Statistical Analysis

Statistical analysis techniques can be used to describe data, generate hypotheses, or test hypotheses. Techniques that summarize and describe characteristics or a group or make comparisons between groups are known as descriptive statistics.

Three types of statistical analyses are as follows:

1. Univariate: when one variable is analyzed
2. Bivariate: analysis of two variables
3. Multivariate: analysis of three or more variables

The following are examples of evaluation questions answered using univariate, bivariate, and multivariate Analysis:

Univariate Analysis	Bivariate Analysis	Multivariate Analysis
What percent of Minnesotans reported smoking cigarettes in the past 30 days?	Is there a difference in smoking behavior between males and females in Minnesota?	Can the risk of lung cancer be predicted using smoking behavior, gender, occupation, and heredity?
In 2008, how many motor vehicle crashes were alcohol-related?	Are motor vehicle crashes more likely to be alcohol related in rural or metro areas?	Can mortality risk among motorcycle drivers be predicted from helmet use, speed, alcohol use and location?
How many Minnesotans were admitted to treatment programs for methamphetamines as their primary substance of abuse?	How do treatment admission trends for methamphetamine abuse compare to treatment admission trends for abuse of opiates?	How effective have social marketing campaigns, policies restricting the purchase of over-the-counter cold medication, and meth lab busts been in reducing the number of Minnesotans in need of treatment for meth addiction?

Adapted from the Minnesota Department of Health's Community Health Promotion Guide website at www.health.state.mn.us/divs/hpcd/chp/hpkit/index.htm

Data Manipulation—Do's and Don'ts

If you manipulate data long enough, they will tell you anything you want! A sticky ethical dilemma can arise if you start analyzing data with specific findings in mind. Perhaps one of your program objectives is to reduce underage drinking among program participants by 25% at the end of the first program year. Don't ask yourself "how can I manipulate the numbers to get 25% or higher?" It's important to report real outcomes, even if they don't meet expectations.

Similarly, your hypothesis may not always be correct. You may believe that binge drinking rates are dropping among students in your school, but survey results show that rates have remained flat over time. Don't keep administering surveys over and over until you see the results you were expecting.

If you are having a difficult time viewing your data objectively, seek help from an outside expert.

Also keep in mind—even though the data might not say what you expected them to say, they can still be useful! Though you may wish to see declining rates of use due to your program, flat rates may simply mean your program needs more time. Or perhaps some small program improvements are needed. On the flip side, declining rates don't necessarily mean there's not a problem anymore.

The same information can be framed in different ways: In Minnesota, the percent of 12th graders who used alcohol at least once a month in the previous year decreased from 54% in 1989 to 41% in 1992. You could highlight this as a positive trend, focusing on the message that prevention work may have been effective. Or you could focus on the work that still needs to be done, emphasizing that alcohol is illegal for 12th graders until age 21, yet 41% are using it. But why not present all of the facts and let the data speak for themselves: a) in 1992, 41% of 12th graders reported alcohol use in the past month, b) this is a decline from 54% in 1989, c) alcohol is illegal for 12th graders until age 21.

Some situations may be less obvious...and less ethically charged.

Calculating Rates

A rate can be calculated by taking the number of cases or incidents (such as cirrhosis deaths or DWI arrests) and dividing by the population size. Care should be taken when choosing the appropriate population to divide by. If you are calculating a county's prostate cancer rate, you wouldn't want to include the number of women living in the county in your denominator! Likewise, if you are reporting an adult smoking rate you would want to use only the number of residents aged 18 and older.

Rounding

Always do any calculation before *rounding*! For example, if you are calculating an average, add the data for each year and divide by the number of years—then round as need. Don't round before adding or before dividing the data.

If the data you obtain from a source are reported to one decimal place (to the nearest tenth), it is not appropriate to report any figures calculated from these data to two or three decimal places. For example, say you obtain the following DWI arrest rates:

2005	63.7 per 10,000 population
2006	72.8 per 10,000 population
2007	67.1 per 10,000 population

When you calculate the three year average, you get: 67.86666667. It's best to report either 67.9 per 10,000

population or 68 per 10,000 population. Do not report 67.867 per 10,000 population.

Synthetic Data

City-level and county-level estimates can be derived from state-level data. For example, a county-level estimate of adult smoking rates can be obtained by adjusting the state-level percentage (from the Behavioral Risk Factor Surveillance System—BRFSS) by the age and gender distribution of the county. If BRFSS data show binge drinking rates to be highest among 18-24 year old males in Minnesota, then a county with a high "synthetic" rate most likely has a high percentage of 18-24 year old male residents.

For example, say that 32,226 residents in your county were aged 65 and older in 2007. According to BRFSS data, about 7.9% of Minnesotan's aged 65 and older were current smokers in 2007. You could then estimate that 2,546 residents in your county aged 65 and older were current smokers in 2007 (32,226 x 7.9%).

Synthetic estimates of current smokers, acute drinking and chronic drinking are available from the Minnesota Department of Health, Center for Health Statistic's County Health Tables: Morbidity and Utilization—Tables 11 and 12. www.health.state.mn.us/divs/chs/countytables/

Synthetic data can be useful for providing a ballpark estimate of how big a problem might reasonably be in your community without having actual community-level data. They give a sense of the level or risk for a behavior. However, care should be taken when using synthetic data. Such estimates do not take into account regional differences in norms, traditions, cultures, etc. A rural county in northern Minnesota could have a very similar gender and age distribution as a metro county, but very different smoking rates. When using synthetic data, always report them as such.

Survey Scales

Many surveys contain Likert scale response options. Examples of Likert scales are "Excellent, Very Good, Good, Fair, Poor" and "Very Satisfied, Somewhat Satisfied, Somewhat Dissatisfied, and Very Dissatisfied." When reporting the results for such a survey question, there are a couple of approaches you can take. One way is to present the number or percent of persons responding to each category, perhaps in a table shown below.

Another ways is to combine categories. For example, you might report that 6% of adults surveyed said that they "often" or "sometimes" drink and drive. Or, you could say that 72% of adults surveyed said that they "rarely" or "never" speed.

Care should be taken when you combine categories, as important detail may be lost! If you simply report that 94% of adults surveyed said that they "rarely" or "never" drink and drive, your audience will not know whether a majority of respondents rarely drink and drive or never drink and drive. If your audience plans to allocate resources towards drinking and driving prevention, it might be important for them to

Adults reporting how often they…				
	Often	Sometimes	Rarely	Never
Drink and drive	2%	4%	10%	84%
Don't wear their seatbelt	9%	13%	32%	46%
Drive over the speed limit	13%	15%	43%	29%

know that 84% never do and 10% rarely do (rather than the reverse situation—only 10% never drinking and driving).

Combining Data

If you wish to combine data, always work with the number or count—not the percent or rate.

For example, say 41% of 12th grade males reported past month alcohol use and 54% of 12th grade females reported past-month alcohol use. If you want to report a percent for all 12th graders, males and females combined, you must use the actual number of males and number of females who report use. When looking at a data table on www.sumn.org for past-month alcohol use, you can use the toolbar at the top of the page to display by 'Number' instead of 'Percent'. In this case, you might find that 59 12th grade males reported use and 85 12th grade females reported use.

To calculate the total number of 12th grade males who responded to the survey question about past-month use in that county, divide 59 by 41% to get 144. For 12th grade females, divide 85 by 54% to get 157.

Now, to combine the data for males and females—sum those reporting past-month use, then divide by the sum of those responding to the survey question:

(59 + 85) / (144 + 157) = 144 / 301 = 48% of 12th graders

Comparing Data from Two Separate Sources

If you want to combine data from two separate sources, you need to first ask a few questions. Are the survey questions exactly the same? For example, one survey may ask respondents to indicate past-year abuse of prescription drugs, while another survey might ask about past-year abuse of prescription pain relievers. The second question would not give you any information on the people who abused prescription sedatives or prescription stimulants.

Is the population group exactly the same? Some statistics from the National Survey on Drug Use and Health (NSDUH) are for persons age 12 and older. These data should not be compared to Behavioral Risk Factor Surveillance System (BRFSS) data on persons aged 18 and older. Also, comparing data from a college survey on 18-20 year olds would not provide the same information as a community-level survey that would also include 18-20 year olds in the workforce or in the military.

Are the time periods the same? Do the survey questions ask about past-month use or past-year use? Also, a 2009 report might provide 2007 data—make sure you're using the same year for comparisons. If you are comparing alcohol use among college students, from the CORE survey for example, find out which semester the survey was implemented in. Rates tend to be higher during spring semester than during fall semester. Comparing one school that administered the survey in the fall to another school that administered the survey in the spring is not an "apples-to-apples" comparison.

Age Adjusting

Some events occur more frequently among younger age groups, while other events occur more frequently among older age groups. Comparisons cannot be done on raw, or crude, numbers of cancer cases or deaths, for example, because the populations may not be comparable with respect to age. Much like finding the common

denominator when working with fractions, epidemiologists must calculate age-adjusted incidence or death rates so that the populations of different states or regions are similar enough for comparison.

For example, consider two states, Florida and Alaska. Florida has a relatively old population and Alaska has a relatively young population. If you look only at each state's crude cancer death rate (total number of cancer deaths divided by the total population) it appears as if Florida has a cancer death rate almost three times that of Alaska. That is because the population of Florida is older and the risk of most cancers increase with age. The age difference makes the overall cancer death rate in Florida appear higher than in Alaska.

Therefore, epidemiologists use a strategy called age-adjustment so that the rates of different states or regions can be compared among people of similar age. Indeed, when the cancer death rates for Florida and Alaska are age-adjusted, they are almost identical. This means that cancer mortality rates are actually similar in the two states, not separated by the threefold difference that appeared at first glance. Getting to the truth in the numbers is important for understanding risk and for planning programs and allocating resources appropriately. This discussion on age adjusting was retrieved from the American Cancer Society on June 12, 2009 from www.cancer.org/docroot/STT/content/STT_1_Age_Adjusted_Backgrounder.asp

A tutorial on how to calculate age-adjusted rates is available from the National Cancer Institute's Surveillance Epidemiology and End Results (SEER) website: http://seer.cancer.gov/seerstat/tutorials/aarates/definition.html