Data Analysis Software

Data analysis tool: RStudio

Kitsap Public Health District's Assessment & Epidemiology program uses RStudio for data analysis. RStudio is a free, open-source Integrated Development Environment (IDE) developed by Posit that serves as a user-friendly interface for the R programming language. The R programming language is used by Kitsap Public Health District (KPHD) to create standardized, reproducible summaries of data and perform statistical analysis.

For more information, see the RStudio IDE webpage at https://posit.co/products/open-source/rstudio/?sid=1.

Trend analysis tool: Joinpoint

Kitsap Public Health District's Assessment & Epidemiology program uses Joinpoint for trend analysis. Joinpoint Trend Analysis Software is statistical software for the analysis of trends using joinpoint models where several different lines are connected at the inflection points or "joinpoints"(for example, zero joinpoints is a straight line).

The software takes trend data (data over time) and fits the simplest joinpoint model that the data allow, starting with the minimum number of joinpoints, and tests whether more joinpoints are statistically significant and must be added to the model, up to the maximum number of joinpoints the data allow. This allows the user to test whether an apparent change in trend is statistically significant. The models can analyze crude and age-adjusted rates and proportions, use a Poisson model of variation, and tests of significance use Monte Carlo permutation methods. KPHD typically uses a log-transformed model which calculates the annual percentage change in the outcome. When an annual percentage change is statistically significantly different from zero for a set of data points over time, we say that the trend is increasing or decreasing. When the annual percentage change is not statistically significant, we say the trend is not changing over time.

Some analyses may be conducted using Joinpoint regression methods in R and some are done in the Joinpoint Trend Analysis Software.

For more information, see the Joinpoint Trend Analysis Software webpage at https://surveillance.cancer.gov/joinpoint/.

Reach out with any questions!

epi@kitsappublichealth.org

Data Analysis Definitions

Rates

Crude rates

A crude rate is the number of events (such as deaths) in a specified time period divided by the number of people at risk of these events in that period (typically, a state or county population unless the event only affects a subset of the population). This figure is generally multiplied by a constant such as 1,000 or 100,000 (the "multiplier") to get a number that is easy to read and compare and is reported as "per 1,000" or "per 100,000." Rates calculated in this manner are called crude rates.

$$Crude\ rate = \frac{\#\ events\ in\ the\ population}{total\ population}\ X\ [multiplier]$$

Age-adjusted rates

Crude rates adjust for differences in population size but not differences in population characteristics. To see if a rate is high due age differences in the population, we need to use age-adjusted rates. These rates are computed by taking crude rates for each age group and applying them to a standard population. Since 1999, the standard has been the 2000 U.S. population.

The major use of age-adjusted rates is to allow comparisons between different areas and/or different time periods, especially for metrics highly reliant on age. Users should be aware that an age-adjusted death rate has no absolute meaning; it is an artificial number based on a hypothetical population and is only useful for comparing with other rates calculated in the same manner.

Incidence

The number of newly diagnosed cases of a disease or condition during a specific time period.

Prevalence

The number of cases of a disease or condition, including both newly diagnosed and pre-existing, for people alive on a certain date.

Confidence Interval

An estimated range of values that are likely to include the true unknown population parameter. The confidence interval is a measure of the variability in the data. If the confidence interval is 8.1% - 15.8%, then we are 95% confident that the actual percentage is between these numbers.

Statistically Significant

A mathematical measure of the difference between groups. The difference is said to be statistically significant if the difference between the groups is greater than what is expected to happen by chance alone 95% of the time.

- When the confidence intervals for two groups do not overlap (include any values in common), then the two groups are statistically significantly different. When the confidence intervals overlap just slightly, a statistical test can be performed to check for a statistically significant difference.

Median

The value exactly in the middle of all the values in the sample or in the data. Half of the values are above and half are below.

Percent Change

The amount of change in a statistic over a given time interval. A positive percent change corresponds to an increasing trend, while a negative percent change corresponds to a decreasing trend.

$$\%\ change = \left(\frac{final\ value - initial\ value}{initial\ value}\right) * 100$$

Trend over Time

The change in rate, or number of events, over time, which is usually expressed as an increasing or decreasing trend over a period of years. Although there may be some change over time, the trend is not identified unless it is statistically significant (i.e. greater than what is expected to happen by chance alone 95% of the time). A change from one year to the next is not a trend. There must be at least 3 years (or periods) of data to show a trend.

Correlation

When two things have a mutual relationship or connection between them, they are said to be correlated. A strong correlation, association, or connection between two things does not mean that one of things caused the other.

Example: there is a correlation between eating ice cream and getting sunburned, but neither one influences, causes, or affects the other. They are both associated with a separate factor, sunny weather. The sunny weather is the confounder.

Causation

The relationship between two events or two variables (cause and effect), where a change in the cause results in a change in the effect.

Confounding

A variable that is associated with two other variables, and because of its association, the confounder may mask a true association between the variables or suggest an association where there is none.

If you have any questions about definitions, calculations, or wording, ask an epidemiologist!