STATISTICAL ANALYSIS AND INTERPRETATION OF DATA COMMONLY USED IN EMPLOYMENT LAW LITIGATION

C. Paul Wazzan
Kenneth D. Sulzer

ABSTRACT

In employment law litigation, statistical analysis of data from surveys, observational studies, and company databases can help determine whether a generalization is possible as to employees, managerial or non-managerial status, or to inform damage calculations. We describe the general methodology of statistical analysis, including sampling, and provide a vocabulary for legal scholars and practitioners.

I. INTRODUCTION

Employment law litigation, such as "wage and hour" misclassification class action lawsuits and "overtime" or "off the clock" lawsuits, has increased in frequency over the past few years as a result of many factors, including increased awareness by the Plaintiffs' bar and new developments in the law. In many of these cases, the use of statistical analysis is increasingly common. The California Supreme Court, in *Sav-on Drug Stores, Inc. v. Superior Court of Los Angeles County* ("*Sav-on*") encouraged trial courts to adopt "procedurally innovative" methods, including survey results and the statistical analysis thereof, for evaluating class certification. In *Dukes v. Wal-Mart Store Inc. ("Dukes")* plaintiffs sought certification of a class of as many as one million women who asserted gender discrimination claims. The large number of potential plaintiffs in ("*Dukes*"), virtually requires statistical analysis.

This article provides legal scholars and practitioners with a vocabulary to address the statistical analysis and interpretation of

1. C. Paul Wazzan is a Principal with LECG, Corp., and specializes in providing financial and economic analysis. Dr. Wazzan can be contacted at pwazzan@lecg.com or 310-556-0622. Kenneth D. Sulzer, is a Partner with Seyfarth Shaw LLP, and specializes in employment law. Mr. Sulzer can be contacted at ksulzer@seyfarth.com or 310-201-5223.

2. 96 P.3d 194 (Cal. 2004).

3. 222 F.R.D. 137 (N.D. Cal. 2004).

data commonly used in employment and labor law litigation. We describe tests of significance, random sampling, and required sample size, and note that practitioners often incorrectly rely simply on the *mean* (*i.e.*, average) when drawing inferences – without considering the *standard deviation* of the distribution of the data being analyzed. This error often leads to false conclusions.

II. DATA GENERATED FROM SURVEYS, OBSERVATIONAL STUDIES ARE LIKELY TO BE MORE INFORMATIVE THAN DECLARATIONS

For many years most employment law litigation battles have consisted largely of "dueling" declarations. For example, plaintiffs would supply declarations stating that a proposed class of managers spent relatively little time performing management tasks while defendants submit equal numbers of declarations stating the converse. In many cases a more effective method to determine the amount of time spent by a proposed class of managers in various tasks is the use of an appropriately designed and administered survey, an observational study and/or an analysis of employee records.

Note that these other methods of determining whether employees are managerial can be (or at least seem) more reliable than declarations, in part because a proper survey, observational study or analysis of employee records has no direct involvement by either the defendant's or plaintiff's attorney. Perhaps more importantly, these methods can also be superior to declarations in that they can be specifically designed to allow for the systematic collection, quantification, and interpretation of the results – as opposed to the general interpretation required when using declarations (*e.g.*, half the declarations claim one thing and half claim another).

For the purpose of this paper we may refer to survey data by way of example, but in general, the analysis and interpretation of any data follows the same methodology.

III. THE BASICS OF STATISTICAL ANALYSIS: CONSTRUCTING CONFIDENCE INTERVAL

Consider a "wage and hour" class action. Whereas complete (but not necessarily the most reliable) results would be obtained by collecting data for every single current and former employee working at any time during the entire class period (potentially thousands or tens of thousands of employees, depending on the

---

4. It is not the purpose of this paper to discuss the correct construction and administration of surveys, but rather the interpretation of the results generated.

specifics of the case), it will often be necessary to draw conclusions from somewhat less data. For example, one may attempt to survey the entire relevant population but simply obtain a lower than 100% response rate for any number of reasons (*e.g.*, lack of participation, faulty contact information, time constraints).

In statistical sampling, a sample (the data that actually is observed) is drawn from the relevant population (all the data that can be observed) and analysis is performed to determine the characteristics of the sample. The characteristics of this sample are then applied to the overall population with a certain statistically specified level of confidence – often referred to as "significance at the X% level". It is generally accepted in academic literature and common practice that X should be 95 or 99.

The sample parameters (*e.g.*, mean, standard deviation, sample size) in addition to the specified level of statistical significance will allow for the construction of a "confidence interval." The confidence interval is simply a range of values around the sample mean. For example, if the sample mean weight of the oranges within a single bag of oranges is 1lb., one would not be able to say with certainty that all oranges in the orchard are therefore 1lb. Rather one would create a confidence interval indicating that it is reasonably certain that 95% of the oranges in the orchard will be between .75lb and 1.25lbs.

The width of a confidence interval is based on "Student's t-statistic." A t-statistic table would indicate the following, where (μ) is the sample mean, (σ) is the sample standard deviation, and ("n") is the sample size:

Table 1. Computing a Confidence Interval

Sample Size("n")	95% confidence interval	99% confidence interval
2	$\mu \pm 4.30\left(\sigma/\sqrt{n}\right)$	$\mu \pm 9.93\left(\sigma/\sqrt{n}\right)$
10	$\mu \pm 2.22\left(\sigma/\sqrt{n}\right)$	$\mu \pm 3.17\left(\sigma/\sqrt{n}\right)$
40	$\mu \pm 2.02\left(\sigma/\sqrt{n}\right)$	$\mu \pm 2.70\left(\sigma/\sqrt{n}\right)$
∞	$\mu \pm 1.96\left(\sigma/\sqrt{n}\right)$	$\mu \pm 2.57\left(\sigma/\sqrt{n}\right)$

5. Values are by way of example.

6. *See*, *e.g.*, MORRIS H. DEGROOT, PROBABILITY AND STATISTICS 393 (Addison-Wesley Publishing 2d ed. 1986).

IV. GENERATING AN APPROPRIATE SAMPLE

In constructing a sample upon which to perform statistical analysis, it is critical that the sample be "representative" of the population. For example, suppose we want to know the average strength of the UCLA student body. The relevant population is the entire student body. For this purpose, the football team would not be a representative sample; we would say that the football team is a "biased sample" in that it is clearly atypical. Likewise, if we just wander around campus and haphazardly try to pick students who appear to be typical, we are likely to end up with a sample that fails to reflect the diversity of the population in terms of gender, athletic experience, height, weight, and so on.

In a wage and hour context, a sample could be considered biased if it consisted only of managers with ten or more years of experience, or only managers in one particular jurisdiction. If certain members of the population have different characteristics than other members of the population (*e.g.*, managers with less than two-years of experience as compared to managers with more than ten-years of experience), then "strata sampling" may be appropriate.

In strata sampling, the entire population is divided into non-overlapping groups called "strata" and then a random sample is selected from each stratum, using a random number generator as described above. It is important that each individual in the population be associated with one and only one stratum.

There are several benefits to using stratified sampling: 1) the cost of collecting and analyzing the data could be reduced (*e.g.*, one could take smaller samples from strata with high sampling costs); 2) the variance of the estimator is often reduced since the variation within strata is likely to be less than the variance across the entire population; and most importantly, 3) separate estimates would be available for the parameters of each stratum. "Stratified random sampling" in a wage and hour misclassification context might lead to subclasses being certified.

Regardless of whether one wishes to develop a single sample, or a stratified sample, one of the best ways to produce a fair sample

---

7. Based on an example provided by: GARY SMITH, STATISTICAL REASONING 254 (Allyn and Bacon 1985).

8. This is not to say that this sample will always be biased. Each survey and test must be structured to inform the specified hypothesis or question. In addition to ensuring a non-biased sample, practitioners should ensure that results are not tainted (*e.g.*, attorneys administering their own surveys are highly suspect).

is through a technique known as "random sampling." In a random sample, each member of the population must be equally likely to be selected for the sample, and the chances of being selected must not depend on the selections. Random sampling will nearly always result in a representative and unbiased sample. Random sampling is easily accomplished. Suppose the population size is 1,000 and one wishes to generate an unbiased sample of 100. One would then assign a unique number from 1 to 1,000 to each of the 1,000 members of the population. Then, using a random number generator (available in statistical texts or common software) one draws 100 unique numbers. Those members of the population corresponding to these 100 draws become the sample.

V. DETERMINING THE APPROPRIATE SAMPLE SIZE

Now that we have seen how a confidence interval is constructed and how a sample can be constructed through random sampling, we turn to the related issue of how large a sample size should be. The sample size is determined by: 1) how narrow the confidence interval (or estimated range) needs to be; and 2) the Central Limit Theorem.

Table 1 shows that a larger sample size will reduce the operator following the plus/minus sign, thereby producing a smaller or (more precise) confidence interval. The larger sample size will also reduce the value of the element within the parentheses, thereby also producing a smaller (or more precise) confidence interval.

Table 1 also shows that the confidence interval is dependent on (σ), the sample standard deviation. The (σ) parameter will be large or small depending on the variation in our sample. If the variation of the sample is too large to produce a useful confidence interval then one may have to increase the sample size or divide the population into subsets.

Table 1 also shows that the specified level of statistical significance (*i.e.,* 95% or 99%) affects the operator following the plus/minus sign. The more precise the statistical level, the larger this

---

9. Note that when analyzing the statistical properties of a sample, it is generally assumed that drawing of the sample is with replacement; that is, the probability of being picked remains constant, and in theory the same item can be picked two or more times. In practice, however, this constraint may be relaxed without affecting the results in any significant way.

10. For a more detailed discussion of random sampling, *see*, *e.g.*, ALEXANDER M. MOOD ET AL., INTRODUCTION TO THE THEORY OF STATISTICS 222-24 (McGraw-Hill 3d ed. 1974).

operator will be and consequently the larger the confidence interval will be.

Lastly, and in addition to the foregoing, the Central Limit Theorem requires the sample size to be at least 30 or 40 if possible, to ensure that the sample distribution will be acceptably close to the normal distribution, thereby allowing use of the t-statistic which is predicated on the assumption that the sample is normally distributed.

In short, the researcher must consider all these parameters in determining the appropriate sample size for a desired level of significance, while simultaneously producing a useful confidence interval. This is generally an iterative procedure as the (σ) value is unknown prior to actually sampling. In practice one often starts with a relatively small sample size and works up towards a larger sample size as needed.

VI. The Binomial Distribution and the Assessment of "Managerial" Behavior

Each of the categories could arguably be answered with a simple "yes" or "no." This type of yes/no outcome is commonly known as a *Bernoulli* or *binomial* trial from which repeated testing (or answering) will yield a binomial distribution.

More formally, a statistical test based on a series of draws (in this case, questions answered) conforms to a binomial distribution when: 1) the experiment consists of n identical trials (surveys), where n is fixed in advance; 2) each trial (survey question) has two possible outcomes ("yes" or "no"); 3) the trials (surveys) are independent so the outcome of one trial has no effect on the outcome of another; and 4) the probability of success (*e.g.*, a "yes" answer) is constant from one trial to another.

To interpret these yes/no answers, one can use "1" to code each "yes" and "0" to code "no." The sample mean (μ) for a binomial distribution is the number of successes divided by the number of observations: μ/n. The standard deviation for a binomial distribution is by definition: $[\mu(1-\mu)/n]^{1/2}$. Using these values for the mean and standard deviation of the sample, one can then test the relevant hypothesis through the construction of the appropriate confidence interval as described above.

11. *See, e.g.*, GEORGE G. JUDGE ET AL., INTRODUCTION TO THE THEORY AND PRACTICE OF ECONOMETRICS 86 (John Wiley & Sons 2d ed. 1982).

13. Additional questions or a more extensive survey may often be required to further refine answers depending upon the application.

Example 1: estimating worked overtime hours

Suppose that one wants to know how much overtime a particular work force has worked. Suppose that the workforce consists of 10,000 employees for whom electronic data on hours worked is not readily available and that one can only recreate the overtime history at significant expense of time and money. One might then determine that a statistical sampling of this population of 10,000 is an efficient way to determine how much overtime has been worked without having to actually recreate the overtime history for all 10,000 employees.

One would first want to determine whether all 10,000 employees could be considered "the same" or "similar" (*e.g.*, one may not want to mix different job functions and/or titles). For simplicity assume that all 10,000 workers are generally similar (*i.e.,* perform the same functions, work the same number of shifts, have similar tenure).

The next step would be to draw a random sample of at least 40 (because of the Central Limit Theorem; see above) from this population. After recreating the overtime history for these 40 employees, one would know on a monthly/weekly/daily basis, as appropriate, how much overtime each of these employees had worked. Assume that the sample mean is 10 overtime hours per week with a standard deviation of 5 overtime hours. From Table 1 above, one could conclude that weekly overtime for 95% of the 10,000 employees would fall between the range of 8.4 hours and 11.6 hours – a spread of approximately 3.2 hours. If this range is suitably small for the required purpose, then one can stop at this point. If the resulting range is too large, then one can either increase the sample size, or stratify the sample to try and reduce the standard deviation

Example 2: determining whether a proposed class warrants certification

Assume we wish to know how many of 10,000 employees can be considered managerial. One could administer an appropriately designed survey and then use statistical analysis to interpret the data generated. Note that each of the four criteria, as described above in Section VI, can be answered with a simple "yes" or "no" coded as a "1" or "0" respectively.

For example, suppose our survey consists of a single question "are you a manager – yes or no?" and that we receive responses from 300 employees out of the total population of 10,000. Assume that 200 answer "yes" and 100 answer "no". We would conclude that the mean (*i.e.,* the percentage who answered "yes" is 200/300 or 66.67%. The standard deviation is 2.72%. With these parameter values, Table 2 indicates that the 95% confidence interval is 61.31% to 72.02%.

Applying these percentages to the total population of 10,000 indicates that at least 61.31% (or 6,131), but at most 72.02% (or 7,202) of the 10,000 employees are managerial (at the 95% statistical significance level). If this confidence interval is too wide, then one can try to increase the sample size or stratify the data (in essence trying to certify sub-classes).

Table 2. Computing a 95% Confidence Interval

Total Population Size	10000	
Total Surveys Returned (n)	300	
Respondents indicating "YES"	200	
Respondents indicating "NO"	100	
Sample Mean (% answering "YES") (μ)	66.67%	2/3
Sample Standard Deviation (σ)	2.72%	σ = [μ(1-μ)/n]^1/2

95% Confidence Interval	61.31%	< û <	72.02%	$\mu \pm 1.96\sigma$
Interpretation of Confidence Interval	6131	< û <	7202	

VII. Legal Analysis

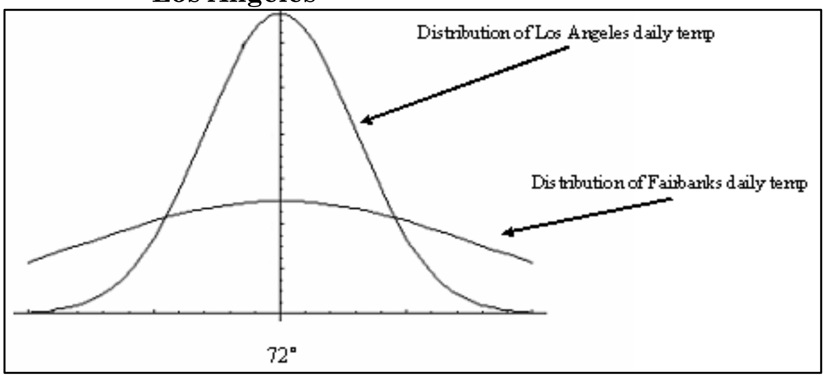
If the resulting range provided by a statistical analysis is too wide to inform a strong conclusion then one is said to reject the tested hypothesis. For example, if an estimated confidence interval is too wide, then one might conclude that a class should not be certified. There is some caselaw regarding class certification in disparate impact cases that also adopts this logic. If the variance is too great, then plaintiffs can try to increase the sample size or stratify the data (in essence trying to certify sub-classes). In practice, both defendants and plaintiffs will argue, respectively, that the observed confidence interval is too great (and a class should not be certified) or is small enough (so that a class should be certified).

A statistical analysis of survey, observational study or employee record data can lead to either a large or small confidence interval. The confidence interval largely depends on the variance of the underlying data. For example, a sample indicating that managers spend on average 90% of their time performing managerial tasks may show either substantial variance, or it can show that the distribution is clustered very narrowly around the mean 90%. In the former case, class certification may be inappropriate. In the later case, class certification may be appropriate.

Here is an extremely important point: While all statistical examples can generate a mean, unless the distribution of responses is clustered around this mean, then strong conclusions (*e.g.*, class certification) with respect to the group sampled is arguably inappropriate. Often, plaintiffs simply argue that because a data can generate a mean, a particular conclusion (*e.g.*, class certification) is appropriate. This argument is badly flawed, because one must also consider the underlying distribution of results, as described by a confidence interval. This extremely important statistical distinction is illustrated with one final example.

Assume that the average annual temperatures in Fairbanks, Alaska, and Los Angeles, California, are both 72º. Further assume that the variance of daily temperature in Alaska is large (it ranges from very cold to very warm) whereas in Los Angeles the variance is small (the temperature is always around 72º). Consequently, one cannot predict, with any precision, the temperature for a random individual day in Fairbanks, while one could do so for Los Angeles. Figure 1 graphically represents the distribution of each city's daily temperature.

Figure 1: Distribution of daily temperature for Fairbanks and Los Angeles



It is clear that the distribution of daily temperatures in Los Angeles are closely grouped around 72º. On the other hand, the distribution of daily temperatures in Fairbanks are more widely spread – the "tails" of the distribution are flatter.

The paradigm for this simple example is as follows: the "litigated" temperature in Los Angeles is a certifiable class whereas Fairbanks is not. Bell curves with a clear center suggest commonality and a flat bell cure would tend to signify a lack of commonality.

VIII. CONCLUSION

This article provides legal scholars and practitioners with a vocabulary to address the statistical analysis and interpretation of data that are increasingly common in employment and labor law litigation. This paper shows that simply relying upon the mean is likely to yield incorrect results or inferences. Careful application of statistical techniques (including sampling and the construction of confidence intervals) and analysis of the distribution of the underlying data are required to accurately draw inferences from the data.