Adverse Impact Toolkit

Version 4.2c

December 2009

Biddle Consulting Group, Inc.
193 Blue Ravine, Suite 270
Folsom, CA 95630
800-999-0438
www.biddle.com

Contents

Program Overview

System Requirements

Before opening and using the Adverse Impact Toolkit, be sure that your computer is equipped with at least the following:

- CD ROM drive
- Pentium III processor (or higher)
- 512 mb of RAM minimum **(1 gig or more recommended)**
- 20 mb of free hard drive space
- Super VGA monitor, 800x600 resolution minimum, 1024x768 recommended
- Microsoft Office 2000/XP/2003
- Excel must also be configured so that **macros are enabled** for the Program to operate

Overview

This Program is designed for use by human resource and EEO professionals to aid in **evaluating practices, procedures, and tests for adverse impact**. Adverse impact is a statistical concept that is generally used to mean there is a "substantially different passing rate" between two groups on a practice, procedure, or test (please see our website, www.uniformguidelines.com for a more detailed discussion). Since the late 1970s, the U.S. courts have defined, re-defined, limited, and expanded the use of adverse impact analyses for evaluating various personnel transactions such as layoffs, hires, promotions, transfer, and (more recently) compensation. Because of this ever-changing and dynamic condition, and because the courts typically consider the statistical evidence provided by adverse impact analyses in a context with other types of data and evidence, a thorough legal discussion about adverse impact is not provided in this Manual. However, because some of the methods in this Program are actually derived from court cases, some brief case references are provided.

The Program consists of a Microsoft® Excel® Workbook that includes five visible sheets and several hidden sheets (these sheets perform calculations and are irrelevant to the user). The worksheets include numerous features that use embedded programming and macros to complete the calculations. Microsoft Excel was chosen as a development platform for this Program to allow the user to readily import and export data and relevant analysis results.

This Manual is designed to provide the user with instructions for operating the worksheets and interpreting their results. It should be noted that some degree of background in statistics and human resources is needed to operate these worksheets and effectively interpret their results.

The Concept of Statistical Significance

Karl Pearson formed the foundation of statistical significance testing as early as 1901, and now a century later researchers are still applying the concept as the acid test for deciding whether the results of their studies are "on to something meaningful." While the concept of statistical significance can be applied to literally hundreds of different types of statistical analyses, the meaning is essentially the same: If the result of a statistical test (i.e., the final product of its calculation) is "statistically significant," it is *unlikely to have occurred by chance*. Said another way, a "statistically significant finding" is one that raises the eyebrows of the researcher and surfaces the thought "I think I've found something here, and it is not likely due to chance."

Statistical significance tests result in a ***p***-value (for ***p***robability). P-values range from 0 to 1. A p-value of 0.01 means that the odds of the event occurring by chance is only 1%. A p-value of 1.0 means that the there is essentially a 100% certainty that the event is "merely a chance occurrence," and cannot be considered a "meaningful finding." P-values of .05 or less are said to be "statistically significant" in the realm of EEO analyses. This .05 level (or 5%) corresponds with the odds ratio of "1 chance in 20."

Statistical Power

When a statistical test is calculated to evaluate whether the event being analyzed is "statistically significant," there is always a "power" associated with that test which can be used to describe its ability to reveal a statistically significant finding *if there is one to be found*. Thus, a *powerful* adverse impact analysis is one that has a *high likelihood* of uncovering adverse impact if it exists. For most statistical tests used in this Program, there are three factors that impact the statistical power of the analysis:

(1) **Effect size**, which in the "selection rate comparison" type of adverse impact analyses means the *difference in selection rates between the two groups* (e.g., if the male passing rate on a test was 80% and the female passing rate was 60% there is a 20% "gap" or *effect size* between the two groups), and for the "availability comparison" type of adverse impact analysis means the difference between a group's current representation versus their availability;

(2) **Sample size** of the group(s) in the analysis (e.g., the number of men and women); and

(3) The **type of statistical test** used, which includes the actual calculations of the adverse impact analyses and whether a one-tail or a two-tail test for significance is used.

Because the effect size is totally out of the researcher's control, only the latter two factors will be discussed here.

Sample Size

Amassing a large sample size is perhaps the single most effective way to increase the power of an adverse impact analysis. There are at least five ways this can be accomplished:

(1) **Widen the timeframe of the events being analyzed**. For example, rather than analyzing a selection process for just one year, several administrations from previous years can be included to make an overall historical analysis. In EEO litigation settings, including several events into a single analysis requires the plaintiff to successfully argue a "continuing violation" of the employer's practices over a longer timeframe than just the single at-issue event (e.g., a written test administered in a certain year).

(2) **Combine various geographic regions together into a "regional" analysis**. For example, a plaintiff group could challenge all of the employer's sites in a particular state and demonstrate that these sites operated under similar policies or hiring/promotion strategies.

(3) **Combine events from several jobs, job groups, or divisions**. Combining jobs or divisions into an overall adverse impact analysis is a complicated issue that requires considering several factors. Some of these factors are: the similarity in work behaviors between jobs, job groups, or divisions; statistical homogeneity (discussed in the Multiple Events section of this Manual); and the consistency and similarity of the patterns and practices.

(4) **Combine various selection procedures**. Under some circumstances the courts will allow combining various practices, procedures, and tests into a combined "overall selection process" analysis. While this type of analysis may sometimes yield meaningful results, an event-by-event analysis should be the primary comparison in most circumstances [the 1991 Civil Rights Act requires that a "particular employment practice" needs to be identified as the source of adverse impact for a plaintiff to establish a disparate impact case, unless the *results are not capable for separation for analysis*—see Section 2000e-2(k)(1)(A)(i)].

(5) **Combine ethnic groups**. Some EEO litigation circumstances allow plaintiff groups to form a class of two or more ethnic groups who share similar allegations (e.g., "total minorities").

Important Note! The **first four of the five** aggregation techniques described above require using the Multiple Events Programs in the Toolkit because statistical anomalies can occur when combining data across multiple strata.

When is a Sample "Too Small" for Adverse Impact Analysis?

Despite years of debate among statistical practitioners, there is no bottom-line threshold to the minimum sample size for conducting statistical investigations. If one had to pick a number for adverse impact analyses, it would be 30 with at least 5 expected for success (i.e., hired, promoted, etc.) (OFCCP, 1993). However, in many circumstances a sample with fewer than 30 can still allow for a meaningful statistical analysis to be conducted (e.g., consider a testing situation where 10 men and 10 women applied for a position and all 10 of the men were hired and none of the women).

One characteristic, however, is common to all statistical analyses where small numbers are involved: they suffer from having higher "sampling error" than analyses involving larger data sets. Analyses with "high sampling error" are prone to change in ways that would likely lead to different statistical outcomes if the event hypothetically occurred again. For example, consider a situation where 10 men and 10 women applied for a job and 8 men were hired (80%) and 4 women were hired (40%) (which equates to a substantial selection rate difference of 40% between men and women). It is very likely that subsequent selection processes would have slightly different results. A difference of just a few hires could change the 40% selection rate difference to 10%, 20%, or 30%. In fact, the very next selection process could result in the complete opposite extreme (i.e., 80% selection rate for women and 40% for men). For this reason, adverse impact analyses that are based on small numbers should be viewed as "less stable" than analyses that involve larger sample sizes.

The Uniform Guidelines offers some guidance regarding data sets with small numbers:

> ***Uniform Guidelines, Section 4D***
>
> "Where the user's evidence concerning the impact of a selection procedure indicates adverse impact but is based upon numbers which are too small to be reliable, evidence concerning the impact of the procedure over a longer period of time and/or evidence concerning the impact which the selection procedure had when used in the same manner in similar circumstances elsewhere may be considered in determining adverse impact."

One-tail Versus Two-tail Tests

A one-tail statistical test investigates the possibility of discrimination occurring *in just one direction* (e.g., against women when a "men versus women" comparison is being made). A two-tail test takes the assumption that discrimination *could have occurred in either direction* (e.g., against men or against women) and hence investigates discrimination in *both directions*. A one-tail test is more powerful than a two-tail test when investigating discrimination cases because it requires only a 5% level of significance level in *one direction in the probability distribution*; whereas the two-tail test requires 2.5% on *each end of the probability distribution* (i.e., it assumes that the at-issue group's selection rate could have been either more or less than the comparison group's). Thus, because a one-tail test is focused on finding discrimination only against one group, it will always find statistical significance before a two-tail test will. For this reason, a one-tail test is a more plaintiff-oriented test than a two-tail test. While statisticians themselves disagree on the validity of the one-tailed test in discrimination cases[1], the courts have been almost totally consistent in their requirement of using a two-tail test for significance. The cases cited below have discussed the one-tail versus two-tail issue (and some rather extensively—e.g., the Palmer case), they almost unanimously agree that a two-tail test is the more appropriate technique for adverse impact cases:

Brown v. Delta Air Lines, Inc., 522 F.Supp. 1218, 1229, n. 14 (S.D.Texas 1980).

Chang v. University of Rhode Island, 606 F.Supp. 1161, D.R.I.,1985.

Csicseri v. Bowsher, 862 F.Supp. 547, D.D.C.,1994.

E.E.O.C. v. Federal Reserve Bank of Richmond, 698 F.2d 633, C.A.N.C., 1983.

[1] Compare Harper (1981), Statistics as Evidence of Age Discrimination, 32 Hastings L.J. 1347, 1355 & n. 65 with D. Kaye (1982), The Numbers Game: Statistical Inference in Discrimination Cases, 80 Mich.L.Rev. 833, 841.

Hoops v. Elk Run Coal Co., Inc., 95 F.Supp.2d 612, S.D.W.Va.,2000.

Moore v. Summers, 113 F.Supp.2d 5, D.D.C.,2000.

Mozee v. American Commercial Marine Service Co., 940 F.2d 1036, C.A.7 (Ind.), 1991.

Palmer v. Shultz, 815 F.2d 84, C.A.D.C., 1987.

In addition to the above-referenced cases, the U.S. Office of Federal Contract Compliance Programs (OFCCP) suggests using a two-tail test[2]. While these cases have clearly endorsed the two-tail method, two cases seem to allow a one-tail investigation under certain circumstances:

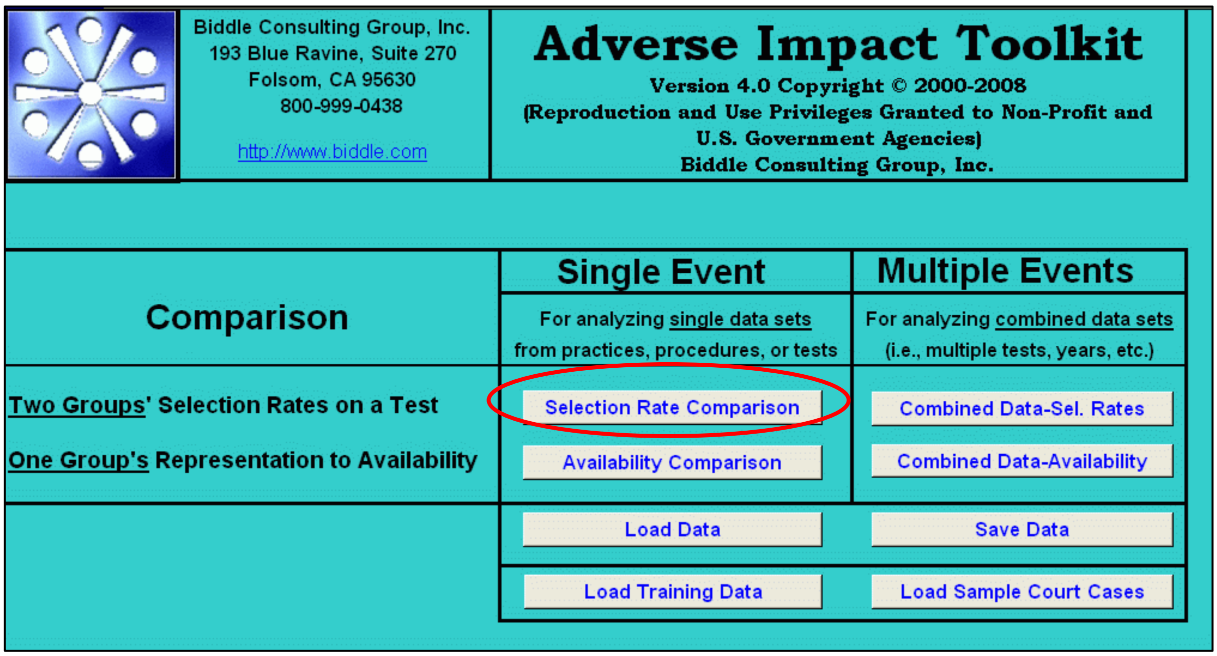
Ottaviani v. State University of New York at New Paltz, 679 F.Supp. 288, S.D.N.Y.,1988 (note that this case did not involve a typical adverse impact analysis, but rather involved using regression analysis for evaluating differences in salary).

Police Officers for Equal Rights v. City of Columbus, 644 F.Supp. 393, S.D.Ohio,1985.

It would seem appropriate to use a one-tail test for adverse impact analyses if the plaintiff can make a showing that the employer's actions were overtly discriminatory. Such actions on the part of the employer would justify a one-tail test because the researcher would have good cause to suspect that the overt discriminatory actions lead to some unequal outcome, hence the use of a one-tail test would only look for what was suspected to exist based on credible evidence.

[2] The "Pooled Two-Sample Z-Score" test specified in http://www.dol.gov/esa/regs/compliance/ofccp/how2/ofcpch3.htm outputs a Z-score value that is a two-tail Z value.

Section 1: Single Event Selection Rate Comparison



Overview

This part of the Program is designed to compare the passing rates of each gender and ethnic group on a single practice, procedure, or test. It may also be used to compare group passing rates on an overall selection or promotion process, although an event-by-event analysis should be the primary comparison in most circumstances [the 1991 Civil Rights Act requires that a “particular employment practice” needs to be identified as the source of adverse impact for a plaintiff to establish a disparate impact case, unless the results are not capable for separation for analysis—see Section 2000e-2(k)(1)(A)(i)].

This type of analysis can be regarded as the “most typical” type of adverse impact analysis, and is specifically explained in the Uniform Guidelines as a “rates comparison” (see Section 4D) that compares the passing rates between two groups (e.g., men and women) on a practice, procedure, or test. This Program can also be used to analyze the outcome of layoffs, demotions, or other similar personnel transactions where there are only two possible outcomes (e.g., promoted / not promoted; hired / not hired, etc.).

Important Note! This part of the Program (the "Selection Rate Comparison") <u>**should not**</u> be used to analyze <u>**combined**</u> sets of data (e.g., analyzing the passing rates of men and women over several years combined). The Multiple Events Selection Rate Comparison Program is necessary when multiple years or tests are placed into a combined analysis. This is because statistical anomalies can occur when combining data across multiple strata.

Data Entry

To input data in this Program, identify the "passing" and "failing" numbers for each gender and ethnic group on a single practice, procedure, or test and enter these numbers into the yellow cells. After the numbers have been entered and double-checked, click the "Calculate" button to complete the exact version of the analysis (the "estimator" calculations will operate without clicking the "Calculate" button). Use the "Clear Data" button to remove the data entered. The analysis results can be printed directly from the screen, or copied and pasted into a word processing file (using the Paste Special command when pasting results into another program).

Analysis Descriptions

This program outputs several types of statistical test results, including:

80% Test

The 80% Test is an analysis that compares the passing rate of one group to the passing rate of another group (e.g., men vs. women). An 80% Test "violation" occurs if one group's passing rate is less than 80% of the group with the highest rate. For example, if the male pass rate on a test was 90% and the female pass rate was 70% (77.7% of the male pass rate), an 80% Test violation would occur. The 80% Test is described by the Uniform Guidelines as:

<u>***Uniform Guidelines Overview, Section ii***</u>

". . . a 'rule of thumb' as a practical means for determining adverse impact for use in enforcement proceedings . . . It is not a legal definition of discrimination, rather it is a practical device to keep the attention of enforcement agencies on serious discrepancies in hire or promotion rates or other employment decisions."

<u>***Uniform Guidelines, Section 4D***</u>

". . . a selection rate for any race, sex, or ethnic group which is less than four-fifths (4/5) (or eighty percent) of the rate for the group with the highest rate will generally be regarded by the Federal enforcement agencies as evidence of adverse impact, while a greater than four-fifths rate will generally not be regarded by Federal enforcement agencies as evidence of adverse impact. *Smaller differences in selection rate may nevertheless constitute adverse impact, where they are significant in both statistical and practical terms* . . ." (emphasis added)

The 80% Test has been scrutinized in Title VII litigation because it is greatly impacted by small sample sizes and does not consider the "statistical significance" of the passing rate disparity between the two groups (see, for example, Bouman v. Block, 940 F.2d 1211, C.A.9 [Cal., 1991]; and Clady v. County of Los Angeles, 770 F.2d 1421, 1428 [9th Cir., 1985]). More typically, courts consider the statistical significance of the passing rate disparity between groups:

Bouman v. Block (citing Contreras, 656 F.2d at 1274-75)

> "Rather than using the 80 percent rule as a touchstone, we look more generally to whether the statistical disparity is 'substantial' or 'significant' in a given case."

Clady v. Los Angeles County, 770 F.2d 1421, C.A.9 (Cal., 1985).

"There is no consensus on a threshold mathematical showing of variance to constitute substantial disproportionate impact. Some courts have looked to Castaneda v. Partida, 430 U.S. 482, 496-97 n. 17, 97 S.Ct. 1272, 1281-82 n. 17 n. 17, 51 L.Ed.2d 498 (1977), which found adverse impact where the selection rate for the protected group was 'greater than two or three standard deviations'[3] from the selection rate of their counterparts. See, e.g., Rivera, 665 F.2d at 536-37 n. 7; Guardians Association of the New York City Police Dept. v. Civil Service Commission, 630 F.2d 79, 88 (2d Cir.1980), cert. denied, 452 U.S. 940, 101 S.Ct. 3083, 69 L.Ed.2d 954 (1981).

Statistical Significance Tests

There are two types of statistical significance tests used by the Program: exact and estimated. The exact test provides the precise probability value of the analysis. Because Title VII is neutral with respect to adverse impact possibly occurring against either group (the group with the highest rate, called the reference group, or the comparison group, called the focal group), this program calculates the two-tail Fisher exact probability. To generate the first tail probability, it begins by calculating the hypergeometric probability for the specific set of entered data (e.g., pass/fail, men/women) and then adds to it every subsequent probability equal to or less than the original in the direction of more extreme results until no members of the focal group are selected.

To generate the second tail probability, the program begins by calculating the hypergeometric probability for the first data set where the selection rate differences between the two groups is larger than the original data set but in the opposite direction (i.e., favoring the reference group), and adds to it every subsequent probability lesser than or equal to the original set in the direction of more extreme results until no members of the reference group are selected. The summed probabilities for each tail are then combined for display purposes.

The widely-endorsed Lancaster (1961) correction has been included as a sensible compromise that mitigates the effects of conservatism of exact methods while continuing to use the exact probabilities from the small-sample distribution being analyzed.

[3] While the courts have ruled that statistical significance occurs when the standard deviation representing the difference in passing rates is "two or three standard deviations . . ." the actual threshold statistically speaking is 1.96 (which equates to a 5% significance level).

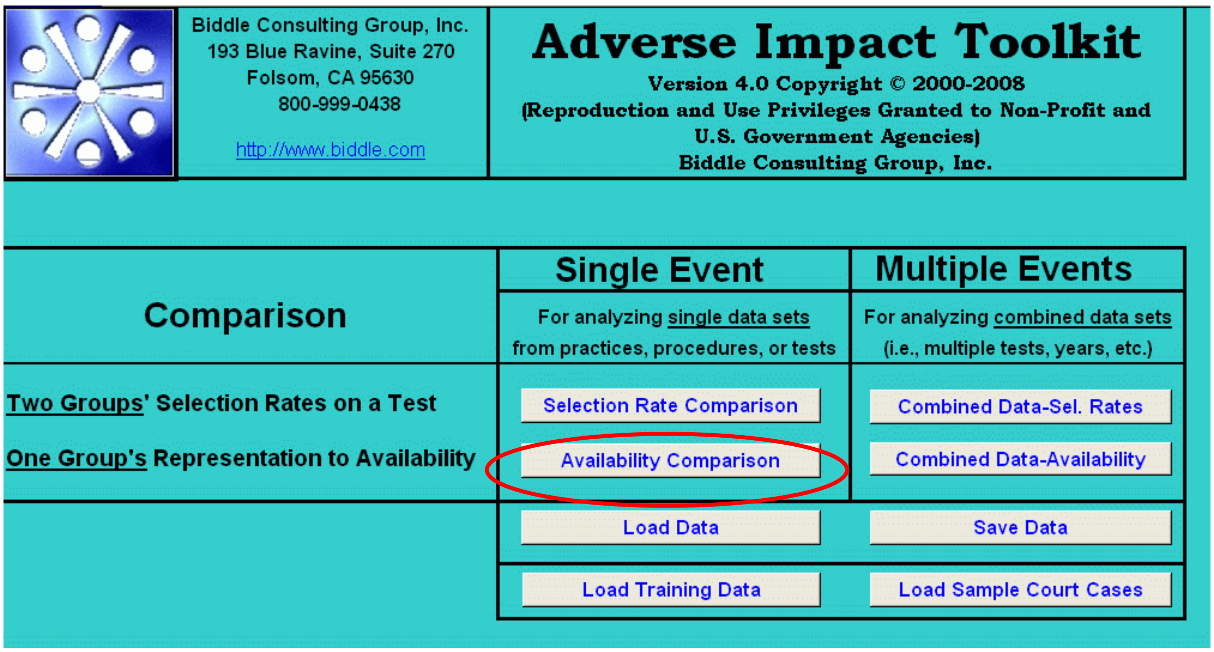
Interpretation of Statistical Tests

These two outputs describe the degree of the statistical test findings. For example, if the "**Likelihood (One Chance In)**" output shows a value of "20," this means that the difference in passing rates between groups (e.g., men vs. women) is so extreme that the odds of it occurring by chance is only 1 in 20, or about 5%. The "**Probability as Std. Deviations**" output describes the probability value (from the statistical test) in terms of standard deviation units, which are sometimes easier to interpret than small probability values. A standard deviation of 1.96 corresponds with a (two-tail) probability value of .05, and a likelihood of 1 chance in 20. A standard deviation of 2.58 corresponds with a probability value of .01, and a likelihood of 1 chance in 100.

The two-tail Fisher Exact Test conducted by the Adverse Impact Toolkit may be replicated using StatXact® and computing the mid-p value as defined in the StatXact user's manual (simply subtract ½ of the point probability of the observed table from the two-tail Fisher Exact Test value). For example, given a dataset where 30 men pass and 21 fail and 22 women pass and 30 fail, the mid-p value may be computed by subtracting ½ of the point probability (.0391 / 2 = 0.01955) from the two-tail Fisher Exact value (0.1162), which equals 0.09665. These data can be obtained from the StatXact output line below:

Two-sided:Pr{FI(X) .GE. 2.782}= Pr{P(X) .LE. 0.0391}= 0.1162

Section 2: Single Event Availability Comparison



Overview

This section of the Program is designed for comparing one group's representation (i.e., the percentage of incumbents in a given position who belong to the gender or ethnic group) to that group's availability in the relevant labor market (using availability data from inside or outside the organization). This type of comparison is useful for determining whether a group is underutilized in a particular position, or group of positions.

This analysis should be differentiated from the "Selection Rate" comparison because (under most circumstances) a statistically significant underutilization does not automatically constitute a finding of adverse impact. The reason for this is straight-forward: a selection rate comparison directly evaluates how two groups fared on a particular employment practice, so if one group significantly outperforms the other, direct evidence is gathered regarding the impact of a *particular employment practice* on the group of interest. Then the attention can shift toward evaluating that particular employment practice for job relatedness (i.e., validity).

By contrast, the Availability Comparison does not (necessarily) consider the impact of one employment practice. Because the comparison is an overall evaluation that considers one group's makeup in a given position compared to their availability outside of position, it does not consider all of the practices, procedures, or tests that may have been used to select or promote individuals for that position. Further, it does not take into consideration other factors such as "job interest" or qualification levels of the at-issue group. For example, if outside availability data shows that men are statistically significantly underutilized for a group of clerical jobs, the underutilization could possibly be explained by either lack of interest on the part of men to pursue these positions, or the fact that men performed poorly on the multitude of qualification screens required for entry into the position (or likely some combination of these two factors and others). For these reasons, the Availability Comparison should be considered as a "threshold" or "initial inquiry test."

Data Entry

To input data into this Program, two sources of information are needed. In the "# Selected/Represented" row, identify the number of individuals in each gender/ethnic group who are currently in the job or group of jobs being evaluated (for an incumbency-to-availability comparison). In the "% Available" row, identify the availability percentage (i.e., the group's representation in the qualified labor market/area) for each group. **Please note that <u>complete data</u> is required for both men and women (when conducting an analysis for women) or all ethnic groups (when conducting an analysis for any ethnic group) for the calculations to operate correctly.** This is because the Program calculates the total number selected/represented by adding the gender and/or ethnic groups together.

After the numbers have been entered and double-checked, click the "Calculate" button to complete the exact version of the analysis (the "estimator" calculations will operate without clicking the "Calculate" button). Use the "Clear Data" button to remove the data entered. The analysis results can be printed directly from the screen, or copied and pasted into a word processing file (using the Paste Special command when pasting results into another program).

Analysis Descriptions

This worksheet includes three statistical outputs, described below:

% Represented

This output shows each group's representation percentage. Note that the Program requires entering data for each gender and/or ethnic group to complete this calculation.

% Over/Under Represented

This output shows the percentage that each group is either over or under represented compared to their respective availability. For example, if the availability for women is 15% and they only make up 10% of the workforce, this output would show "-5%" in red.

Statistical Test Results

There are two types of statistical significance tests computed by the Availability Comparison: exact and estimated. The exact test provides the precise probability value of the analysis. The estimated techniques used by the Program approximate the exact results without requiring lengthy calculations. The exact test uses the (two-tail) exact Binomial Probability Test to assess whether the degree of underutilization is extreme enough to be considered “beyond chance.” Values less than .05 (indicated in red) are “statistically significant.” Because this test compares one group’s representation against their availability (rather than comparing the selection rates of two groups, like the Statistical Test on the “Selection Rate Comparison”), statistically significant findings (without other evidence) should not be considered as direct evidence of adverse impact (because both discriminatory and non-discriminatory reasons can possibly account for the group’s underutilization).

The exact test in this procedure includes the mid-p adjustment that mitigates the effects of conservatism of exact methods while continuing to use the exact probabilities from the distribution being analyzed (see Agresti & Gottard, 2007; Simonoff, 2003; and Armitage, et. al., 2001).

To replicate the Adverse Impact Toolkit two-tail exact mid-p output using Excel®, follow the steps below:

> **Step 1:** Take the observed N, X, P values and compute the exact probability value for that table only. For example, N = total incumbents; X = # focal group members, P = Percentage available, use =BINOMDIST(X, N, P, 0). Be sure to use “0” in this formula because that is needed here is the probability value of the very first (observed) table.
>
> **Step 2:** Take the value from step 1 and multiply it by .5 (i.e., reduce it by one-half).
>
> **Step 3:** Take the observed N, X, P values and create a new table using X-1. For example, for the set N=30, X = 18, P = 75%, use N=30, X = **17**, P = 75%. In other words, just subtract 1 from the X value. If X = 0 then be sure to use 0, and not -1 (the formula needs to be floored with 0 as the lowest value for X).
>
> **Step 4:** Use the *cumulative* BINOMDIST function to compute the one-tail probability value for the N, X-1, and P values defined in step 3 (=BINOMDIST(X-1), N, P, 1). Be sure to use the *cumulative* probability value for the BINOMDIST function by using 1, not 0.
>
> **Step 5:** Add the values from Step 2 and 4 together.
>
> **Step 6:** Double the value obtained in step 4 (Step 4 X 2) and report this as the total two tail exact probability value.

Interpretation of Statistical Tests

These two outputs describe the degree of the statistical test findings. For example, if the “**Likelihood (One Chance In)**” output shows a value of “20,” this means that the degree of underutilization is so extreme that the odds of it occurring by chance is only 1 in 20, or about 5%. The “**Probability as Std. Deviations**” output describes the probability value (from the statistical test) in terms of standard deviations units, which are sometimes easier to interpret than small probability values. A standard deviation of 1.96 corresponds with a (two-tail) probability value of .05, and a likelihood of 1 chance in 20. A standard deviation of 2.58 corresponds with a probability value of .01, and a likelihood of 1 chance in 100.

Important Note! A “**#NUM!**” error message will occur if your computer has insufficient memory to calculate the exact test procedure in this Program with the sample sizes that have been entered. The upper limit (regardless of computer memory) for the sample size used in this Program is 4,000.

As discussed above, the user is cautioned against interpreting statistically significant results in this part of the Program as a "concrete finding of adverse impact" or "evidence of discrimination." Rather, results that yield probability values of less than .05 should be regarded as "statistically significant under-representation," but not necessarily as adverse impact. These findings are sometimes called a "manifest imbalance" or "statistically significant underutilization" in litigation to simply mean that there is a "statistically significant gap" between the comparison group's availability for employment or promotion (typically derived from a group's representation in the qualified applicant pool for entry-level positions and the "feeder" positions for promotional positions[4]) and the group's current representation in the at-issue job or group of jobs. When such a statistically significant "imbalance" or "underutilization" exists, any of the five (5) circumstances listed below can possibly lead to a court's finding of disparate impact:

(1) The employer **failed to keep applicant records** (sometimes referred to as an "adverse inference"—see Section 4D of the Uniform Guidelines). If an employer fails to keep applicant data, the government has reserved the right to infer disparate impact on the selection or promotion process if the agency has an imbalance in a job or group of jobs.

(2) The employer **failed to keep adverse impact data** on the selection or promotional processes (Section 4D of the Uniform Guidelines). Similar to #1 above, if employers have an imbalance in a job or group of jobs and does not have information regarding the adverse impact of the various practices, procedures, or tests used in the selection or promotion process, an "adverse inference" can be made. Employers should maintain passing rate data for their various selection and promotional processes, and practices, procedures, or tests that have adverse impact should be justified by evidence of job relatedness and business necessity.

(3) The employer's **recruiting practice was discriminatory** toward the protected group [see Section 4D of the Uniform Guidelines and Hazelwood School District v. United States, 433 U.S. 299 (1977)]. For example, if the employer recruits for certain jobs only by "word of mouth," and the only applicants who are informed about the job opportunities are of a certain race and/or gender group, the employer could be held liable in a discrimination lawsuit. Plaintiff groups may also argue that minorities and/or women were "funneled" by the employer's systems and processes into filling only certain position(s) in the employer.

(4) The employer **maintained a discriminatory reputation that "chilled" or "discouraged"** protected group members from applying for the selection process (Section 4D of the Uniform Guidelines). This argument has successfully been

[4] If the target position is not underutilized when compared to the relevant feeder position(s), yet the relevant feeder position(s) is underutilized when compared to those with the requisite skills in the relevant labor area (called an "outside proxy group"), it can be argued that the proxy group can be used to compare to the target position. This process is referred to as a "barriers analysis."

made in several discrimination cases[5], and is a viable argument for plaintiffs to make in some circumstances.

(5) The employer **failed to conduct a formal selection process for the position** and instead hired or promoted individuals through an “appointment only” process. This “promotion by appointment” practice would certainly lend itself to a viable plaintiff argument because the practice was exclusionary to qualified individuals who were not allowed an equal opportunity to compete for a position. Further, this type of promotional practice could make the use of conventional adverse impact analyses impossible (because there are no clear “promotional processes” or “events” that can be analyzed by comparing the passing rates between two groups), which could limit the adverse impact analysis to a comparison between the disadvantaged group’s representation in the promotional position to their availability in the “feeder” positions. While informal selection procedures are not directly prohibited under the various civil rights laws, they are much more difficult to defend against claims that they were used unfairly than are more standardized selection processes.

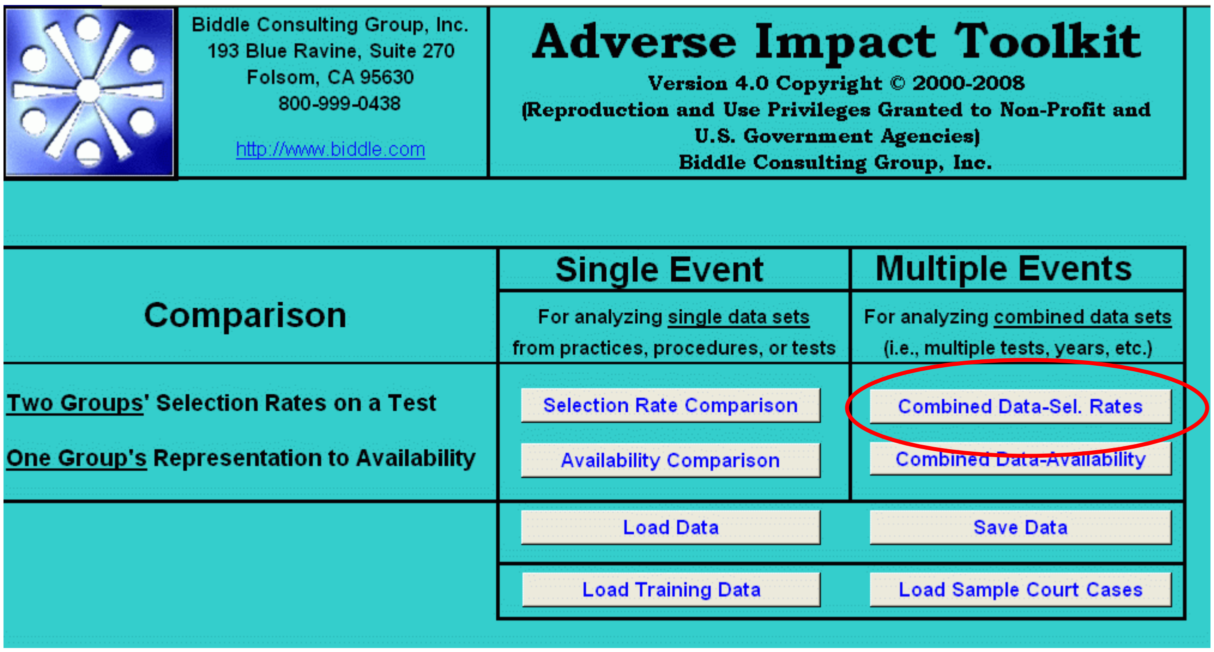
Unless one of these five (5) situations exists, a plaintiff group will be required to pinpoint the specific practice, procedure, or test that caused the adverse impact. The only exception is if the agency’s practices cannot be “separated for analysis purposes[6].”

---

[5] See Donnel v. General Motors Corp, 576 F2d 1292 (8th Cir 1978); Dothard v. Rawlinson, 433 U.S. 321 (1977); Williams v. Owens-Illinois, Inc, 665 F2d 918 (9th Cir), Cert denied, 459 U.S. 971 (1982).

[6] 1991 Civil Rights Act [42 U.S.C. §2000e-2(k)(ii)(B)].

Section 3: Multiple Events Selection Rate Analysis



Overview

This part of the Program is designed for comparing the passing rates of each gender and ethnic group on several combined "events" or administrations of various practices, procedures, or tests. It may also be used to complete an overall adverse impact analysis on several jobs or groups of jobs with similar skill sets. It may also be used to compare group passing rates on an overall selection or promotion process for multiple years, although an event-by-event analysis should be the primary comparison in most circumstances [the 1991 Civil Rights Act requires that a "particular employment practice" needs to be identified as the source of adverse impact for a plaintiff to establish a disparate impact case, unless the results are not capable for separation for analysis—see Section 2000e-2(k)(1)(A)(i)].

> ***Important Note!*** *T*he "Multiple Events" technique used in this Program is necessary when analyzing several Selection Rate Comparisons (e.g., analyzing the passing rates of two groups on a test) rather than combining multiple years of data or multiple tests into a combined analysis on the "Selection Rate Comparison" worksheet. This is because statistical anomalies can occur when combining data across multiple strata.

This type of analysis has been applied in numerous EEO litigation settings. A partial list is provided below:

- Arnold v. Postmaster Gen. (Civil Action Nos. 85-2571, 86-2291, U.S. Dist. Ct. for the District of Columbia, 667 F. Supp. 6; 1987)
- Covington v. District of Columbia (Nos. 94-7014, 94-7015, 94-7022, 94-7107, U.S. Ct of Appeals, 313 U.S. App. D.C. 16; 57 F.3d 1101, 1995)
- Dees v. Orr, Secretary of the Air Force (No. Civil S-82-471 LKK, U.S. Dist. Ct, Eastern District of California, 1983)
- Dennis L. Harrison v. Drew Lewis (Civil Action No. 79-1816, U.S. Dist. Ct for the District of Columbia, 559 F. Supp. 943, 1983)
- Hogan v. Pierce, Secretary, Housing and Urban Development (No. 79-2124, U.S. Dist. Ct District of Columbia, 1983)
- Johnson v. Garrett, III as Secretary of the Navy (Case No. 73-702-Civ-J-12, U.S. Dist. Ct. for the Middle District of Florida, 1991)
- Manko v. U.S. (No. 79-1011-CV-W-9, U.S. Dist. Ct for the Western District of Missouri, 636 F. Supp. 1419, 1986)
- McKay v. U.S. (Civil Action No. 75-M-1162, U.S. Dist. Ct. for the District of Colorado, 1985)
- Paige v. California Highway Patrol [No. CV 94-0083 CBM (Ctx), U.S. District Court, Central District of California, Order Entered August 12, 1999]
- Trout v. Hidalgo (Civ. A. Nos. 73-55, 76-315, 76-1206, 78-1098, U.S. Dist. Ct. of Columbia, 517 F. Supp. 873, 1981)

Data Entry

To input data into this Program, identify the “passing” and “failing” number for each gender and ethnic group on a single practice, procedure, or test and enter these numbers into the yellow cells. Data must be entered for at least two events for proper calculations to be made. Use the “Clear Data” button to remove the data entered. The analysis results can be printed directly from the screen, or copied and pasted into a word processing file (using the Paste Special command when pasting results into another program).

Analysis Descriptions

This worksheet includes three sets of outputs that should be interpreted in sequence:

Step 1 - Pattern Consistency: Can the Events Be Combined Into An Overall Analysis?

These statistical tests evaluate whether the “trend” in the passing rate difference between groups is consistent between events. When one group continually maintains a passing rate that is lower than the other group, the result of this test will be close to 1.0. When there are “flip-flops” between events (e.g., when one group’s selection rates are below the other group's for four years in a row, then in the fifth year the trend reverses), the value will approach 0. When the result of this test is below .05, a statistically significant "flip-flop" of one or more events has been detected, and the user should consider conducting separate analyses for these events or removing them altogether.

The first test uses the "Breslow-Day" (1980) with Tarone's Correction (1988) test which evaluates the appropriateness of the Mantel-Haenszel; the second uses the "Treatment by Strata Interaction" test (Mehrotra, 2000) which evaluates the appropriateness of the Minimum Risk test (see below).

Step 2 - Statistical Test Results: Do the Combined Events Result in Statistical Significance?

If the "Pattern Consistency" test (above) is not violated (i.e., the result was equal to or greater than .05 and the VALID sign appears in Step 2), the second step can be completed. This step allows for the independent "events" to be combined into an overall analysis to see if the combination of all events results in adverse impact. There are three statistical tests to investigate adverse impact in this step:

Mantel-Haenszel Test (Versions 1 & 2)

There are two versions of this (two-tail) test (Version 1 and 2). Version 1 is likely to be closest to the "exact" [7] probability value and only uses a modest (.125) continuity correction.[8] Version 2 uses a conventional correction (.5) and will typically overestimate the probability value (especially with small data sets). These tests assess whether the selection rate difference between two groups (e.g., men vs. women) for all "events" combined is extreme enough to be considered "beyond chance." Values less than .05 (indicated in red) are "statistically significant"; values between .05 and .10 (in orange) are "close" to significance. Both tests use the Cochran version of the Mantel-Haenszel statistic which weights the events by sample size.

Minimum Risk Test

This test does not weight the events by sample size and can sometimes provide a "more powerful" analysis than the Mantel-Haenszel Test (above). The test uses a moderate correction for continuity (.375). This test assesses whether the selection rate difference between two groups (e.g., men vs. women) for all "events" combined is extreme enough to be considered "beyond chance." Values less than .05 (indicated in red) are "statistically significant"; values between .05 and .10 (in orange) are "close" to significance.

[7] Calculating the exact probability for Multiple Event Selection Rate Comparisons requires using advanced statistical software. Currently, there are at least two programs that can be used: StatXact® or WINPEPI (PEPI-for-Windows: "PEPI" is an acronym for Programs for EPIdemiologists. WINPEPI is available free at: http://www.brixtonhealth.com/pepi4windows.html.

[8] A correction to a discrete probability distribution to approximate a continuous probability distribution. This correction adjusts the estimator formula to better approximate the exact probability value.

Mantel-Haenszel Test EXACT

The Mantel-Haenszel technique used in the Adverse Impact Toolkit provides only a probability *estimate*. If the probability value output by the program is between .05 and .15, we recommend computing an ***exact mid-P probability*** using the commercially available StatXact program. The steps for completing this process are described below:

Step 1: Setup the Table Data. For example, if there are two years of promotion data, setup two 2X2 tables (i.e., each table has two rows and two columns):

2009 Promotions	Women	Men
Promoted	24	29
Not Promoted	26	21

2010 Promotions	Women	Men
Promoted	20	30
Not Promoted	30	20

Important: Be sure to setup the "reference" group (the advantaged group with the highest success rate) in the right column and the "focal" group (disadvantaged) in the left column. Also, setup the successful outcome (in this example "Promoted") in the top row and the unsuccessful outcome (in this example "Not Promoted") in the bottom row. Setting up the data in this fashion allows the odds ratios to be interpreted as "the odds of men succeeding over the odds of women succeeding" (see additional discussion in Step 4 below).

Step 2: Compute the Analysis. Select "Statistics," "Stratified 2X2," then "CI on Common Odds Ratio" and "Exact." The output will vary by the version of StatXact being used.

Step 3: Use the output from the program to compute the **Exact Mid-P Probability:**

1	B	C	D	E	F	G	H
2	StatXact Output						
3	Exact p-values for testing that the Common odds Ratio is 1:						
4	One-sided: Pr { S .GE. 59.00 } =					0.0239	
5	Pr { S .EQ. 59.00 } =					0.0121	
6	Two-sided: Method 1: 2 * One-sided					0.0478	
7	Method 2: (Sum of Probs .LE. 0.0121) =				0.0121	0.0478	0.0467
8	Method 3: Pr{|S-Mean| .GE. | 59.00-Mean|} =					0.0478	
9							
10	Mid-P Calculation Steps						
11	Lower P value						0.0239
12	Point Prob						0.0121
13	Lower mid p value=lower-1/2*point EXCEL = H11-(0.5*H12)						0.0179
14	Upper P value=1-lower EXCEL =1-H11						0.9761
15	Upper mid p value =upper+1/2*point EXCEL =H14+(0.5*H12)						0.9822
16	Two sided mid p value=2*min(lower, upper) EXCEL =2*(MIN(H13,H15))						0.0357

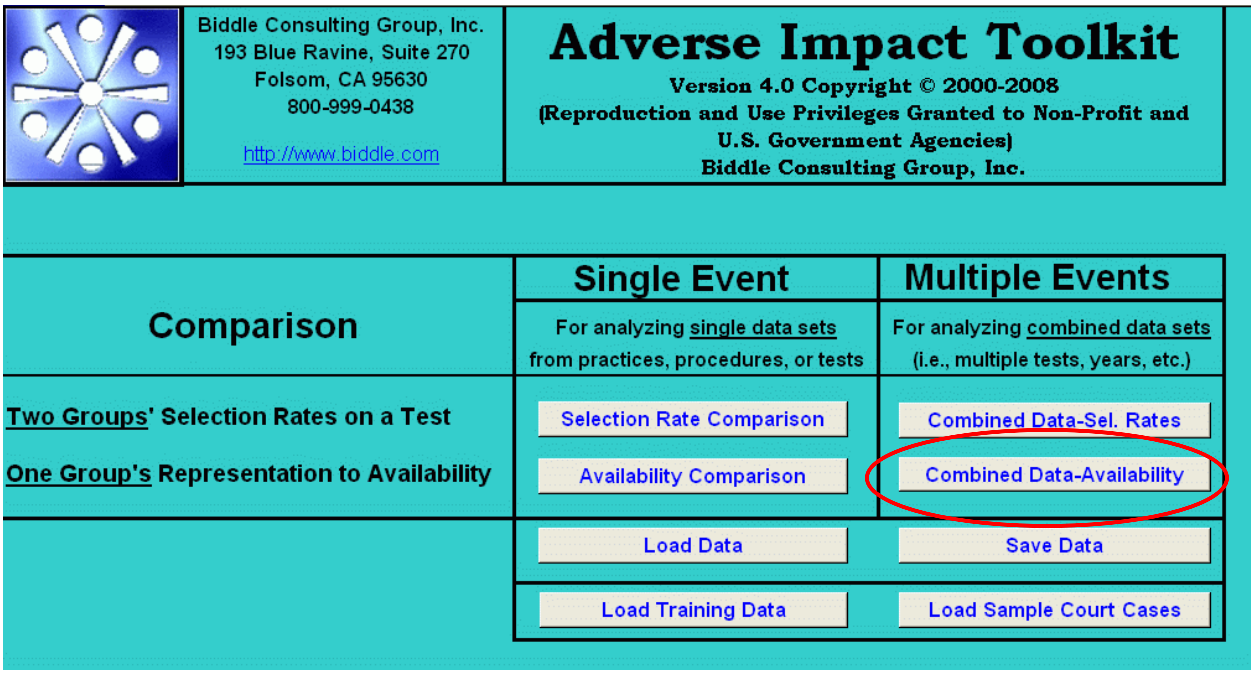
Note that this process uses the *Lower P-value* and the *Point Probability* (the probability of the observed first case) to compute the mid-p value as shown. Using earlier versions of StatXact (e.g., Version 4.0 as shown here), the *Lower P-value* is provided in the "One-sided: Pr { S .GE.}" row and the *Point Probability* is provided in the (One-sided) "Pr { S. EQ. }" row. In more recent version of StatXact (e.g., version 8.0), the *Lower P-value* is provided in the "Exact LE" row and "P-Value, 1-Sided" column and the *Point Probability* is provided in the "Exact: Method 2" row and "Statistic" column.

Step 4: Interpret the Odds Ratio. Using StatXact (and configuring the data in the manner shown above) is useful because the output also includes the "Exact Estimation of Common Odds Ratio." The odds ratio provides the likelihood of a focal group member being promoted compared to a reference group member, which in this example is 1.821 to 1. In other words, the male odds (or likelihood) of promotion is 1.821 higher than females. The range of this odds ratio can be interpreted using the 95% mid-p corrected Confidence Interval provided, which shows the 95% confidence range for this finding is 1.041 (on the low end) and 3.207 (on the high end).

Step 3 – Interpret Degree of Statistical Test Results

These outputs (provided for each of the Statistical Tests above) describe the degree of the statistical test findings. For example, if the "**Likelihood (One Chance In)**" output shows a value of "20," this means that the difference in passing rates between groups (given all of the events combined) is so extreme that the odds of it occurring by chance is only 1 in 20, or about 5%. The "**Probability as Std. Deviations**" output describes the probability value (from the statistical test) in terms of standard deviations units which are sometimes easier to interpret than small probability values. A standard deviation of 1.96 corresponds with a (two-tail) probability value of .05, and a likelihood of 1 chance in 20. A standard deviation of 2.58 corresponds with a probability value of .01, and a likelihood of 1 chance in 100.

Section 4: Multiple Events Availability Comparison



Overview

This part of the Program is designed for making availability comparisons for several different jobs (or groups of jobs) or for the same job across multiple years. The reader is referred to the "Availability Comparison" section of this manual for several caveats that apply to these types of analyses. This type of analysis has been applied in several EEO litigation settings. A partial list is provided below:

- Vuyanich v. Republic National Bank (N.D. Texas 1980). 505 F. Supp. 224
- EEOC v. United Virginia Bank (615 F.2d 147 (4^{th} Cir. 1980)
- Cooper v. University of Texas at Dallas (482 F. Supp. 187, N.D. Tex. 1979)

Data Entry

To input data into this Program, three sources of information are required for each row or "event" (**data must be entered for at least two events for proper calculations to be made**):

(1) **Availability Percentage (%)**: The value entered in this cell should be the **focal group** (e.g., women) percentage of qualified applicants in the relevant labor market. This value can be obtained from "feeder groups" inside the organization or outside labor market data (e.g., census data) depending on how the at-issue job group is filled. At a minimum, job qualifications and the relevant geographical area (e.g., city, county, state, etc.) should be two of the primary considerations when selecting the availability percentage.

(2) **# Selected/Represented**: For each event, enter the number of **focal group** members in the job or job group in this cell.

(3) **Total in Overall Group**: For each event, enter the number of total selections or incumbents (from all groups) in this cell.

The analysis results can be printed directly from the screen, or copied and pasted into a word processing file (using the Paste Special command when pasting results into another program).

Analysis Descriptions

This worksheet includes two sets of outputs that should be interpreted in sequence:

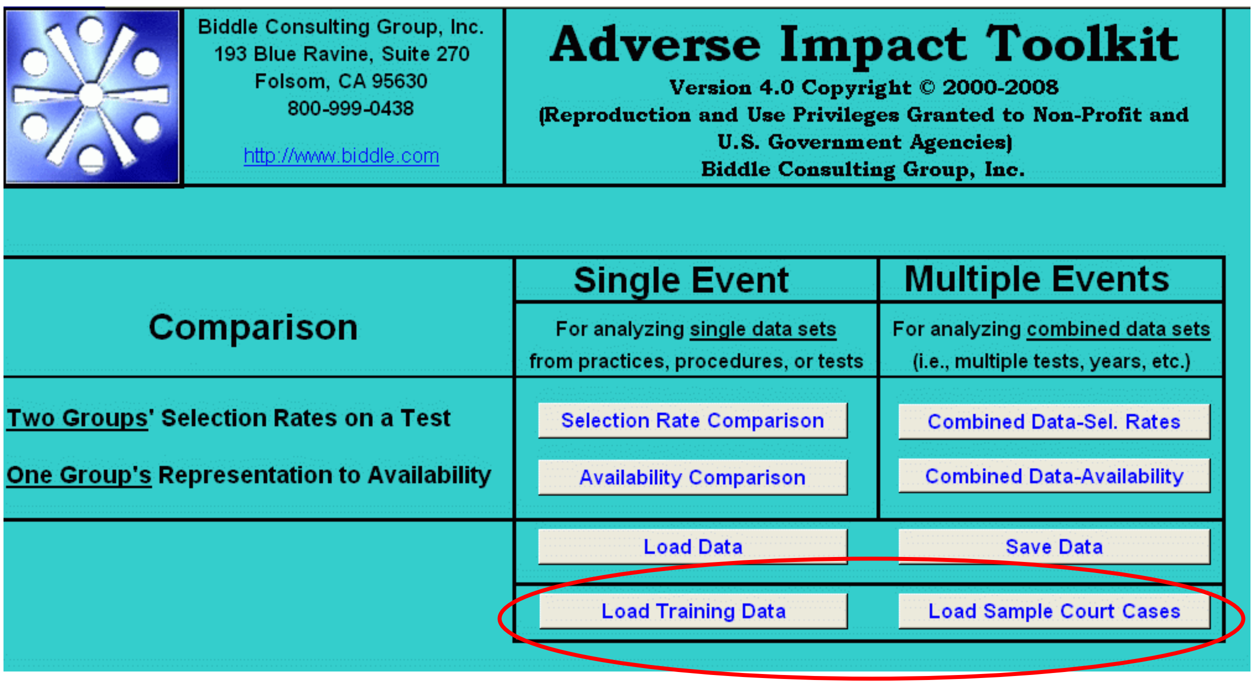
Step 1 - Statistical Test Results: Do the Combined Events Result in Statistical Significance?

This step allows for the independent "events" to be combined into an overall analysis to see if the combination of all events results in statistical significance. This procedure used to calculate statistical significance for the combined set is the Generalized Binomial Test (see Gastwirth, 1987). This (two-tail) test evaluates whether the Focal Group's continual (i.e., across multiple events) under-utilization is statistically significant. Values less than .05 (indicated in red) are "statistically significant"; values between .05 and .10 (in orange) are "close" to significance.

Step 2 - Interpret Degree of Statistical Test Results

These outputs (provided for each of the Statistical Tests above) describe the degree of the statistical test findings. For example, if the "**Likelihood (One Chance In)**" output shows a value of "20," this means that the difference in passing rates between groups (given all of the events combined) is so extreme that the odds of it occurring by chance is only 1 in 20, or about 5%. The "**Probability as Std. Deviations**" output describes the probability value (from the statistical test) in terms of standard deviations units which are sometimes easier to interpret than small probability values. A standard deviation of 1.96 corresponds with a (two-tail) probability value of .05, and a likelihood of 1 chance in 20. A standard deviation of 2.58 corresponds with a probability value of .01, and a likelihood of 1 chance in 100.

Section 5: Training Data and Sample Court Cases



Overview

This part of the Program is designed for training. The Training Data consists of a sample data set for each of the four sections of the Program. Each data set contains values of 30 or less so it can be used with the Demo Version of the Program.

The Sample Court Cases (available in the Full License Version only) contains sample data from actual court cases (cases that were relevant to each type of analysis used in the Program). These cases are briefly described below.

Selection Rate Comparison

The data from this case was taken from Waisome v. Port Authority of New York and New Jersey (758 F.Supp. 171, S.D.N.Y., 1991). This case involved a written test for the position of Police Sergeant that had adverse impact against blacks. The District Court ruled that if just two more black applicants hypothetically passed the test, the statistical significance finding (a value of 2.68 standard deviations, or a probability value of .007, using an "estimator" test) would have changed to a non-significant value (a standard deviation value of 1.89, or a probability value of value of .059 using an “exact” test). This ruling was upheld at the Circuit court level.

Availability Comparison

These data were taken from the infamous Hazelwood School District v. United States (433 U.S. 299, 1977) case (see footnotes 14-17 of this case for a description of the analyses used and related results). This case ruled that black teachers were statistically significantly underutilized in the Hazelwood School District at a level well beyond what could be considered a chance occurrence. This fact, combined with several other factors, was used to draw an inference of discrimination.

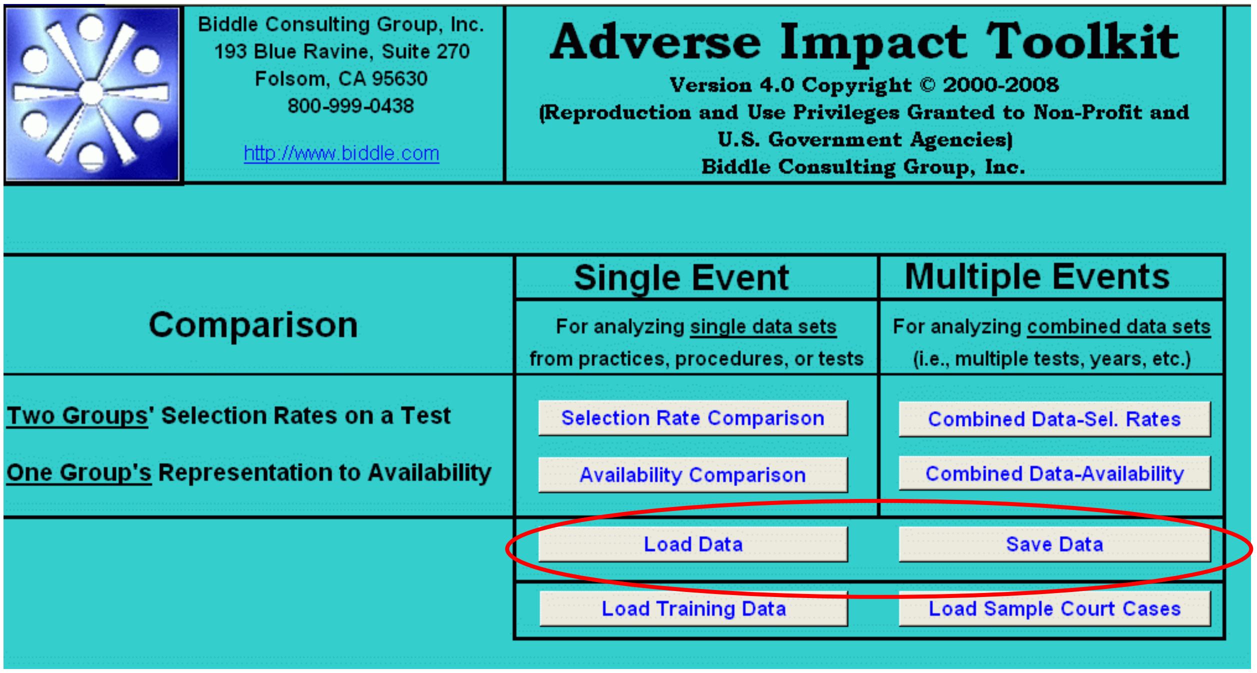
Combined Data-Pass Rate Comparison

These data were taken from the Agarwal v. McKee and Company (16 E.P.D. P 8301, N.D.Cal. 1977) case. In this case, the promotion rates of majority and minority candidates were compared within four different job strata (levels 7, 8, 9, and 10). While the three estimator calculations provide probability values ranging from .035 to .063, the exact probability value (using the Multiple Events Exact Probability test) is .0347.

Combined Data-Availability Comparison

These data were taken from the Cooper v. University of Texas at Dallas (482 F. Supp. 187, N.D. Tex. 1979, aff'd., 643 F.2d 1039, 5th Cir. 1981) case. This case involved an allegation of discrimination in hiring based on gender. Notice the difference in probability values using the two different methods.

Section 6: Save Data/Load Data



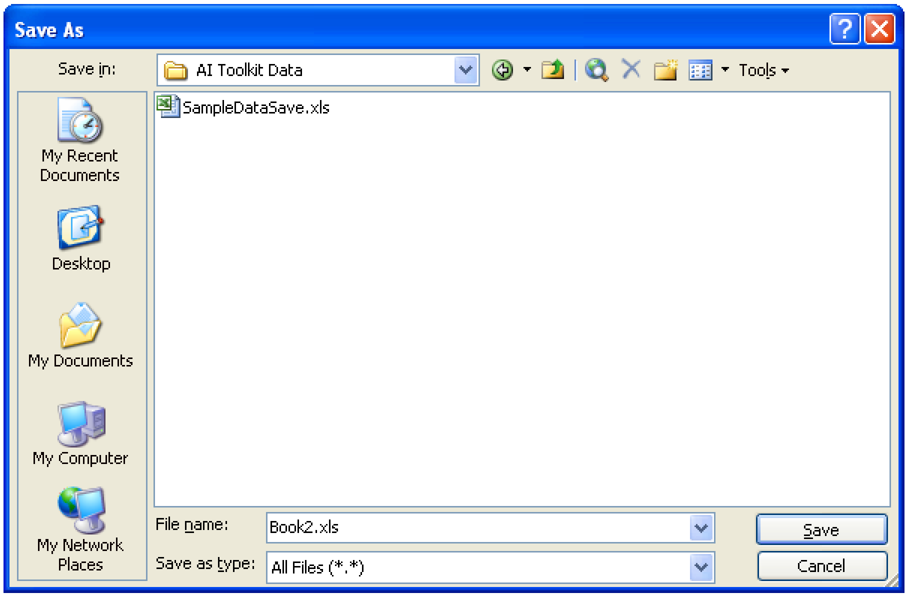
Overview

In the Desktop version of Adverse Impact Toolkit, data you have added can be saved, and restored using the simple **Save Data** and **Load Data** options on the **Main** worksheet page.

Note: This option is not available in the web version of Adverse Impact Toolkit.

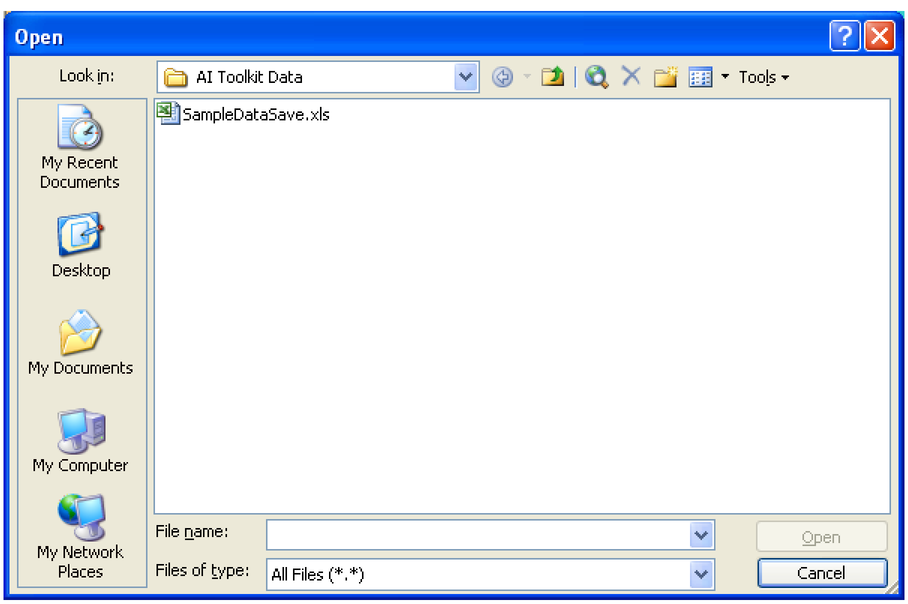
Save Data

Data that you have entered may be saved to a file using the **Save Data** option. When selected, a standard **Save** window is displayed so you can specify a file name and browse to the desired folder location to save the file. The file will be saved as an Excel spreadsheet file (.xls file).



Load Data

Data you have saved previously can be re-loaded to the Toolkit using the **Load Data** option. This displays a standard **Open** window that allows you to browse to the desired folder location where you stored data, then select the xls file to load. Click **Open** when ready.



References & Cases

Agarwal v. McKee and Company (16 E.P.D. P 8301, N.D.Cal. 1977).

Agresti, A. & Gottard, A. (2007). Nonconservative exact small-sample inference for discrete data, *Computational Statistics and Data Analysis.*

Armitage, P., Berry, G., Matthews, J. N. S. (2001). Statistical Methods in Medical Research (4th ed.). Wiley-Blackwell.

Arnold v. Postmaster Gen. (Civil Action Nos. 85-2571, 86-2291, U.S. Dist. Ct. for the District of Columbia, 667 F. Supp. 6; 1987).

Bouman v. Block, 940 F.2d 1211, C.A.9 (Cal., 1991).

Breslow, N.E. & Day, N.E. (1980). Statistical methods in cancer research, Vol. I. "The Analysis of Case-Control Studies." WHO/IARC Scientific Publication, No. 32.

Brown v. Delta Air Lines, Inc., 522 F.Supp. 1218, 1229, n. 14 (S.D.Texas 1980).

Castaneda v. Partida, 430 U.S. 482, 496 (1977).

Chang v. University of Rhode Island, 606 F.Supp. 1161, D.R.I.,1985.

Clady v. County of Los Angeles, 770 F.2d 1421, 1428 (9th Cir., 1985).

Cooper v. University of Texas at Dallas (482 F. Supp. 187, N.D. Tex. 1979, aff'd., 643 F.2d 1039, 5th Cir. 1981).

Covington v. District of Columbia (Nos. 94-7014, 94-7015, 94-7022, 94-7107, U.S. Ct of Appeals, 313 U.S. App. D.C. 16; 57 F.3d 1101, 1995).

Csicseri v. Bowsher, 862 F.Supp. 547, D.D.C.,1994.

Dees v. Orr, Secretary of the Air Force (No. Civil S-82-471 LKK, U.S. Dist. Ct, Eastern District of California, 1983).

Dennis L. Harrison v. Drew Lewis (Civil Action No. 79-1816, U.S. Dist. Ct for the District of Columbia, 559 F. Supp. 943, 1983).

Donnel v. General Motors Corp, 576 F2d 1292 (8th Cir 1978).

Dothard v. Rawlinson, 433 U.S. 321 (1977).

E.E.O.C. v. Federal Reserve Bank of Richmond, 698 F.2d 633, C.A.N.C., 1983.

EEOC v. United Virginia Bank, 615 F.2d 147 (4th Cir. 1980).

Gastwirth, J.L. & Greenhouse, S.W. (1987). Estimating a common relative risk: Application in equal employment. Journal of the American Statistical Association, 82, 38-45.

Guardians Association of the New York City Police Dept. v. Civil Service Commission, 630 F.2d 79, 88 (2d Cir.1980), cert. denied, 452 U.S. 940, 101 S.Ct. 3083, 69 L.Ed.2d 954 (1981).

Hazelwood School District v. United States, 433 U.S. 299 (1977).

Hogan v. Pierce, Secretary, Housing and Urban Development (No. 79-2124, U.S. Dist. Ct District of Columbia, 1983).

Hoops v. Elk Run Coal Co., Inc., 95 F.Supp.2d 612, S.D.W.Va.,2000.

Johnson v. Garrett, III as Secretary of the Navy (Case No. 73-702-Civ-J-12, U.S. Dist. Ct. for the Middle District of Florida, 1991).

Lancaster (1961). Significance tests in discrete distributions. Journal of the American Statistical Association, 56, 223-234.

Louv, W.C. & Littell, R.C. (1986). Combining one-sided binomial tests. Journal of the American Statistical Association, 81, 550-554.

Manko v. U.S. (No. 79-1011-CV-W-9, U.S. Dist. Ct for the Western District of Missouri, 636 F. Supp. 1419, 1986).

McKay v. U.S. (Civil Action No. 75-M-1162, U.S. Dist. Ct. for the District of Colorado, 1985).

Mehrotra, D. & Railkar, R. (2000). Minimum risk weights for comparing treatments in stratified binomial trials. Statistics in Medicine, 19, 811-825.

Moore v. Summers, 113 F.Supp.2d 5, D.D.C.,2000.

Mozee v. American Commercial Marine Service Co., 940 F.2d 1036, C.A.7 (Ind.), 1991.

OFCCP (Office of Federal Contract Compliance Programs) (1993). Federal contract compliance manual. Washington, D.C.: Department of Labor, Employment Standards Administration, Office of Federal Contract Compliance Programs (SUDOC# L 36.8: C 76/993).

Ottaviani v. State University of New York at New Paltz, 679 F.Supp. 288, D.N.Y.,1988.

Paige v. California Highway Patrol [No. CV 94-0083 CBM (Ctx), U.S. District Court, Central District of California, Order Entered August 12, 1999].

Palmer v. Shultz, 815 F.2d 84, C.A.D.C., 1987.

Police Officers for Equal Rights v. City of Columbus, 644 F.Supp. 393, S.D.Ohio,1985.

Simonoff, J.S. (2003). Analyzing Categorical Data (1st ed.). Springer.

Tarone R.E. Homogeneity score tests with nuisance parameters. Communications in Statistics, Series A 1988; 17:1549-1556.

Trout v. Hidalgo (Civ. A. Nos. 73-55, 76-315, 76-1206, 78-1098, U.S. Dist. Ct. of Columbia, 517 F. Supp. 873, 1981).

Uniform guidelines on employee selection procedures. (1978). Federal Register, 43, 38290-38315.

Vuyanich v. Republic National Bank, (N.D. Texas 1980). 505 F. Supp. 224.

Williams v. Owens-Illinois, Inc, 665 F2d 918 (9th Cir), Cert denied, 459 U.S. 971 (1982).