STATISTICS IN RESEARCH: An Introduction

In tobacco control, we often encounter scientific research that describes the public health impacts of tobacco, the effectiveness of cessation therapies, and similar topics. The research almost always involves statistics, which help us interpret the meaning of the data. To understand these studies, we need to be familiar with statistical terms and concepts.

There are two main subsets of statistics:

- **Descriptive statistics** summarize data and describe general characteristics: the total number of tobacco retailers in each California county, for example, or the percentage of residents of each county who use tobacco.
- **Inferential statistics** help us see beyond the surface to uncover latent information in our data that can provide us with inferences about its underlying meaning. If we wanted to study whether high numbers of tobacco retailers might cause higher levels of tobacco use, we would analyze our data using descriptive statistics.

SOME BASIC CONCEPTS

Data can be **quantitative**, or data that is numerical, or **qualitative**, which is not. Quantitative data can be **continuous** or **discrete** variables. With continuous variables, there is a continuity of values: for example, time is a continuous variable. We can count seconds, and we would have no problem dividing seconds into fractions. In contrast, discrete variables take on only whole numbers: the number of people in group, for example, could only be a whole number.

Very often, public health research involves comparing **populations**, or sample groups, one of which is subject to a certain **variable**, **exposure**, or **intervention** while another is not. For example, a study might follow two groups of smokers: an **experimental group** (or "intervention arm") that enrolled in a tobacco cessation program, and a **control group** (or "control arm") that did not, and then measure the **outcome** of how many **subjects** in each group succeeded in quitting tobacco six months later. By collecting this data, we may gain insights about the effectiveness of the cessation program, our "intervention."

Perhaps we expect that a cessation program is not likely to be effective, and the data will show no significant difference between the two groups: this expectation is called the **null hypothesis**. Typically, the null hypothesis assumes that there will not be any expected differences between groups. Or perhaps we have an **experimental hypothesis (or research hypothesis)**,which is an expectation that we will see **statistically significant** differences between the outcomes in each population. Either way, we need to look at things like **sample size** (larger is usually better), numerical averages (both **mean** and **mode**), **variance** (how much individual scores deviated from the average), **correlation** (how strongly one variable was associated with another), and other factors in order to understand our data.

WHAT ARE THE CHANCES?

Probability theory and statistics are related branches of mathematics that seek to uncover patterns that underlie events. It is a basic principle of probability theory that any event could occur by chance. However, by observing a series of events, we may discover patterns that are unlikely to have occurred by chance, which can lead us to insights about why these events occurred.

When we flip a coin, we can't normally predict the outcome—heads or tails—because it is a random event. Therefore, if we flip a coin 100 times, we would expect the outcome to be heads roughly 50 percent of the time. But what if we observed 60 heads out of 100? Or 70? At some point, we would suspect that something other than random chance was at work. Using inferential statistics, we can calculate the odds that an outcome occurred by chance.

Statistical significance means that there is a very low probability that the data resulted from random chance. There are numerous ways to calculate that probability, and many statistics to describe it. This now begs the question of how low of a probability we are comfortable of getting away with. This threshold is called the **significance level**. Typically, researchers use 0.05 as their significance level, which means that 5% of the time we concluded our research hypothesis when in reality it's not true. In public health science, some of the most commonly used probability tools include:

Odds ratio (OR): a number that represents the odds that an outcome will occur given a particular exposure compared to the odds that it would occur by chance. An OR of less than 1 indicates that exposure made the outcome less likely; an OR greater than 1 means exposure made it more likely.

- If a study reports that young adults living in communities with strong tobacco retail license regulations had lower odds (OR 0.74) of initiating e-cigarettes than youth living in communities with weak TRL ordinances, it means that these young adults were 26% less likely to begin using e-cigarettes.

Confidence Interval (CI): the "margin of error," or how much the outcome might be expected to vary if the study were repeated. The 95% confidence interval is used to estimate the precision of the odds ratio, with a small CI representing a higher degree of OR precision.

- If a study reports an odds ratio as "OR 0.80" and the confidence interval as "95% CI=0.56-0.90" means that if the study were repeated 100 times with different sample populations, the OR would fall between 0.56 and 0.90 at least 95 times.

SOME ADDITIONAL TOOLS & TERMS TO KNOW:

ANOVA is an acronym for "analysis of variance," which is a process for extending the comparison between means derived from two groups to means derived from three or more groups.

Categorical variables are a subset of qualitative data in which individuals are sorted into categories according to some qualitative property. For example, ethnicity within a population could be counted as a categorical variable.

Chi-square test is a process that compares the distribution of data from a random sample to the distribution that could be expected if the data conformed to the experimental hypothesis. It can be thought of to measure whether the data "fits" the hypothesis. A two-way Chi-square test is a variant of the Chi-square test, which compares two categorical variables and can assess whether there is dependence between them. For example, we might want to test whether race has an effect on respondents' opinion on smoking. This is when the Chi-square test comes in handy.

Distribution is a list of frequencies for an event happening. A normal distribution is a bell-shaped curve. An example of this is the distribution of height. There are more people who are averaged height while there are fewer people who are very tall or very short.

Effect size measures the magnitude or the impact of an event or phenomenon. Most commonly used is Cohen's d, which quantifies the effect size of an event or outcome between two populations.

Interrater reliability looks at whether two raters are consistent with agreeing or disagreeing on a subject matter. Cohen's Kappa is one of the many tools used to quantify agreement between two raters, with the possibility of chance being accounted for.

Mean looks at whether two raters are consistent with agreeing or disagreeing on a subject matter. Cohen's Kappa is one of the many tools used to quantify agreement between two raters, with the possibility of chance being accounted for.

Mode is a measure of center that refers to the data value with the highest frequency of recurrence.

McNemar test is very much like the two-way Chi-square test but can be used when the observations are related somehow. For example, there is a cessation program intended to curb future smoking behavior (yes or no) in teenagers. A McNemar test can test whether the program had an effect on the teenagers' smoking behavior by looking at their pre- post responses. A McNemar test can only be used on two categorical variables with only dichotomous responses.

P value is the observed significance level. If our significance level was 0.05, "P<0.05" indicates a statistically significant difference between the study groups. "P>0.05" means there was no statistically significant difference.

Proportion test, also known as the large-sample hypothesis test for population proportion, is another test used to compare proportions of a characteristic from two different populations. For example, we might want to compare the proportion of males who smoke in County X to the proportion of males who smoke in County Y in order to tell if there's any significant difference.

Regression refers to a way of analyzing data to understand how an increase or decrease in one measured quantity may correspond with a change in another quantity. For example, we might use regression to analyze data on whether an increase in the number of cigarettes smoked correlated with increased incidence of lung disease, or whether there was a relationship between low personal income and tobacco use. We can also use regression to predict the response variable based on one or more independent variables.

Standard deviation refers to a statistical measure of how the data diverges from the mean. Data that clusters around the mean has a lower standard deviation; data that deviates further from the mean has a higher standard deviation.

T-test is a process used to compare the mean scores from two samples in order to determine whether significant differences exist between the groups. The test considers the different means from each group and the **standard error**, or how much each individual score varies from the group mean, to derive a **T-value**, which can indicate whether or not the differences between group were statistically significant. We use a t-test instead of a z-test when we do NOT know the population variance.

Variance is the square root of standard deviation; and likewise, standard deviation is the square root of variance.

Z-test refers to any of several tests that use statistical models of standard normal data distribution to determine if observed differences within a population or differences between different populations are large enough to be statistically significant. One particularly important distinction from the t-test is that we DO know the population variance in the z-test.

For more information, visit TCEC's website: https://tobaccoeval.ucdavis.edu

Martin, J. & Yang, A. 2019. *Statistics in research*. Davis, CA: Rover Tobacco Control Library.