CS-BIGS 5(1): 32-43
http://www.bentley.edu/centers/csbigs/goodman.pdf


# Measuring Changes in the Distribution of Incident-Outcome Severities: A Tool for Safety Management

**William M. Goodman**
*University of Ontario Institute of Technology, Canada*

*Responsible organizations monitor events that do or could injure their workers, or members of the public or the environment. As now is widely recognized, not all hazards pose the same relative risks that, if there is an incident, the outcomes will be serious or severe. The literature is found to not offer a clear method for assessing, statistically, upward (or downward) shifts in the distributions of incident-outcome severities that arise under different circumstances. An appropriate method could assist analysts who compare risks in diverse contexts—whether from floods, industrial hazards, hurricanes, or even financial risks. Typically, adverse-incident data are dichotomized (e.g. into fatal versus non-fatal outcomes), but advantages will be shown for a method that allows entire distributions of severity-outcomes to be compared. Applied to case-based examples from industry and a government safety organization, two methods are presented (at an Intermediate level) for comparing severity-outcome risks—one based on resampling procedures and the other using parametric approximations. Besides the data, the Excel file for this paper includes automated templates for readers to apply and experiment with the proposed methods.*

Good practice (and the law) generally require that employers monitor incidents that occur during their operations that do or could injure their own workers, as well as possibly members of the public or the environment. Incidents that occur in different contexts or environments can be distinguished by their relative frequencies of occurrence, or by the resulting severities of their occurrences. Presented in this paper are some revised tools and methodologies for addressing that important second aspect of safety - the likely severities of outcomes when incidents do occur.

## Distributions of Incident-Outcome Severities ("Severity Distributions")

When the results of incidents are analyzed over a period of time and, possibly, aggregated by region or industry, the distribution of their severities may look similar to the classical model of an "accident pyramid" (or "safety triangle"), introduced by H. W. Heinrich in 1931 (Heinrich, Peterson, Roos 1980). Figure 1 expands this model, and also reflects suggestions by Bird and Germain (1986). The key detail is that the most serious outcomes are just "the tip of the iceberg," and (assuming relative areas in the figure represent the corresponding risks for severities), we see that most accidents will usually have the less severe outcomes, and many will just be near misses.

Heinrich proposed that for every 300 incidents that occur without an injury, we should expect about 29 or 30 others to involve a minor injury, and one additional outcome to be a severe injury. Bird and Germain's revised numbers reflect their own, subsequent study of insurance claims in North America. They expect the ratios of outcomes, for severity levels 1, 2, 3, and 4 in the figure, to be about 600: 30: 10: 1. Those numbers are still being quoted as near paradigms for 'true' ratios (e.g., in Howe 2007, and

Wausau Insurance Companies 2007). Others retain the visual model, but acknowledge that exact numbers can vary over time, or by type of incident, or by company (e.g., Pinnacle West 2006).

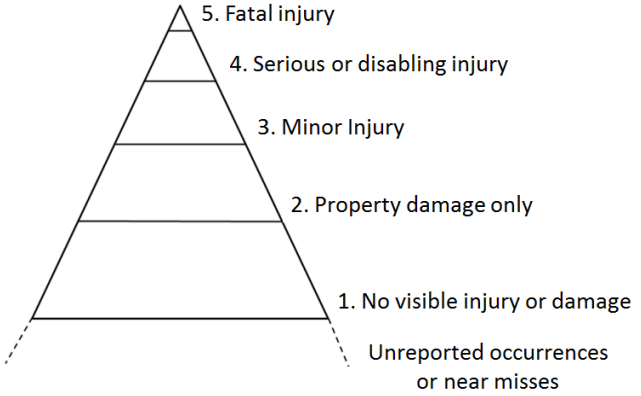



**Figure 1.** A "Safety Triangle" Distribution of Incident-Outcome Severities

An alternative way to represent the distribution of incident-outcome severities (i.e., a "severity distribution") is with a histogram, as illustrated in Figure 2. The objects for counting in this figure, slightly modified from an actual case for confidentiality, are the severity levels that resulted, as outcomes, from each individual incident (such as a fall, trip, shock, or chemical burn) that occurred during a specified time frame in the environment. This approach is better suited than the triangle for comparing severity risks *between* environments, or within one company from year to year. If the segments' areas in a safety triangle stand for relative risks of severity, note that each triangle-segment's area is a function of its squared distance from the base, so relative risks are hard to read from the figure—and in any case, the boundaries are rarely graphed to scale.

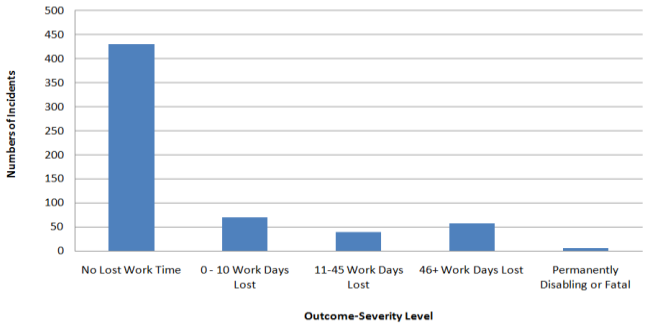



**Figure 2.** A Sample Incident-Severity Distribution

Perhaps a greater concern with the Safety Triangle model is its association, traditionally, with a specific assumption about how accidents and their outcomes' severities are related. This assumption is the "identical causation hypothesis" (Lozada-Larsen & Laughery 1987). It claims that the *same* factors that cause *any* accident to occur will also lead by chance to the few unlucky cases having severe outcomes; so in theory, the ratios among severity levels identified by Bird and Germain would hold, more or less, for *any* type of company or circumstance. Empirically, this claim is unsupportable, as even the living co-authors of Heinrich's fifth edition now acknowledge: "[Different] things cause severe injuries [from those that cause] minor injuries"; thus, "there are different ratios for different accident types, for different jobs, for different people, etc." (Heinrich, Petersen, & Roos 1980, pp. 64-65).

Lozada-Larsen and Laughery recommend a revised interpretation of the above hypothesis: The severity distribution *for a specific* firm or job class or time period *can* be mapped similarly to Figure 2, implying that relative risks for different severities of outcomes *in that context* are generally as shown. But, so far as another environment (or time) has different hazards, and different energy exposures, then a different distribution may apply there (or then). This revised model predicts, and the evidence confirms, distinct severity distributions for any of a variety of industries, whether in manufacturing (Kriebel 1982), meatpacking (Conroy 1989), or electrical generation and transmission (Hotte & Hotte 1990; Goodman 1992).

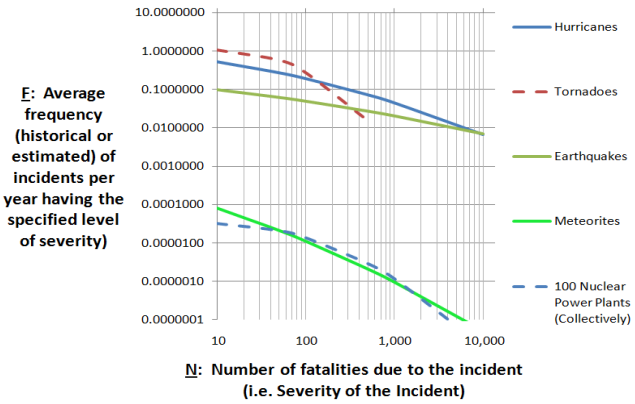



**Figure 3.** F-N Chart for the Distribution of Potential Severity Risks for Catastrophic Events

Note that in Figure 2, the counts are for incident-outcomes for one person at a time. If two people collide in one event, then each person's outcome-severity is counted separately. Compare this with a similar type of chart developed earlier, called an "F-N" chart, whose focus was on possible, public *catastrophes*; severity-levels are measured there in *numbers* of resulting fatalities. (See Figure 3.) For each severity level, its relative risk is in terms of *expected frequency* (e.g., on average, per year) for those severity levels occurring. Figure 3 is adapted from a

well-known chart, often cited and re-drawn, originally published in a study on nuclear reactor safety (U.S. Nuclear Regulatory Commission 1975; compare Foote 2002).

The graphical approach in Figure 4 combines Figure 3's comparisons of different sources' severity risks with Figure 2's focus on *individuals'* outcomes rather than the community's. Observe, in the figure, how it appears that the severity risks for those employed in tasks identified as "Work Category 6" may have shifted (like a rolling wave) towards the more severe end of the scale, compared to those generally in the population. The question then arises whether the degree of this apparent shift is significant in a particular case. This paper will explore procedures for answering that question.

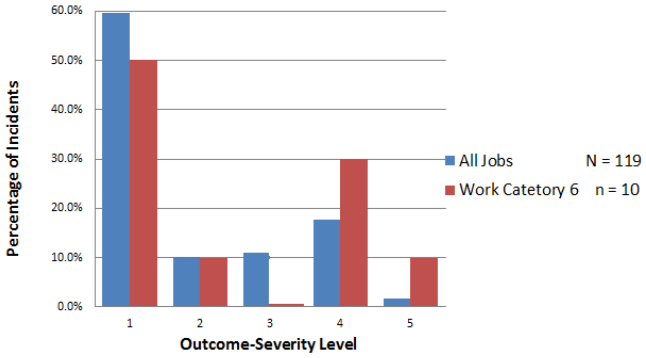



**Figure 4.** A Comparison of Severity Distributions - Overall versus One Work Category

## Assessing Upward (or Downward) Shifts Between Severity Distributions

For a person responsible for safety management, the rejecting of the "one ratio fits all" model for severity risk imposes a new obligation: As Petersen writes, "If we want to control serious injuries, we should try to predict where that will happen" (1998); and this entails monitoring specific types and contexts of work (or play, for recreational hazards), to identify how and where some activities are more dangerous than others (Manuele 2003). To provide something of a baseline, Manuele has compiled a list of the actual frequency ratios for (approximately) the severity levels envisioned by Heinrich, for various industries, and grouped by SIC (standardized industry) codes (Manuele 2004). A particular company might start by comparing its own distributions to the corresponding baselines for its industry.

The literature fails to offer, however, a clear or standard methodology for making the comparison which Manuele and others recommend—between the severity distribution for one's own work group or other relevant context versus some reference population's distribution (e.g. for a whole industry, or from previous years). One idea tried by Manuele (2004) is to sort industries by their respective ratios of fatalities to overall injuries, and to see where one's own organization fits; but this method does not use all the information in the full distribution. For any one company, the number of fatalities may (hopefully) be rather small, and as Manuele acknowledges (2008), the contributors to the most extreme outcomes may be "unique and singular events, having multiple and complex causal factors". So unless one is employing large samples of *aggregated* data, there is typically too much variance, and too little power, to distinguish one distribution from another based solely on the ratios of fatalities to all other injuries.

Another gap in the literature that could use a clear, documented method for identifying shifts between severity distributions is to complement papers that try to *predict* the severities of outcomes in differing contexts. Often, these latter employ ordinal severity classes. Yet for each severity class, they work backwards, in effect, to find clusters of covariate values that best "predict" each severity level (or, at least, increase its relative likelihood of occurrence). For example: What railway or highway features tend to lead to the more severe accidents at railway crossings (Hu *et al*, 2010); or what demographic features associate with the higher-severity skiing accidents (Corra & De Giorgi, 2007; Girardi *et al*, 2010), or to reduced life-functional status among elderly persons over time (Anderson *et al*, 1998).

Clearly, identifying clusters of characteristics that can elevate severity risks is beneficial. For example, predictive methods might have pointed to "Work Category 6" (see Figure 4) as associated with the higher severities. But Figure 4 shows also that the severity risks get "elevated" not as single points, but as upward *shifts in probability distributions*. Procedures are still needed to ask or confirm: How far precisely has an identified cluster-of-workers' severity distribution shifted from the aggregate distribution, with respect to both effect size and significance?

To answer such questions, this paper recommends comparing severity distributions, not by reducing them to dichotomous alternatives like [fatal, not fatal], but by comparing the distributions as entireties. The method is consistent with Lozada-Larsen's and Laughery's premise (1987) that the "identical causation hypothesis" contains a grain of truth—*if* we apply it, not to all accidents generally, but to those occurring in contexts that share comparable hazards, such as similar exposures to high

energies or sharp objects. With reference to Figure 4: If work-category "6" is truly "more hazardous" than others, this is not just because of its one fatality—which might have had very unique causes; but rather it is because, *all along the distribution*, the chances are always tending in the direction of more risk that outcomes will be more severe, compared to the baseline.

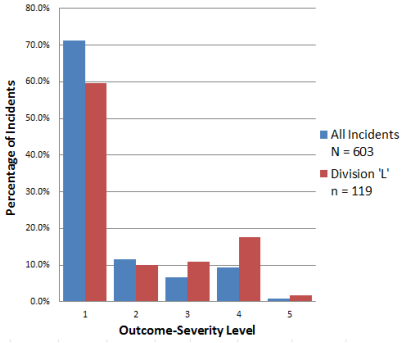
In a study of factors that impact severity for ski injuries, the methodology used by Takakuwa and Endo (1997) takes some steps in the direction proposed in this paper. Using for severity the scale measure AIS (Abbreviated Injury Scale, described in OrlandoHealth 2009), which has a limited number (six) of ordinal severity categories, they draw charts similar to Figure 4 to compare ski accidents' relative severity risks, depending, for example, on whether the skis' bindings were released or not during the event. Yet, their paper fails to explain clearly how to compare the illustrated severity patterns. Instead, the authors choose cut points (not always the same ones) to dichotomize the data (e.g. demarcating [AIS $\leq$ 2, versus AIS $>$ 2] for one test, and [AIS $<$ 4, versus AIS $\geq$ 4] for another test); then, based on these dichotomies, they apply chi-square tests to 2 $\times$ 2 tables.

In Cattermole's analysis of skiing injuries (1999), he similarly bases testing on chi-square analyses of dichotomized injury scores; in this case, using the more continuous Injury Severity Score measure, ISS. In an exploration of truck drivers' injuries, Charbotel *et al* (2003) take a similar approach to Takakuwa and Endo: They compare descriptively the full, differing distributions of severity scores, by AIS, for different groups; but then for formally assessing effect sizes or significance of differences between severities, they revert to dichotomized severities, based on [ISS<9, versus ISS≥9].

In short, when authors need to compare distributions of outcome severities, the methods they choose seem improvised and, collectively, inconsistent.

## Why Not Chi-Square ( $\chi^2$)?

As noted, some authors like Charbotel and Takakuwa & Endo recognize that different accident circumstances often have different distributions of outcome severities. To confirm these apparent differences they use $\chi^2$ tests - but apply these to simplified versions of the distributions, generally reducing them to just two values ("worse" or "better"), by some criteria. If graphically they have considered the whole distributions, why do they not *compare* them as a whole? Problems that can arise with the dichotomizing include (a) how to validate the cut-point decision (such as Takakuwa & Endo's AIS $>$ 2 in one test, and AIS $\geq$ 4 in another test in the same paper), and (b) lack of power and/or model appropriateness for using $\chi^2$ tests where $n$ is relatively small. Figure 5 illustrates both problems. (The data in the figure reflect, like Figures 2 and 4, actual business cases, with slight random modifications for confidentiality.)




| Severities | Division "L" | Population Distribution |
|---|---|---|
| 1 | 71 | 430 |
| 2 | 12 | 70 |
| 3 | 13 | 40 |
| 4 | 21 | 57 |
| 5 | 2 | 6 |

| | Division "L" | *Other* Division |
|---|---|---|
| Sev. < 4 | 96 | 444 |
| Sev. ≥ 4 | 23 | 40 |

| | Division "L" | *Other* Division |
|---|---|---|
| Sev. < 5 | 117 | 480 |
| Sev. ≥ 5 | 2 | 4 |

**Figure 5.** A Comparison of Severity Distributions - Overall versus Division "L"

By eye it appears that the risk of more severe accidents may be greater in Division "L" than in the overall population of a certain large company. If the selected cut point for testing a 2 $\times$ 2 matrix is set at severity levels less than versus equal to 5 (see bottom table in the figure), then the $\chi^2$ result will suggest no significant effect (i.e. no Division-based difference between the ratios of counts between severity levels) ($\chi^2$ = 0.708, DF = 1, p-value = 0.400); whereas if the cut point is set at severity levels less than versus at least equal to 4, then "a significant difference" would be concluded between the Division groups, with respect to ratios of counts between severity levels ($\chi^2$ = 12.495, DF = 1, p-value = 0.0004). Which conclusion is correct? To avoid "fishing" arbitrarily for the "best" cut point, what guidelines should be used? Also note that for the cut-point-at-5 decision in this example, the testing model's guideline that expected values in all cells equal at least 5 is violated. For smaller samples, it is not uncommon that even the more stringent guideline that all expected values equal at least 1 gets violated.

An alternative $\chi^2$ test, which does have the advantage of comparing severity distributions as a whole, is to use $\chi^2$ tests for goodness-of-fit—comparing a given severity distribution against a model for the expected distribution. These are a step in the right direction, but still have some problems.

Table 1 shows the outputs from Minitab Version 16, of conducting a goodness of fit test for Division L's severity distribution (per Figure 5; see "observed" column in the table) compared to the overall population distribution

("historical counts" column in the table). Again we see a violation of the guideline that all expected values be at least 5: In view of sample size, and the population's distribution, only "1.18" outcomes at Category 5 level would be expected. Nonetheless, the test model does clearly reflect a shift in severity pattern: For the lower severities, the observed counts are less than was expected if Division "L" were no different from the population, whereas for higher severities, the actuals are larger than expected. The effect size for $\chi^2$ reflects this relationship, as does the low p-value.

**Table 1.** Chi-Square Goodness-of-Fit Results

**Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Division "L"**

| Category | Observed | Historical Counts | Test Proportion | Expected | Contribution to Chi-Sq |
|---|---|---|---|---|---|
| 1 | 71 | 430 | 0.713101 | 84.8590 | 2.26344 |
| 2 | 12 | 70 | 0.116086 | 13.8143 | 0.23827 |
| 3 | 13 | 40 | 0.066335 | 7.8939 | 3.30290 |
| 4 | 21 | 57 | 0.094527 | 11.2488 | 8.45309 |
| 5 | 2 | 6 | 0.009950 | 1.1841 | 0.56223 |

| N | DF | Chi-Sq | P-Value |
|---|---|---|---|
| 119 | 4 | 14.8199 | 0.005 |

Table 2 illustrates a potential problem of using $\chi^2$ to test for differences between two severity distributions. Applied to safety, a mere change of severity pattern from some reference—in no particular order—is not usually what draws attention; alarms sound when the distribution tends to shift upward, systematically, to higher severities. The data for Table 2 are approximated from Takakuwa and Endo's Figure 5 (1997) (wherein the exact frequencies are not displayed). Severities of ski injuries for cases when ski bindings are unreleased during the incident ("Observed") are compared to the overall ski-injury distribution ("Historical Counts"). At first glance, the $\chi^2$ and *p*-values appear to support the authors' claim that "binding release [or more accurately, the failure of bindings to release] was significantly correlated with a higher [severity as measured by] AIS." But "correlation" generally implies a directional consistency, which is not found in the table: One of the two *lower*-severity categories for the Observed column has a larger than expected count, while one of the two *higher*-severity categories has a lower than expected count; these data do *not* suggest that non-release consistently tends to raise the severity. Because $\chi^2$ does not distinguish between on-trend and counter-to-trend variances from the expected values, both types of variance get added, spuriously, which can inflate the apparent effect size.

A third limitation of the $\chi^2$ goodness-of-fit approach, beyond the spurious effects and the often-violated minima requirement for expected values, lies in the difficulty of interpreting and communicating the effect size. As the scientific and publishing communities increasingly recognize, statistical results are more convincing if effect sizes (ideally with confidence intervals) are reported along with conventional *p*-values for significance (Ellis, 2010; Goodman, 2010). The $\chi^2$ measure can be inflated artificially by various circumstances; not just by spurious effects, but also for example if the data representing the population distribution includes a zero-frequency severity category, and the frequency increases by even one in the obtained sample. This inflationary effect could occur arbitrarily, due to boundary choices for the severity classes. So, even if reporting to an audience who are familiar with $\chi^2$, the extent to which its magnitude measures a real effect may not be intuitive or obvious.

**Table 2.** Spurious Effects in Chi-Square Goodness-of-Fit Results

**Chi-Square Goodness-of-Fit Test for Observed Counts in Variable: Unreleased**

| Category | Observed | Historical Counts | Test Proportion | Expected | Contribution to Chi-Sq |
|---|---|---|---|---|---|
| 1 | 105 | 285 | 0.437788 | 131.774 | 5.44004 |
| 2 | 160 | 285 | 0.437788 | 131.774 | 6.04592 |
| 3 | 28 | 53 | 0.081413 | 24.505 | 0.49836 |
| 4 | 8 | 28 | 0.043011 | 12.946 | 1.88976 |

| N | DF | Chi-Sq | P-Value |
|---|---|---|---|
| 301 | 3 | 13.8741 | 0.003 |

## Proposed Procedures to Assess Upward (or Downward) Shifts Between Severity Distributions

We conclude that new procedures are needed to validly and reliably assess whether, in comparing a sample's severity distribution to a reference severity distribution (for a putative population), the sample's distribution has meaningfully *shifted* towards each individual having greater probabilities of assuming higher (or lower) severity values than is modeled in the reference. The proposed method would not assess goodness (or non-goodness) of fit, in general, but rather, in the sense illustrated in Figure 4, whether there has been a systematic, upward (or downward) *shift-from-fit*. The intended procedures should also reflect the importance to safety practitioners of (a) being as straightforward as possible to use in the field, and (b) appearing intuitive to all stakeholders evaluating the results. Two such methods are presented.

The main method presented here ("Method 1") employs resampling techniques, to minimize requirements for parametric assumptions about the distributions. The Excel file conjoined with this article includes templates, with macros, that implement the procedures described; and the macro scripts and formulas are all fully accessible for inspection. (For quicker run-times, a commercial resampling add-in or program could be partly substituted.) The second approach ("Method 2") provides

parametric approximations for the resampling-based results from Method 1. Some may find this method more accessible, and it is found to be reasonably robust to variations from its ideal assumptions.

*Method 1*

The basic inputs and outputs for the proposed procedure are illustrated in Figure 6. This figure combines elements from three separate worksheets in the conjoined file: An input screen; a screen to generate a point estimate for effect size and a *p*-value; and a third screen estimating the 95% confidence interval for the effect size. The source data in the figure are the same as in Table 2. Observe that the spurious effect described previously, wherein $\chi^2$ was inflated by *non-systematic* changes across the distribution from the expected distribution, does not occur here: The *p*-value calculates as non-significant, and moreover the effect size appears to be negligible, as will be explained.

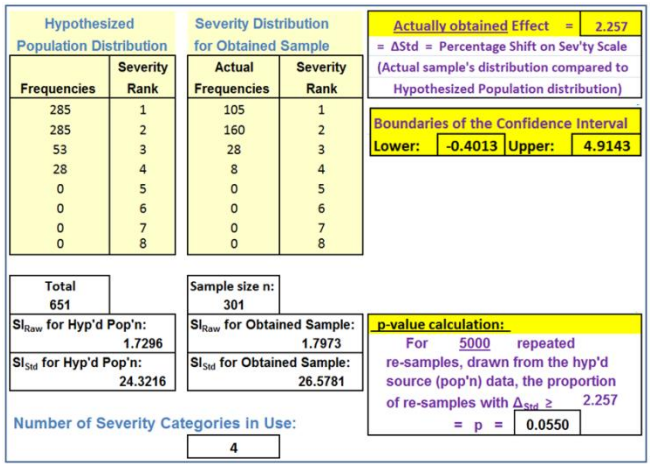



**Figure 6.** Inputs and Outputs of the Procedure

In Takakuwa and Endo's original example, the *hypothesized population distribution* is the marginal distribution of ski-injury severities, regardless of whether or not the victim's bindings were released during the incident. In other words, for a test of whether non-release of bindings impacts severity, the pooled severity distribution, for both groups combined, provides an estimate for the null population's distribution. The data labeled *severity distribution for the obtained sample* is, in this case, the distribution of injury severities specifically for those whose bindings were not released. The test question is: Has the obtained sample's pattern significantly shifted from the hypothesized population's?

The statistic underlying the effect size calculation is the *raw severity index* ($SI_{raw}$), calculated for each of the hypothesized population and the obtained sample. $SI_{raw}$ is simply a frequency-weighted mean for the severity levels:

**Formula 1.** $$SI_{raw} = \frac{\sum_{i=1}^{k}(SevRank_i \times Freq_i)}{\sum Freq}$$

where each $SevRank_i$ is one of the severity levels, from 1 to $k$, that have been defined for the application, and $Freq_i$ is the corresponding frequency for that severity level. In Figure 6, we see that four severity levels have been defined for the hypothesized population. If every level of severity were equally likely for every incident, then $SI_{raw}$ would be simply the mean of the severity level numbers (1, 2, 3, 4), and so equal to 2.5. The frequency column conveys the additional information that *not* all severities are equally likely to occur; so the frequency-weighted mean incorporates this extra distributional information.

Technically, Formula 1 (which is finding a mean severity level) presumes that the data are interval or ratio level, so that distances between any two, successive severity values (e.g. from severity level 1 to 2, versus from level 3 to 4) are all the same. Yet, it would seem that the assigned severity categories are ordinal data (simple ranks), which lack the required property. However, the author observes that, as assigned in practice, severity categories *can* be interpreted as quasi-interval-level in nature. They are certainly not assigned subjectively, as for example Likert-scale values might be. There are certain objective milestones or thresholds that must be reached or crossed to elevate a severity assignment to higher levels; in a sense, the severity level is like a count of milestones that have been met. For example: In industrial accident contexts, the 'distance' from Severity level 1 to level 2 might be the crossing of the threshold from no lost time incidents to *some* lost time. Further up the scale, the 'distance' from Severity level 3 to level 4 might be crossing of the threshold from 'merely' a serious, lost-time injury to an outcome for which (by policy or regulation) 'long term disability' status must be applied for. Thus, an increase in mean severity level for a severity distribution is a measure of how far up this scale of 'serious milestones reached' an organization's incidents are tending to fall.

Referring back to Figure 6, note that $SI_{raw}$ for the null distribution is about 1.73 and $SI_{raw}$ for the obtained sample is about 1.80. By eye this difference looks rather small; but the units are not well-suited for communication: On the one hand, the weighted mean for severity levels ($SI_{raw}$) is not widely familiar as a unit (although this could be addressed by education). More seriously, $SI_{raw}$ is a relative measure, dependent on the number of severity categories that happen to be defined

for an application. In a severity scale from 1 to 3, for example, an $SI_{raw}=2.4$ would suggest a marked trend towards more severe outcomes, yet in a severity scale from 1 to 7, an $SI_{raw}=2.4$ would seem more trended towards low severities. Therefore, $SI_{raw}$ values will be transformed to *standardized severity index* ($SI_{std}$) values, which are comparable between different contexts.

$SI_{std}$ can be interpreted as a measure of "percent-distance-along-the-severity-scale" (i.e. from the lowest to the highest possible values).

**Formula 2.** $SI_{std} = \frac{SI_{raw}-1}{K-1}\times 100$

where K is the defined number of severity categories. In the case of all ski injuries (see Figure 6 at the upper left), if all outcomes had severity level equal 1, then $SI_{raw}$ would equal 1, its lowest possible value; if all outcomes had the maximum severity level (here equal to 4), then $SI_{raw}$ would equal 4. Because in fact, $SI_{raw}$ equaled 1.73 - which is 24.3% along the distance scale, from the lowest- to highest-severity possibilities - $SI_{std}$ equals 24.3.

On this basis, *the point estimate for effect size* (ΔStd) is defined as the difference between the two standardized severity indexes - the obtained sample's and the hypothesized population's:

**Formula 3.** $\Delta Std = SI_{std,\ obtained\ sample} - SI_{std,\ hypothesized\ population}$

For the ski-injury example (Figure 6), this effect size is only about 2.3. That means that, even discounting margin of error and supposing "statistical significance", there is only a 2.3 percentage point shift along the severity scale, from the null case to the sample group's. Clearly, this shift, even if it is real, is very small. The implication is that, in general, outcome severities have scarcely changed from the null model; and if any specific incident turns out by chance to have an extreme outcome, it is likely associated with some unique, unmodeled circumstance, rather than being explainable by the small, general severity shift.

The *p*-value displayed in Figure 6 is determined by resampling, as follows (for more details about the Excel algorithm, see the Appendix which follows the References): Under the null hypothesis, the actually obtained sample is just one random instance of the size-*n* samples that could have been drawn from the hypothesized population. The actually obtained effect size ΔStd is the sample statistic. We interpret the relative frequencies of the severity categories in the null as giving the expected probability distribution for the severities to occur within particular samples. Thus, we can simulate the drawing, sequentially, of *many* random, size-*n* samples from the hypothesized population, and calculate the effect size (sample statistic) for each case. The distribution of a set of many such simulated effect sizes approximates the sampling distribution for the sample statistic, as expected for samples of size *n* taken from the null population. For one-tail testing, the proportion of those sample statistic (effect-size) values in the sampling distribution that equal or exceed the actually obtained value in the sample is the *p*-value, which can be used (with caution) as a possible indicator of significance.

The tests are usually right tail because, when testing for a significant difference in a severity distribution, compared to a baseline, one is generally concerned with shifts towards greater severity. It is possible, of course, to consider a factor that is associated with an apparent *downward* shift of severities, in which case a left tail test would apply. In the attached template in Excel, if a downward severity shift is detected, then the display also adds the *p*-value from a left tail perspective.

The final output when using the proposed method is a 95% confidence interval for the effect size. Just the point estimate for effect size does not account for the variability in the sampled data. In Figure 6, this additional information provides a further indication that the sampled group may not be meaningfully different from the general population: In the confidence interval, we see that the effect size might actually be negative, and if positive, only by a little.

As in the example just described, it is often the case that if the *p*-value is "not significant" (e.g. *p*-value > 0.05), then the confidence interval (CI) for effect size straddles the value zero (0), and vice versa. But exceptions can occur for two reasons: (1) The *p*-value is calculated from a one tail model, whereas the confidence interval is here splitting the variability onto two tails. (2) The variance underlying the *p*-value calculation is based on data in the model for the full, hypothesized population, whereas the variance for calculating CI is found directly from the sample data. The two estimates may differ.

Procedurally, the proposed method requires a separate cycle of bootstrapping, from the sample data of interest, to determine the confidence interval. The procedure cannot directly use the results from the foregoing *p*-value procedure. As mentioned earlier, the CI estimate draws its data from the sample itself (there is no 'null' premise for the population distribution), so the procedure re-simulates the sampling distribution for the effect size - but this time using the sample itself as the best model

available for the true probability distribution of outcome-severities. The set of all the effect-size outputs following many repetitions simulates the sampling distribution for the effect-size sampling statistic; and the values lying on 2.5th and 97.5th percentiles, respectively, of that distribution, are the estimates for the lower and upper bounds of the 95% confidence interval for the true parameter.

*Method 2*

For those without access to the attached templates or other resampling program, an alternative version of the proposed method is provided, based on the familiar *t*-distribution. This works because the sample statistic is really a mean (i.e. the group mean of the standardized severity levels), so for large enough samples, the central limit theorem applies to its sampling distribution. (You can test this by experimenting with the attached templates; the sampling distribution is simulated by the output set of a program run, which the computer then graphs as a histogram.) The *t*-distribution model is imprecise (a) when sample sizes are small, because severity distributions are typically quite skewed, and (b) due to the data's discreteness (i.e. the defined severity levels are discrete.). Nonetheless, experiments with the template, which readers can replicate, suggest the relative robustness of the model.

For this parametric approach, the only real innovation is the construct of the standardized severity index - for each of the null population and the obtained sample. The effect size measure is simply the difference between sample's $SI_{std}$ (which is now the sample statistic) and the null's $SI_{std}$. Note that, for convenience, the calculations based on Method 2 are also automated and included in the attached Excel file, in the fourth tabbed worksheet.

Figure 7 highlights some steps for parametrically obtaining the *p*-value. In this approach, the sample statistic is the standardized severity index for the obtained sample. But we convert the sample statistic to a *test statistic*, based on how many standard errors the sample statistic is from the expected parameter value (i.e. from $SI_{std}$ for the hypothesized population).

**Formula 4.** $$\text{Test statistic } t = \frac{SI_{\text{std, obtained sample}} - SI_{\text{std, hypothesized population}}}{\text{(Standard Error for the Sampling Distribution of the Sample Statistic)}}$$

That distance, in the numerator, we have calculated previously (per Formula 3) as the effect size (ΔStd), shown at the top right in Figure 7.

Typically, the standard error for the denominator would be estimated as $s/\sqrt{n}$—because *s* is usually the only estimate available for the hypothesized population's standard deviation σ. In this application, we can find a better estimate for σ: $\sigma_{est.}$—based on the explicitly hypothesized frequency distribution for the null. This is the formula for estimating the standard deviation of the hypothesized population, based on its frequency distribution:

**Formula 5.** $$\sigma_{est} = \sqrt{\frac{\sum_{i=1}^{k}(w_i \times (x_i - SevMean_{std})^2)}{n_{null} - 1}}$$

where $x_i$ = the standardized equivalent to the *i*th severity level (of the *k* levels available) (standardized per Formula 2)

$SevMean_{std}$ = the frequency-weighted mean for the population's standardized severity levels*

$w_i$ = the weight for the *i*th severity-level—using the frequencies as the weights

$n_{null}$ = the sum of the frequencies**

* $SevMean_{std}$ is equivalent to the previously calculated $SI_{std, hypothesized population}$, shown at the upper right in Figure 7. In a previous section, we calculated the frequency-weighted mean for the severities, then standardized the result; as interpreted here, we standardize the severities first, then find the weighted mean for the standardized values.

** $n_{null}$ includes all counts in the hypothesized population distribution. It is *not* a population size *N*, because generally we do not have access to a full population of all relevant cases. Most likely, the population is larger than $n_{null}$, and the standard deviation formula reflects (by using $(n_{null} - 1)$) this sample-like aspect of the available reference distribution. Nonetheless, the null assumption is that the relative frequency distribution of the full population *is* reasonably accurately represented in the available data distribution being used as the null model.

Acknowledging the sample-like aspect of the distribution used for the null model, one might suggest using a two-sample t-test for this parametric approach (with the reference distribution being one sample), rather than interpreting the test as a one-sample t-test, in which a sample is compared to a hypothesized population distribution. In the author's experience, however, in the field, the latter model seems to better reflect the conditions for conducting these tests. For example, in a uranium mine, one might query whether the miners' exposures to elevated radiation levels are greater in a particular mine level (the sample) compared to in the mine, generally (the population). But there is no really independent way to measure radiation exposures 'in general' in the mine; one only has exposure data from actual workers' dosimeters, which are dispersed in particular places in that environment - widely dispersed, but not into every possible location or condition in the mine. Under the null hypothesis (that workers in all levels in the mine are subject to the same radiation-exposure-distribution), the aggregated distribution from

| | {Use as the **w**eights} | {Use as the **x**'s | | | |
|---|---|---|---|---|---|
| Sev'ty Levels | Freq's (in $H_0$) | Standardized Severity Level | w · (x - $SevMean_{std}$)$^2$ | *Calculate these values the same as for Method 1:* | Point Estimate for Effect Size: |
| 1 | 285 | 0.0 | 168588.3626 | Standardized Freq'cy-Weighted Mean Severity Level for Population | |
| 2 | 285 | 33.3 | 23145.45427 | (= $SevMean_{std}$) = $SI_{Std}$ for Hyp'd Pop'n = 24.321557 | |
| 3 | 53 | 66.7 | 95034.74246 | Sample size *n* = 301 | 2.26 |
| 4 | 28 | 100.0 | 160362.3504 | | |
| 5 | 0 | | | Formula-estimated standard error = $\sigma_{est.}/\sqrt{n}$ = 1.5117411 | |
| 6 | 0 | | | | |
| 7 | 0 | | | *Calculate the p-Value based on t distributed sampling distribution:* | |
| 8 | 0 | | | *Test statistic t* = {Point Estimate for Effect Size / Std.Error}) = 1.49266 | |
| Totals: | 651 | | 447130.9097 | *Degrees of Freedom* = 300 ) | |
| Estimated Pop'n Variance for St'dized Sev. levels: | | | 687.8937072 | = (column total / ($n_{null}$-1)) | Formula-estimated, one-tail: p-Value ≈ 0.06829 |
| Estimated Std. Dev'n ($\sigma_{est.}$) for St'dized Sev. Levels: | | | 26.22772783 | = squareroot(variance) | {Compare this Method1-based value, once resampling is run: 0.055 } |

**Figure 7.** Parametric Calculations for the *p*-Value for a Difference in Severity Distribution

| | {Use as the **w**eights} | {Use as the **x**'s | | | |
|---|---|---|---|---|---|
| Sev'ty Levels | Freq's (in Sample) | Standardized Severity Level | w · (x - $SevMean_{std}$)$^2$ | *Calculate these values the same as for Method 1:* | Point Estimate for Effect Size: |
| 1 | 105 | 0.0 | 74171.36676 | Standardized Freq'cy-Weighted Mean Severity Level for Sample | |
| 2 | 160 | 33.3 | 7301.366553 | (= $SevMean_{std}$) = $SI_{Std}$ for Sample = 26.578073 | |
| 3 | 28 | 66.7 | 44998.66938 | Sample size *n* = 301 | 2.26 |
| 4 | 8 | 100.0 | 43126.23481 | | |
| 5 | 0 | | | Formula-estimated standard error = s/√n = 1.3704589 | |
| 6 | 0 | | | | |
| 7 | 0 | | | Margin of Error for the Confidence Interval: ± 2.697 | |
| 8 | 0 | | | Based on: SE x t[critical(α,n-1)] = 1.3705 x t[critical(0.05,301-1)] | |
| Totals: | 301 | | 169597.6375 | = 1.3705 x 1.9679 | |
| Estimated Sample Variance for St'dized Sev. levels: | | | 565.3254583 | = (column total / (n-1)) | |
| Sample standard deviation *s* for St'dized Sev. Levels: | | | 23.77657373 | = squareroot(variance) | |
| | | | | Confidence Interval = [Point Est.] ± [Margin of Error] = 2.26 ± 2.7 | |
| | | | | Formula-Based C.I. Boundaries: -0.44 up to 4.95 | |
| | | | | {Compare these Method1 results, once resampling is run: | |
| | | | | C.I. Boundaries: Lower: -0.4013 Upper: 4.9143 | |

**Figure 8.** Calculations for 95% Confidence Interval for Difference in Sev. Distributions

all available dosimeters, from whatever work area, is taken to offer the best approximation for the 'true' population distribution, if the null is true. The distribution in the one mine area of special interest can be compared to that reference.

Calculations reflecting Formula 5 are illustrated on the left in Figure 7; the calculated $\sigma_{est.}$ appears near the bottom, centre. As illustrated near the right, middle, in the figure, we divide $\sigma_{est.}$ by the square root of the sample size *n* (for the obtained sample) to estimate the standard error for the sampling distribution. We can now complete the calculation for the test statistic *t*; and knowing the degrees of freedom (*n*-1), we calculate the *p*-value conventionally. Observe at the bottom right of the figure that, for the illustrated data, the calculated *p*-value is close to the result obtained earlier by resampling.

A parametric model can also be used to calculate a confidence interval for the effect size. Some basic steps are illustrated in Figure 8. At the right, middle, of the figure, we see the conventional formula for calculating the Margin of Error for a 95% confidence interval:

Margin of Error = ±[(standard error for the sampling distribution) × ($t_{critical,\alpha=0.05,df=n-1}$)]

The standard error is calculated conventionally, as $s/\sqrt{n}$, but in this case the standard deviation *s* is determined by formula from the grouped data in the sample itself; there is no null assumption for the distribution.

**Formula 6.** $$s_{\text{(based on the obtained sample)}} = \sqrt{\frac{\sum_{i=1}^{k}(w_i \times (x_i - SevMean_{std})^2)}{n-1}}$$

where $x_i$ = the standardized equivalent to the *i*th severity level (of the *k* levels available)

$SevMean_{std}$ = the frequency-weighted mean for the *sample's* standardized severity levels (this is equivalent to $SI_{std,obtained\ sample}$)

$w_i$ = the weight for the *i*th severity-level—using the frequencies as the weights

*n* = sample size = the sum of the frequencies

For the final steps of finding a confidence interval for the mean, we apply the conventional formula:

CI = [Point Estimate] ± [Margin of Error]

But note that the point estimate of current interest is *for the effect size*, not for the population mean, per se. The effect size, we recall, is:

$(SI_{std,\ obtained\ sample} - SI_{std,\ hypothesized\ population})$.

For a one-sample estimate of the mean, we would simply wrap the margin of error around the sample mean which equals $SI_{std,obtained\ sample}$. Mathematically, however,

$(SI_{std,\ obtained\ sample} \pm \text{Margin of Error}) - (SI_{std,\ hypothesized\ population})$

is equivalent to a calculation for effect-size CI:

$(SI_{std,\ obtained\ sample} - SI_{std,\ hypothesized\ population}) \pm (\text{Margin of Error})$

## Cases

Included in the Excel file for this paper are five cases, each on a separate worksheet tab, that can be used for experimenting with the methods described here. Both the re-sampling approach (Method 1) and the parametric-approximations (Method 2) can then be run. (Additional documentation is included on the spreadsheets, and amplified in the Appendix.)

Each case page includes, for comparison purposes, the result of using a chi-square test (for goodness of fit, or for independence (i.e. using a 2 × 2 reduction of the distributions), as appropriate.) The chi-square results were generated by Minitab® 16.1.1, as also were Minitab's "warnings" where the ideal assumptions for running the chi-square are not met.

Four of the cases are drawn from real industrial-safety contexts, but with exact values and descriptions modified for confidentiality. The fifth case invites a comparison between the severity data for a particular (hypothetical) company, versus the benchmark of severity distributions for a corresponding administrative region and time period (from data in Association of Workers' Compensation Boards of Canada, 2010).

## Conclusions

This paper has described the need for new procedures than can validly and reliably assess differences between severity distributions. As shown, not all hazardous situations have the same proportional risks that, given occurrence of an incident, its outcomes will be severe. Clear methods have not been available for assessing whether, and by how much, the severity-risk distributions differ from one case of interest to another. When quantitative analyses are reported, they often have focused on occurrences of just one severity level at a time (e.g. fatalities), which may or may not reflect a general, across-the-board increase in severities between contexts.

In response to this need, two related methods have been presented for assessing differences between severity distributions, both in terms of their effect size (including a confidence interval) on a readily understood scale, and a *p*-value. The resampling-based method has the advantage of limiting requirements for parametric assumptions. The attached Excel file includes a template for running such analyses. An alternative, parametric metric does not require automation (although this is also provided, for convenience), and it is demonstrably robust to most circumstances. Real-world based cases have been provided to enable readers to experiment with the methods.

## References

Anderson, R.T., James, M.K., Miller, M.E., Worley, A.S., & Longino, C.F., Jr. 1998. The timing of change: Patterns in transitions in functional status among elderly persons. *Journal of Gerontology: Social Sciences*. 53B(1), S17-S27.

Association of Workers' Compensation Boards of Canada (AWCBC) 2010. Key statistical measures for 2008. Web-published sub-page of Key Statistical Measures (KSMs) - Data Tables presentation, updated February 2010.

Bird, F.E., Jr. & Germain, G.L. 1986. Practical loss control leadership. Loganville, Ga: Institute Publishing; International Loss Control Institute.

Blank, S., Seiter, C., & Bruce, P. 2001. Resampling Stats in Excel. Version 2.. Arlington,VA: Resampling Stats Inc.

Cattermole, T.J. 1999. The epidemiology of skiing injuries in Antarctica. *Injury: International Journal of the Injured*, 30(7), 491-495.

Charbotel, B., Martin, J.-L., Gadegbeku, B., & Chiron, M. 2003. Severity factors for truck drivers' injuries. *American Journal of Epidemiology*, 158(8), 753-759.

Conroy, C. 1989. Work-related injuries in the meatpacking industry. *Journal of Safety Research*, 20(2), 47-53.

Corra, S. & De Giorgi, F. 2007. Sledding injuries: Is safety in this winter pastime overlooked? A three-year survey in South-Tyrol. *Journal of Trauma Management and Outcomes,* 1(5), [no page numbers].

Ellis, P.D. 2010. *The essential guide to effect sizes*. Cambridge, et al: Cambridge University Press.

Foote, AJ.. 2002. Is probabilistic risk assessment the answer? Presented at the 11th International Process & Power Plant Reliability Conference. Houston, Texas.

Girardi, P, Braggion, M, Sacco, G., De Giorgi, F., & Corra, S. 2010. Factors affecting injury severity among recreational skiers and snowboarders: An epidemiological study. *Knee Surgery, Sports Traumatology, Arthroscopy*, 18(2), 1804-1809.

Goodman, W.M. 1992. Fatality-risk reduction: Measuring progress through comparative, analytic techniques. Ontario Hydro Report #HSD-ST-92-15.

_____ 2010. The undetectable difference: An experimental look at the 'problem' of p-values. JSM Proceedings. Alexandria, VA: American Statistical Association.

Heinrich, H.W., Petersen, D., and Roos, N. 1980. *Industrial accident prevention: A scientific approach. Fifth edition*. New York et al: McGraw-Hill Book Company.

Hesterberg, T., Monaghan, S., Moore, D.S., Clipson, A., & Epstein, R. 2003. *Bootstrap methods and permutation tests: Companion chapter 18 to the practice of business statistics.* New York: W.H. Freeman and Company.

Hotte, P.W. & Hotte, V.E. 1990. An epidemiological study of electrical accidents. Report for the Canadian Electrical Association. Report #293-T-592.

Howe, J. 2007. Risk in adventure/outdoor education activities. Web-published presentation posted on the University of Wisconsin website, at http://www.uwlax.edu/sah/ess/pe/files/Risk_in_Adventure.ppt

Hu, S.-R., Li, C.-S., & Lee, C.-K. 2010. Investigation of key factors for accident severity at railroad grade crossings by using a logit model. *Safety Science*, 48, 186-194.

Kriebel, D. 1982. Occupational injuries: Factors associated with frequency and severity. *International Archives for Occupational and Environmental Health*, v.50, 209-218.

Lozada-Larsen, S.R. & Laughery, K.R.,Sr. 1987. Do identical circumstances precede minor and major injuries?" In Proceedings of the Human Factors Society: 31st Annual Meeting, Human Factors Society. pp. 200-204.

Manuele, F.A. 2004. Injury ratios: An alternative approach for safety professionals. *Professional Safety*, 49(2), 22-30.

_____ 2008. Serious injuries and fatalities: A call for a new focus on their prevention. *Professional Safety*, 53(12), 32-39.

_____ 2003. Severe injury potential. *Professional Safety*, 48(2), 26-31.

Minitab 2010. Software: Minitab® 16.1.1. Minitab, Inc.

OrlandoHealth Surgical Critical Care Fellowship 2009. Injury severity scoring. Web published at: http://www.surgicalcriticalcare.net/Resources/injury_severity_scoring.pdf

Petersen, D. 1998. *Safety Management: A Human Approach. Third Edition*. Des Plaines, IL: American Society of Safety Engineers.

Pinnacle West Capital Corporation 2006. 2006 PNW/APS safety pyramid. In 2006 Corporate Responsibility Report. Web-published at http://www.pinnaclewest.com/main/pnw/AboutUs/commitments/ehs/2006/charts/charts06_22.html

Sormani, M.P., Molyneux, P.D., Gasperini, C., Barkhof, F., Yousry, T.A., Miller, D.H., & Filippi, M. 1999. Statistical power of MRI monitored trials in multiple sclerosis: New data and comparison with previous results. *Journal of Neurology, Neurosurgery, and Psychiatry*, 66(April), 465-469.

Sutherland, A.G., Johnston, A.T., & Hutchison, J.D. 2006. The new injury severity score: Better prediction of a functional recovery after musculoskeletal injury. *Value in Health*, 9(1), 24-27.

Takakuwa, T. & Endo, S. 1997. Factors determining the severity of ski injuries. *Journal of Orthopaedic* Science, v. 2, 367-371.

U.S. Nuclear Regulatory Commission 1975. Reactor safety study. USNRC Report (NUREG-75/014), WASH-1400

Wausau Insurance Companies 2007. Accident pyramid: Guide to a proactive approach. Focus on Safety: Safety Tips from Wausau Insurance Companies. Web-published at http://www.bbr.org/safety/002_Accident Pyramid.pdf

## Appendix

**Notes on the Attached Excel Worksheet**

All data are in the file **Severity Shift Calculations_ Effect Size_p_&_CI.xls**.

**Tab: Inputs Worksheet**

There are three data Inputs required to run all the procedures described in this paper: (1) The severity distribution for the hypothesized population or baseline; (2) the severity distribution for the obtained sample; and (3) the number of defined severity categories. The Inputs worksheet instructs where to make those inputs and provides preliminary outputs, such as the raw and standardized severity indexes.

**Tab: Effect Size & p-value**

The left portion of this page re-caps the inputs, and displays the point estimate for the Effect Size, in the units described in this paper. Once the macro button ("Click to Generate Outputs") is clicked, resampling is used to produce and display a (right-tail) *p*-value in Row 22. (If the system detects a negative effect-size, then the screen also displays a left-tail *p*-value in Row 23.)

Intermediate steps are recorded and documented in columns BA to BL. A single resample, of the size of the actually obtained sample, is drawn randomly from a probability distribution consistent with the hypothesized population distribution. The effect size for this one resample is calculated and recorded in cell BL9. When the Macro is run (see above), 5000 such re-samples are calculated, independently, and their effect sizes are recorded in column BM, as well as graphed as a histogram in Columns S to AA. (Following Hesterberg *et al.*(2003), the number of resamples used was larger than the common choice of 1000 (compare Blank, Seiter & Bruce (2001), and Sormani *et al.* (1999)), since the greater number helps to smooth and remove gaps in the distribution, and generally "introduces little variation".) The *p*-value finally displayed in cell K22 is the proportion of the 5000 resampled effect sizes that are at least as large as the actually obtained effect size, in L2.

**Tab: Confidence Interval**

This page is organized similarly to the preceding. Press the Macro ("Click") button to begin estimating the boundaries of the confidence interval for the effect; they will be displayed in the range I22:L23. Again, intermediate steps are columns BA to BL, which randomly generate a single sample of the size of the actually obtained sample, and find its corresponding effect size. This time however, the probability distribution used to generate the resample is the original, obtained sample; there is no null assumption for the distribution. Resampling repeats this process 5000 times. The resulting, estimated sampling distribution of effect sizes is graphed, and the $2.5^{th}$ and $97.5^{th}$ percentiles from the distribution provide the CI boundaries.

**Tab: Formula-Based Versions**

This page applies and illustrates the procedures, from the *Method 2* section of this paper, to find parametric approximations for the *p*-value (cell J17) and the confidence interval for Effect Size (range G35:J35). In neighboring cells, for comparison, the resampling-based results are shown (assuming those procedures have been run).

**Tabs (5) for Cases**

(These pages are described in the *Cases* section of this paper.)

Correspondence: Bill.Goodman@uoit.ca