)$$

† This example is a simplified version of an analysis used in the appeal of Watson v Fort Worth Bank & Trust (798 F.2d 791 (5th Cir. 1986)). While paralleling the data in that case, the data presented in this paper are fictious, and are intended to illustrate the problems of the analysis. The case is discussed in full in Gastwirth and Grier(1988).

§ The Wilcoxon would be a better statistic in this case, but the sign statistic was used in this case and is commonly used in other cases because it has a meaning that is readily understood.

where $\Phi(x)$ is the cdf of the standard normal distribution, and $\rho = \text{Corr}(R_i, R_j)$, $1 \le i < j \le N_B$.

While this is a valid solution, the formula does not easily generalize to more complicated models such as regression models. Furthermore, the assumption of a normal distribution is too restrictive for courtroom use, as it would be immediately challenged by an opposing attorney. Except in cases where the data clearly follow a normal curve, the evidence will be weakened.

The bootstrap is one method we can use to estimate the distribution of the sign statistic. We estimate the distribution of the sign statistic by doing a two stage random sample from the white data only. The algorithm is as follows:

**Algorithm 1.** *First Bootstrap estimate of the Distribution of the Sign Statistic.*

I. Randomly draw, with replacement, a sample of size $N_W$ from the white data and compute the average.
II. Randomly draw, with replacement, a second sample of size $N_B$ from the white data and compute $R_i \quad 1 \le i \le N_B$.
III. From newly computed $R_i$, compute the sign statistic and record its value.
IV. Repeat steps I to III until variance of results is appropriate low.
V. Compute the probability density by dividing the accumulated number of times each sign statistic appeared by the number of times that you iterated steps I to III.

This algorithm will estimate the distribution of the sign statistic under the assumption that the salaries of the Black population have exactly the same distribution as the salaries of the White population. No further assumptions are made about the nature of that distribution. This is an advantage for the legal work, in that the underlying distributions are often not normal, as was true in the case of Watson v Fort Worth Bank, and the sample sizes are too small for normal approximations.

The bootstrap estimator is a consistent estimator of the probability (Bose, 1988), and has all the other usual properties of the bootstrap (Ephron, 1983). However, we can improve the accuracy of this estimator by applying the conditional monte carlo variance reduction transformation. Instead of doing our second random sample in step II of the algorithm, we note that we are simply doing binomial sampling with the probability that each $R_i$ is positive equal to the number of white salaries greater than the estimated mean. This leads to an improved algorithm.

**Algorithm 2.** *Bootstrap estimate of the Distribution of the Sign Statistic using Conditional Variance Reduction.*

*I. Randomly draw, with replacement, a sample of size $N_W$ from the white data and compute the sample average.*
*II. Estimate the probability that $R_i$ will be positive by calculating the fraction, $P_k$ of white salaries $W_i$, that are greater than the sample average.*
*III. Repeat steps I and II until variance of results is appropriate low.*
*IV. For each $P_k$, compute the binomial distribution of the sign statistic*
*V. Do a weighted average of the distributions calculated in IV by multiplying each distribution by the fraction of times it occurs.*

This algorithm is just the conditional variance reduction transformation applied to the first algorithm. The variance reduction can be substantial. If $E_1$ is the estimate from the first algorithm and $E_2$ is the estimate from the second algorithm, then

$$\text{Var}(E_1) = \int \frac{(p_\mu - \int p_\mu f(p_\mu) dp_\mu)^2 f(p_\mu)}{N_W} dp_\mu + \int \frac{p_\mu(1-p_\mu) f(p_\mu)}{N_W N_B} dp_\mu$$

where $p_\mu$ is the probability that a single observation is less than the sample mean, $f(\cdot)$ is the density of the fraction of observations falling below the sample mean.

We also have the variance of $E_2$:

$$\text{Var}(E_2) = \int \frac{(p_\mu - \int p_\mu f(p_\mu) dp_\mu)^2 f(p_\mu)}{N_W} dp_\mu$$

and hence the difference is:

$$\text{Var}(E_1) - \text{Var}(E_2) = \int \frac{p_\mu(1-p_\mu) f(p_\mu)}{N_W N_B} dp_\mu$$

The difference term, $\text{Var}(E_1) - \text{Var}(E_2)$, can account for a large fraction of the variability in the $E_1$ estimator, especially if the original sample is tightly clustered. For a quick example, consider the case where $p_\mu$ can take one of three values $\{\frac{1}{3}, \frac{1}{2}, \frac{2}{3}\}$, with equal probability. If we use $N_B = 11$, then

$$\frac{\text{Var}(E_1) - \text{Var}(E_2)}{\text{Var}(E_2)} = 1.136.$$

By using the second algorithm, we halve the variability of the estimate.

## II. A More General Example

It is a rare company in which all employees have approximately the same salary. Salary is usually related to experience, education, performance and other factors. As was done in Watson v Fort Worth Bank & Trust, one common way to include this information is to fit a linear regression model to the white salaries, compare the final model to the black salaries and compute the sign statistic from the number of black salaries that are below the value predicted by the regression model. (Belson, 1956) This technique also induces a correlation structure on the sign statistic. This correlation structure can be much more complicated than that for the simple example. If we $R_i$ be the difference between the black salary $B_i$ and the value predicted by the linear regression model, then

$$\text{Corr}(R_i, R_j) = \frac{\mathbf{x}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j \sigma_W^2}{\sigma_B^2 + \sqrt{\mathbf{x}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \mathbf{x}_j (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_j \sigma_W^2}}$$

This value can be either positive or negative, depending on the value of the value of the covariates. Furthermore, the $R_i$ are not equicorrelated, since two difference pairs of $R_i$ may have different correlations. This complicates the problem. In

the simple example, we can calculate the distribution of the sign statistic, provided that we assume that all the data came from independent normal distributions with a common variance. Even with those assumptions, we cannot calculate the distribution if we have more than about 5 blacks in the model. Calculating such a distribution would require use to evaluate a 5 dimensional normal distribution CDF with arbitrary covariance matrix, something that cannot be done accurately.

The bootstrap presents a reasonable alternative to calculating the distribution of the sign statistic. For the general regression setting the algorithm is:

**Algorithm 3.** *Bootstrap estimate of the Distribution of the Sign Statistic using Conditional Variance Reduction for a Regression Model*

*I. Randomly draw, with replacement, a sample of size $N_W$ from the white data (keeping covariates and response variables matched) and fit the correct regression model to that data.*

*II. Estimate the probability that $R_i$ will be positive by calculating the fraction, $P_k$ of white salaries $W_i$, that are greater than the sample average.*

*III. Repeat steps I and II until variance of results is appropriate low.*

*IV. For each $P_k$, compute the binomial distribution of the sign statistic*

*V. Do a weighted average of the distributions calculated in IV by multiplying each distribution by the fraction of times it occurs.*

## III. Power Calculations and Alternative Null Hypotheses

Power calculations are important in legal work, especially for rebuttal evidence in a trial. To rebut the claim of discrimination, a defendant will often present statistical evidence that does not reject the hypothesis of not no discrimination. To properly weigh such evidence, it is important to know what magnitude of difference the methods can detect. Results have been admitted in court that could not detect a one million dollar a month salary difference (Gastwirth,1988). This method can be easily modified to calculated the distribution of the sign statistic, given that the two groups have a known difference in mean, $\Delta$.

**Algorithm 4.** *Bootstrap estimate of the Distribution of the Sign Statistic when the two groups differ by a constant $\Delta$ using Conditional Variance Reduction for a Regression Model*

*I. Randomly draw, with replacement, a sample of size NW from the white data (keeping covariates and response variables matched) and fit the correct regression model to that data.*

*II. Estimate the probability that $R_i$ will be positive by calculating the fraction, $P_k$ of white salaries $W_i$ plus the constant $\Delta$, that are greater than the sample average.*

*III. Repeat steps I and II until variance of results is appropriate low.*

*IV. For each $P_k$, compute the binomial distribution of the sign statistic*

*V. Do a weighted average of the distributions calculated in IV by multiplying each distribution by the fraction of times it occurs.*

The correlation between the residuals of both the simple model and the regression model are dependent on the ratio of the variances of the two populations. If the variance of the Black salaries is considerably larger than the variance of the White salaries, then the correlations are close to 0, and the regular sign test distribution may be used will little ill effects. However, if the Black salaries have a smaller variance than the white population, then the correlations between the residuals will be close to 1, and both the standard sign statistic distribution and the distribution estimated by Algorithm 2 will be far too liberal, suggesting a difference when there is none. The distribution of the sign statistic under these circumstances may still be estimated with the bootstrap technique.

**Algorithm 5.** *Bootstrap estimate of the Distribution of the Sign Statistic when the two groups differ possessing different variances, using Conditional Variance Reduction for a Regression Model*

*I. Fit the correct model to the White data using the entire data sample.*

*II. Create a set of pseudo observations, $W_i^*$, from the original data to reflect the difference in variances between the two groups. Let $\hat{W}_i$ be the fitted values from the regression. Make the pseudo observations $W_i^* = \hat{W}_i + \hat{W}_i \frac{\sigma_B^2}{\sigma_W^2}$.*

*III. Randomly draw, with replacement, a sample of size $N_W$ from the white data (keeping covariates and response variables matched) and fit the correct regression model to that data.*

*IV. Estimate the probability that $R_i$ will be positive by calculating the fraction, $P_k$ of white salaries $W_i^*$ that are greater than the sample average.*

*V. Repeat steps III and IV until variance of results is appropriate low.*

*VI. For each $P_k$, compute the binomial distribution of the sign statistic*

*VII. Do a weighted average of the distributions calculated in VI by multiplying each distribution by the fraction of times it occurs.*

The underlying assumptions for this last algorithm are under the null hypothesis, are that both the White and the Black salaries have a distribution $F(\frac{x-\mu}{\sigma_k}) \quad k \in \{B, W\}$, and that the sample variance of the Black and White salaries are good estimators of $\sigma_B$ and $\sigma_W$ respectively.

## IV. Summary

This is a problem that demands a nonparametric solution. The nature of legal statistics requires as few extra assumptions as possible. Furthermore, a good legal strategist should evaluate the quality of the result by calculating the power. The bootstrap is a method that fulfills those requirements. In the regression setting, where an arbitrary correlation structure may exist, it presents the only tractable solution.

The bootstrap is often applied naively, preparing the program to exactly model the sampling experiment. In this case, as simple conditional variance reduction transformation can be easily applied.

## AUTHOR BIOGRAPHY

DAVID ALAN GRIER is an Assistant Professor at George Washington University in the Department of Statistics, Computers and Information Sciences who has recently taken an interest to the problems of Legal and Public Policy Statistics. He received his PhD in Statistics at the University of Washington in Seattle and spent 5 years working as a computer designer for Unisys Corporation.

David Alan Grier
Department of Statistics, Computer and Information Sciences
The George Washington University
2201 G Street NW
Washington DC 20052

Bitnet: DAGRIER @ GWUVM
(202)994-6359

## Bibliography

Belson,W. A. (1956) A Technique for studying the effects of a television broadcast, *Applied Statistics*, V, 195-202.

Bose, Arrup, (1988) Assymptotics of the Bootstrap, *Proceedings of the 1988 Winter Simulation Conference,*, Institute of Electrical and Electronic Engineers, San Francisco, CA.

Bratley, Paul, Fox, Bennett, and Schrage, Linus (1983), *A Guide to Simulation*, Springer-Verlag, New York.

David, (1970) *Order Statistics*, Wiley, New York.

Ephron, Bradley (1983), *The Bootstrap, Jackknife and other Resampling Plans*, SIAM, Philadelphia, PA.

Gastwirth, Joseph and Grier, David (1988), A Note on Judge Goldberg's Dissent in Watson v Fort Worth Bank, submitted to *Jurymetrics.*

Gastwirth, Joseph (1988), *Report on an EEOC Case*, personal communication.

Gastwirth, Joseph (1988), *Statisical Reasoning in the Law and Public Policy*, Academic Press.

Nelson, Barry, (1987) Control Variates for Quantile Estimation, *Proceedings of the 1988 Winter Simulation Conference,*, Institute.

Rubenstein, Reuven (1983), *Simulation and the Monte Carlo Method*, Wiley, New York.