2021 Data Analysis & Technical Assistance Manual

Sponsoring Washington State Agencies:
Health Care Authority - Division of Behavioral Health and Recovery
Department of Health
Office of Superintendent of Public Instruction
Liquor and Cannabis Board

Prepared by:
Looking Glass Analytics, Inc.
July 2022

Washington State Healthy Youth Survey 2021

Data Analysis & Technical Assistance Manual

Prepared For:

Health Care Authority
Division of Behavioral Health and Recovery
626 8th Avenue SE
Olympia, WA 98501

Department of Health
Town Center East
111 Israel Road S.E.
Tumwater, WA 98501–7835

Office of Superintendent of Public Instruction
Old Capitol Building
600 S. Washington
P.O. Box 47200
Olympia, WA 98504–7200

Liquor and Cannabis Board
1025 Union Ave SE
P.O. Box 43075
Olympia, WA 98501

Prepared By:

Looking Glass Analytics, Inc.
101 Capitol Way N, Suite 203
Olympia, WA 98501
July 2022

This manual is available online at: http://www.AskHYS.net/Training

Suggested Citation:

Healthy Youth Survey 2021 Data Analysis & Technical Assistance Manual. Washington State Health Care Authority, Department of Health, Office of the Superintendent of Public Instruction, and Liquor and Cannabis Board, July 2022.

The Healthy Youth Survey was administered by the Washington State Health Care Authority Division of Behavioral Health and Recovery, the Department of Health, the Office of the Superintendent of Public Instruction, and the Liquor and Cannabis Board. The Healthy Youth Survey Planning Committee includes members of these state agencies and oversaw the implementation of the 2021 survey.

Washington State funding for the 2021 survey and for this report was provided by the Dedicated Cannabis Account, as specified in Initiative 502. Additional support for HYS trainings and other reports were provided by WA State Department of Health and the U.S. Center for Substance Abuse Prevention, Substance Abuse Block Grant.

Appendix A: County-level Analysis Coding by Year updated 7-24-2023.

Contents

Introduction

Purpose

- Establish standard methods for simple frequency and crosstab analysis of Healthy Youth Survey (HYS) data.
- Support STATA programming of HYS analysis. These concepts can be translated into other software languages by users.
- This manual will focus on STATA, Version 16.

Audience

People who conduct or request analysis of HYS data:

- Department of Health (DOH) epidemiology/research staff
- Health Care Authority (HCA)/Division of Behavioral and Health Recovery (DBHR) or other state agency research
- Local health jurisdiction staff
- Other community partners
- Researchers and graduate students

Uses

This manual was developed to be used by a variety of people:

- Experienced or novice STATA users new to the HYS
- People familiar with the HYS but new to STATA

While this manual provides basic information about analyzing the HYS, it is by no means exhaustive. Nor does it present the only way or the best way to run analyses. As STATA users know, there are multiple ways to program to achieve the same results.

Manual Layout

This manual is accompanied by examples of STATA coding and tables and charts. In the manual STATA coding and output are formatted as:

STATA coding is highlighted in grey STATA results are in STATA output format

This manual includes references to other sections of this document and to outside websites. References to outside websites do not imply endorsement by the state agencies involved in HYS.

The manual is divided up into the following sections:

1. HYS Overview
    - Provides a brief overview of the survey, its history and goals
2. Getting Access to HYS Data
    - Describes our data sharing agreements and terms of use
3. Getting to Know your Data
    - Describes common HYS variables including demographic, 30-day and lifetime substance use, calculated and computed variables, and risk and protective factors
    - Provides computed variable coding
4. Getting to Know STATA
    - A table with commonly used STATA commands
5. HYS Data Analysis in STATA
    - Describes how to set up STATA for different types of data, how to explore your data, transform it and run some simple analyses
    - Includes a "do file" in the appendix with coding
6. HYS Data Analysis – Quick Examples
    - Provides an example of how to run crosstab analyses in STATA using state data, county sample, census or mixed data, and ESD data
    - Includes a "do file" in the appendix with coding
7. Comparing State and Local Data
    - Describes how to combine state and local data and compare local data to the rest of the state sample
8. Comparing Years of Data
    - Describes how to combine years of data and compare one year to another
9. Combining Grade Levels
    - Describes how to combine grades and create "high school" estimates for comparison with other surveys
10. Adding Additional Data
    - Describes how to add additional data to your HYS dataset by merging.
11. Checking Findings and Significance Online
    - Describes the information available on the www.AskHYS.net website and how to use it to verify your analysis results
12. Displaying Results
    - Provides some tools to help you display the results of your STATA analysis
    - Includes a "do file" in the appendix on graphing in STATA
13. Web Resources

Special Considerations for HYS 2021 and the COVID-19 Pandemic

The COVID-19 pandemic led to several important changes in the usual HYS process and structure that may have an impact on the results. Due to the unexpected shift to primarily remote learning, the HYS was not administered in fall 2020 as it was originally intended. Instead, the HYS Planning Committee determined it would be best to delay the survey to fall 2021. This ensured a process could be created for students who needed to take the survey remotely and it allowed the Planning Committee to add a few COVID-19-related questions to support future preparedness and response in schools and communities. In addition, the decision was made to expand e-survey/online survey administration across the state. All of this led to a number of factors that may or may not have had an impact on the results:

1) Delaying the survey by a year means a change in the cohort of students being surveyed. HYS has historically been offered in Fall of even years to students in grades 6, 8, 10 and 12, So, roughly the same cohort of students were ultimately being surveyed every two years as they advanced.

2) The Planning Committee chose to halt plans for a more extensive evaluation of the e-survey mode compared to paper that was scheduled for HYS 2020. Instead, the shift to an e-survey mode without the in-depth comparison makes it more difficult to determine whether the survey mode (paper vs e-survey) has an effect on how students answer questions. Only a very small number of schools elected to do the survey on paper in 2021.

3) Schools were allowed to administer the e-survey remotely in Fall 2021 to accommodate students who may be doing hybrid or fully distanced learning. The vast majority of students took the survey in-person at school, though a small number did take the survey remotely. The potential impact of having students complete the survey remotely is still being assessed.

4) Finally, the pandemic itself has led to massive changes in the lives of Washington youth. Increases and decreases in HYS 2021 data may be more a reflection of the pandemic and its effect on our lives than a change that would have happened if the pandemic had not occurred. This means that trend data from before the pandemic and during/after the pandemic should be interpreted with tremendous caution. For example, a large decrease in one particular risk behavior on school property may be explained by a new school education campaign or program or it may be explained by the fact that students are doing more remote learning.

Due to concerns about the impacts of survey administration changes in 2021 and COVID-19, we recommend using caution when analyzing changes from previous HYS administrations and trends.

Other Issues in Analyzing Healthy Youth Survey Data

The Healthy Youth Survey is a large-scale effort and involves a number of complexities which affect data analysis. These issues are discussed throughout this manual and are also summarized

below. They include:

- Complex sampling designs and survey designs that vary between geographic areas
- Comparisons of state and county data
- Multiple Forms of the questionnaire
- Surveying particular grades
- Response rates and valid survey rates, which are estimated based on available data before final enrollments become available
- Cell size

Sampling Designs

The Healthy Youth Survey is intended to provide information about students in public schools at a variety of geographic levels: state, county, Educational Service District (ESD), district, and school (or in the case of small schools, groups of schools). The design for these different geographic levels varies. For such as schools, school districts, and small counties, a census design is used in which all students in that area are asked to participate. For larger groups, in order to increase efficiency, we use a complex sampling design in which we select random samples of schools and then recruit all students in the grades of interest in participating schools. In the absence of drawing a sample, we assume a census design for the purpose of analysis.

State level

At the state level, in order to efficiently provide information that is representative of students in public schools statewide, we select three simple random samples of public schools in the state containing grades 6, 8, and 10/12, and recruit those schools for the state sample. In 2021, Tribal schools and charter schools were also included the state sampling frame. All of the students in these sampled schools in the surveyed grades are asked to participate. This "clustered" sampling design reduces student to student variability because students in the same school may tend to answer survey questions in similar ways; that is, the data are correlated within schools. We adjust for the clustered design by using a statistical program developed to analyze data from complex sampling designs. Since the sample is drawn by randomly selecting schools within grades, the grade/school combination (schgrd) is the primary sampling unit (PSU). (On non-identified datasets, schgrd is replaced by a sequential variable called "psu" that is converted from schgrd to remove identifying information.)

Using a statistical analysis that incorporates the design used and designating the PSU is necessary in order to obtain correct standard errors, confidence intervals, and significance tests. Using an analysis that adjusts for the clustered sampling design compensates for the reduced variability due to intra-correlation within schools and provides error estimates that should approximate what would have been obtained with a simple random sample. Not accounting for PSUs will generally underestimate the variability in the sample and give you lower standard errors and narrower confidence intervals.

Local levels

To produce local results, schools not selected for the state sample are also invited to participate in the survey. Most local data assume a census design, in which all students in the grades of interest in that area would ideally participate. In order to use these data to generalize beyond the particular students surveyed (e.g., to intervening years or to students who may have not been surveyed) these data can be analyzed and confidence intervals obtained by using a random sample design. In these cases, the PSU is the individual student so you do not need to set a PSU and you can use any statistical program.

In large counties, where there will be a gain in efficiency by drawing a sample instead of a census design, county-level samples are drawn. The criteria for drawing a sample at a particular grade in a county is that there need to be at least 30 schools in the county containing that grade. County samples are drawn by beginning with schools selected for the state sample in that county and adding an additional random sample of schools. When analyzing counties with county samples, the school building is designated as the PSU to compensate for the clustering effect (as in the state sample).

Combined local levels.

In order to combine data from geographic areas that used different sampling designs (such as ESDs which can include both sampled and census counties, or comparisons between a census county and the state sample) more complicated approaches may be necessary. Also, although the individual samples are designed to be self-weighting, combining data using different designs may require weighting the data. Depending on your data and analysis goals, there are multiple ways to weight your data. See the instructions below for setting up your data for analysis and how to designate the PSU depending on your specific data.

NOTE: For more information on sampling design see the Sampling section. For more information on data analysis depending on your sampling design see the General Set Up for Survey Analysis section.

Comparisons of County and State Data.

Schools in the state sample are also in the county, and so there is overlap between county and state data. Thus, when comparing a county to the state, there are two ways to do the comparison:

1. Compare the county to the *entire* state sample, including schools in the state sample that are also in the county.
2. Compare the county to the *rest of* the state sample with all of the schools from the county removed from the state sample.

For most counties, the number of schools that are in the state sample is generally too small to raise a concern and both options will give similar results. But technically, to conduct statistical comparisons, it is necessary to have the county and state respondents independent of each other. Thus we recommend the following when conducting comparisons between the county and the state:

- To determine statistical significant differences, remove the county schools from the state sample dataset prior to comparing.
- To report state point estimates and confidence intervals, use the full state sample (not the results with the county schools removed). This way your results will not contradict the previously published state results.

NOTE: For more information see the Comparing State and Local Data section.

Multiple Forms of the Questionnaire

In order to include a large number of questions on the surveys, we use three versions of the questionnaires: one version for grade 6, and two versions (A and B), distributed randomly for grades 8, 10, and 12. Thus, questions that are only on one version cannot be crossed with questions that are only on another version. Also, some items are removable questions that schools can remove prior to administration (2002-2008 and 2016-2021) or on enhanced versions of the surveys that schools could request upon registration (in 2012 and 2014). These removable or enhanced items have much smaller numbers of responses than other items. These factors also mean that while there are a relatively large number of participants in the HYS, the number available for detailed breakdowns may be much smaller. For some analyses, it may be necessary to combine data from two or more years to obtain adequate numbers.

NOTE: More detail on these issues is provided in the Survey Questionnaire section.

Surveying Particular Grades

The Healthy Youth Survey is conducted in grades 6, 8, 10 and 12. It is highly recommended that analysis be limited to a single grade. However, there are situations in which combining grades may be desirable, for example when comparing to the high school estimates from the Youth Risk Behavior Surveillance System (YRBSS), or if there are very small numbers that cannot be reported.

NOTE: More detail on this is provided in the Stratified Analysis and Subpopulations section.

Starting in 2014, schools from small districts could also survey additional grades – 7th, 9th, and 11th. School districts were considered small if they had fewer than 150 students enrolled in one of the surveyed grades 6, 8, 10, or 12. If schools surveyed these grades they could get individual grade level results and combined grade results for Middle School (grades 6, 7, and 8 combined)

and for High School (grades 9, 10, 11, and 12 combined).

Combined grade results are weighted so that each grade contributes equally, using the "smallschoolstateweight", "smallschoolcountyweight", "smallschooldistrictweight" or "smallschoolschoolweight" variables.

Survey Participation

Participation rates for the Healthy Youth Survey are calculated by the number of valid surveys returned divided by the total enrollment for region (i.e., the school, district, county, ESD, or state). Adequate participation rates are necessary to help ensure that the results are representative of the larger region.

There are a number of factors that may influence participation rates, including non-participating schools, participating schools not surveying all students in the grade, students absences, students opting out of taking the survey, and the loss of surveys during the data cleaning process.

NOTE: More information about the response rates and about analyses conducted to examine possible sources of bias in the data are available in the Participation Rates section.

Cell Size

To report results, you must have at least 5 observations per cell when running cross tabulations of state level data, or at least 10 observations per cell when running sub-state level cross tabulations.

HYS Overview

This section provides a brief overview of the survey, its history and goals.

Survey History

The first "Healthy Youth Survey" to assess student risk/protective factors and health behaviors was administered to Washington students in October 2002 and was scheduled for administration every two years, in the Fall of even-numbered years. This document provides a brief description of the survey's purpose and implementation, to help provide a common understanding for community and school stakeholders.

Nationally, trends in youth behaviors and risk/protective factors have been measured using federally developed school-based surveys such as the Monitoring the Future Survey (MTF), and the Centers for Disease Control and Prevention's Youth Risk Behavior Surveillance Survey (YRBSS) and Youth Tobacco Survey (YTS).

State agencies have organized seventeenth statewide youth surveys between 1988 and 2021. The most recent survey in 2021 was administered a year later than originally planned due to COVID-19. It was taken by about 210,000 students in 877 schools, in 215 school districts, and in all 39 of Washington's counties. The next survey will be administered in October 2023.

The implementation and content of Washington youth surveys have changed over time. For example, prior to HYS, four of the previous seven surveys were given to youth during fall months, and three were given during spring months. The surveys in 1988, 1990, and 1999 had a health-risk focus, whereas surveys in 1992 and 1998 were centered on risk and protective factors. More recent versions of the survey—1995, 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014, 2016, and 2018—are a combination of both. The years of administration were also not systematic (that is, there was no pattern for which years the surveys were given). The lack of consistent survey attributes meant that surveys were not necessarily comparable to each other over time and school personnel being asked to participate in each state survey had to learn what was uniquely expected or included in each survey process.

In the late 1990's, interest in youth surveys and need for data for planning and evaluation of science-based programs to support youth both increased. School administration and staff were receiving requests to participate in the various state surveys, national surveys, research studies, and community-generated or school system-generated surveys.

Simultaneously, beginning in 1997, Washington began to implement required student achievement testing as part of evaluating educational systems. The Washington Assessment of Student Learning (WASL) test was required for administration to students in grades 4, 7, and 10. Implementation of this test disrupted several days of instruction for schools in the spring of each year. The names for these achievement tests have changed over time, but up until 2021, they still took several days and occur in the spring. The 2021, achievement testing occurred during the fall, since the testing was cancelled for 2020 due to COVID-19.

State Superintendent of Public Instruction Terry Bergeson determined in 1998 that state agencies must cooperate to administer only one survey of youth behaviors every two years in the fall. In response, staff from OSPI, DBHR, DOH, the Department of Commerce (CTED) and the Governor's Family Policy Council formed the Joint Survey Planning Committee (JSPC). In 2014, the committee was renamed the Healthy Youth Survey Planning Committee (HYSPC) and included OSPI, DBHR, DOH and the Liquor and Cannabis Board.

Common Goals

The HYSPC first identified issues of interest to each agency and to local constituents. These included:

- Describing school, community, family, and peer-individual risk and protective factors (similar to the "Communities that Care" model developed by the University of Washington Social Development Research Group – including Dr. Hawkins and Dr. Catalano)
- Describing youth health habits, risks, and outcomes
- Gathering state-level data in a consistent way (with predictable timing and using comparable measures over time)
- Supporting local-level data collection and use for planning/assessment and evaluation of programs to serve youth

Agreement about Survey Features

After agreeing on common goals, agencies negotiated specific features of the survey – to be called the "Healthy Youth Survey"—necessary to achieve these goals. Agreed features of the survey are as follows:

1. Only one statewide school-based survey of youth will be administered, supported by all state agencies. State agencies in the planning committee agreed to not conduct independent surveys of schools to gather youth data. This agreement should increase efficiency of surveys that are conducted and reduce the burden on schools for surveys. Agencies understood that this would mean challenges in coming to agreement on content for a unified survey.
2. A simple random sample of schools will be recruited at the state level, and county samples will be provided (as appropriate). Methods used to identify a sample of schools to be included in state surveys had changed over time. These changes can have some impact on results, and also complicate year-to-year comparisons of data. Identification of a simple sampling plan makes the survey easier to manage and analyze. The disadvantage of this method is that few schools in any particular area would be included in the state sample, but the planning committee agreed that local schools would be provided some way to "piggyback" (voluntarily participate) to gather local-level data, and county samples could be drawn for counties that are large enough to do so.
3. The survey will be consistently administered in the Fall of even years (2002, 2004, 2006, etc.). This predictable timeline will avoid conflict with student achievement testing, to allow schools and communities to have data available for spring grant writing/needs assessment activities, and help school administrators to plan ahead for participation.

4. Gathering of data in the fall does make comparison to some national surveys (YRBS, YTS) more difficult, because those surveys are conducted in spring months, when youth are older and more likely to engage in risky behaviors.
5. The survey will mainly be given to 6th, 8th, 10th, and 12th graders. Data collection of these grades on a two-year cycle will enable communities and state agencies to watch "cohorts" of youth over time. In other words, the 6th graders who take the Fall 2014 survey will participate as 8th graders in the Fall 2016 survey, and so on. In comparison to national surveys such as the YRBS and YTS, which are given to 9-12th graders, this method will collect more data from younger youth, which is important for early prevention efforts.
6. The survey will be given to youth using survey booklets with a one-page tear-off answer sheet. In comparison to past school surveys, which were given as scannable booklets, having a separate scannable answer sheet dramatically increased the speed of scanning and delivering results. It also decreased the cost of printing. This layout also makes it easier to provide the survey in Spanish. It is possible that using a separate answer sheet may increase the number of mistakes that youth might make as they "bubble" their answers on a separate page from the questions, and might also increase the time it takes for youth to complete each question. The HYSPC investigated this administration change in a small HYS pilot prior to the first administration in 2002, and used results to identify a maximum number of questions that most students could complete during a class period using a separate answer sheet. They also found that the number of illogical answers (either due to mistakes or to students purposely "drawing patterns" on answer sheets rather than answering questions) was not excessive and could still be managed using logic checks during regular quality control screening.
7. The survey will be given to 8th, 10th, and 12th grade youth as a two-form "interleaved" administration. To manage the length of the survey with the breadth of information desired by agencies and stakeholders, there will be a "Form A" and a "Form B" for the survey. Alternately-seated students will receive "Form A" and "Form B" but it will not be obvious to youth sitting next to each other that they have different versions. All youth will have the same "core" questions in their surveys. Youth who complete "Form A" will go on to answer additional questions about sexual orientation and gender identity, while youth who complete "Form B" will answer additional questions about sexual orientation, gender identity, sexual behaviors, and sexual abuse. All sixth graders will have a single version ("Form C") that includes similar items to A and B, to be negotiated among the agencies. However, it is shorter, and in some cases, includes simplified wording to assure that younger students can successfully complete it. Removable questions on sexual behaviors, sexual orientation, gender identity, and sexual abuse are not included on "Form C".

Twenty years later, many of these common goals and survey methods are still core features of the HYS.

HYS 2021 Survey Questionnaires

The HYS includes 3 different survey questionnaire – Form A, B, and C. Forms A and B are for students in 8th, 10th, and 12th grade. Form C is for students in 6th grade.

Forms A and B

Questions for students in 8th, 10th, and 12th grade students are split onto two forms to increase the number of questions that are asked on HYS, but without making it too long for students to complete during a class period. Forms A and B are randomly distributed, so that about half of the students receive Form A and half receive Form B. In 2021, Form A included 144 questions and Form B included 130 questions plus 6 questions on sexual behavior and violence that could be removed by schools if they chose not to ask them.

In 2021, there were about 50 "core" questions that were asked on both Form A and B. These questions included:

- Student demographics including sexual orientation and gender identity questions
- 30-day and lifetime use of alcohol, tobacco, and other drugs
- Key violence-related questions (weapon-carrying, fighting, gangs, perceived safety)
- School climate questions (activities, bullying, safety, school engagement)
- Mental health and depression
- COVID-19 concerns and prevention

In 2021, when schools registered for the survey, they could indicate if they wanted to ask the 6 removable questions on sexual behavior and violence. This was similar to 2010, 2012, and 2014 when schools could register for paper versions of the survey without those questions, but different from 2002 to 2008 and 2016, when schools could not request that the questions be removed and they had to physically tear off the questions from the back page of the survey forms.

In 2018, the two questions on sexual orientation and gender identity were asked on the portion of the removable survey. In 2021, those questions were moved to the main survey on both Form A and Form B. Schools could request an exemption to have the sexual orientation and gender identity questions removed.

Form C

Questions on Form C are mostly consistent with the longer Form A and Form B questionnaires but includes fewer questions and some questions have been simplified. These differences are because 6th grade youth do not have reading skills to complete a longer survey, because some questions applicable to older youth are not appropriate for younger youth, and because there are more small buildings for 6th graders than for older grades where giving results would be affected by having only half the youth take a particular version. In 2021, there were 99 questions on Form C.

For details about which questions are on a specific form, see the Data Dictionary and Crosswalk available at: https://www.askhys.net/Analyzer

HYS 2021 Implementation Schedule

- **January 2021:** state and county samples identified by the Department of Health and approved by the WA State Institutional Review Board
- **March 2021:** survey content finalized; recruitment letters sent to Washington school administrators
- **March 2021-June 30, 2021:** recruitment of schools to participate
- **June 30, 2021:** last day for schools to sign up for the survey
- **September:** Survey coordinators complete an online training, work with school IT support staff to prepare for the survey, notify parents and students about the survey, and train classroom teachers to administer the survey
- **October 11-29, 2021:** schools administer survey to youth
- **March 2022:** reports of results and fact sheets for schools, districts, counties, ESDs and the state will be posted on www.AskHYS.net.

HYS 2021 Administration

There were two major differences in the 2021 HYS administration from previous surveys.

The 2021 survey was giving primarily as an electronic survey instead of a paper and pencil survey. Grade 8, 10, and 12 schools received an e-survey link the Form A and B survey. Students were randomly given either Form A or Form B. Grade 6 schools received an e-survey link to the Form C survey. Four schools in the state sample (about 1%) requested an exemption to administer the paper and pencil survey.

Schools also had the option to administer the survey in a remote setting instead of in the traditional classroom setting on school property. For the state sample, about 1 percent of surveys were administered remotely.

State and County Sampling

Most public schools in Washington State with grades 6, 8, 10, and 12 are eligible to participate in HYS, including charter schools and tribal schools. Some schools that don't have typical classrooms where students can take the survey, including online schools, parent support/home school programs, and special education, aren't automatically included as eligible for HYS, but can request to take HYS as long as they can ensure student anonymity. Institutions and correctional facilities are not eligible for HYS.

A simple random sample of eligible public schools with at least 15 students per grade is drawn (based on the most recent enrollment figures from OSPI). Within the participating schools, all students in the surveyed grades are asked to participate.

- Three samples are drawn: one for 6^{th} grade schools, one for 8^{th} grade schools, and one 10^{th} and 12^{th} grade schools combined (since the grades are often in the same school).
- Non-sampled schools are also invited to participate in the survey; participation allows these schools to obtain their own school results and to contribute to district-, county-, and ESD-level results.
- County samples are drawn for counties with more than 30 schools in a grade. In 2021,

King, Pierce, and Snohomish (grades 6, 8 ,10 and 12), and Spokane (grades 6 and 8) had county samples drawn. The responses from these sampled schools were used to produce county-level estimates.

- For all other counties, the responses from all schools are used to produce the county-level estimates, whether they are in the state sample or not.

For the 2021 and prior HYS administrations, funds were available to support non-sampled schools to register for the survey at no cost. This funding is not guaranteed for future HYS administrations. The combination of schools in the state sample and schools that participate that are not in the sample are considered the statewide census.

Survey Participation Rates

Calculating response rates for the Healthy Youth Survey is complicated by a number of factors:

Loss of data to non-response and during cleaning.

Reasons for data being unavailable included 1) refusals to participate by some schools, 2) students being absent, parents opting out their students, students opting out, or students being away from their school during survey administration, and 3) cases discarded during cleaning based on an algorithm that includes the amount of missing and inconsistent responses, responses to a question asking about fictitious drug use, and responses to a question asking about honesty of responding.

Levels of aggregation.

Response rates for local data were calculated by dividing the number of valid surveys in the sampled schools by the total enrollment in schools selected for the sample. Although issues affecting data lost to non-participation and data discarded during cleaning may be different, the vast majority of unavailable data was due to non-participation in the survey, and only about 2% of data collected is discarded during cleaning. Thus, these figures (actually the valid survey rates) provide estimates of the response rates.

In 2021, state and sampled county response rates were calculated by dividing the number of participants in the sampled schools by the total enrollment in schools selected for the sample. Valid survey rates were calculated by dividing the number of valid surveys in the sampled schools by the total enrollment in schools selected in the sample.

Non-sampled county, school district, and school building response rates were calculated by dividing the number of participants in all relevant schools by the total enrollment in those schools. Valid survey rates were calculated by dividing the number of valid surveys in those schools by the total enrollment in those schools.

2021 response rates were calculated using fall 2021 OSPI enrollment data.

See Tables below for information about the state response rates and participation from HYS administrations.

HYS Student Response Rates for State Sample by Year

Grade	2002	2004	2006	2008	2010	2012	2014	2016	2018	2021
Grade 6	61%	68%	78%	76%	76%	76%	79%	77%	76%	72%
Grade 8	65%	73%	70%	77%	77%	77%	79%	80%	76%	71%
Grade 10	44%	58%	63%	60%	60%	60%	67%	69%	66%	70%
Grade 12	40%	49%	51%	50%	50%	50%	50%	49%	46%	44%
Total	50%	61%	65%	66%	66%	66%	68%	69%	66%	64%

Number of HYS Participants (with Valid Surveys) by Year, State Sample and Census Schools

Grade	2002		2004		2006	
	Sample	Census	Sample	Census	Sample	Census
Grade 6	7,952	32,588	7,862	46,178	8,825	46,031
Grade 8	7,473	32,788	8,466	45,942	8,912	47,970
Grade 10	5,127	26,847	8,059	36,564	8,514	41,458
Grade 12	4,133	20,299	5,876	26,024	6,280	30,308
Total	24,685	112,522	30,263	154,708	32,531	165,767
Total*	137,207		184,971		198,298	

Grade	2008		2010		2012	
	Sample	Census	Sample	Census	Sample	Census
Grade 6	9,068	48,566	11,549	45,756	8,229	48,690
Grade 8	8,730	50,687	9,723	48,119	10,202	46,994
Grade 10	6,907	46,181	6,889	45,997	8,372	42,779
Grade 12	5,641	35,071	5,908	37,390	6,467	33,521
Total	30,346	180,505	34,069	177,262	33,270	170,984
Total*	210,851		211,331		204,254	

Grade	2014		2016		2018	
	Sample	Sample	Sample	Census	Sample	Census
Grade 6	9,129	50,250	9,722	53,614	9,604	55,910
Grade 8	10,673	48,944	8,662	53,812	8,895	53,224
Grade 10	8,821	45,296	10,835	44,766	8,096	47,665
Grade 12	6,639	33,479	7,590	31,392	5,676	33,459
Total	35,262	177,969	36,809	183,584	32,271	190,258
Total*	213,231		220,393		222,529	

Grade	2021	
	Sample	Census
Grade 6	8,426	43,937
Grade 8	7,691	50,287

Grade	2021	
	Sample	Census
Grade 10	9,378	40,910
Grade 12	5,672	27,962
Total	31,167	163,096
Total*	194,263	

*Total does not include 7th, 9th, and 11th grade respondents that took the survey in 2014 through 2021. Total also does not include respondents who answered the wrong form.

Availability of enrollment figures

The denominators used for calculating response rates and valid survey rates are drawn from OSPI October enrollment figures (available online at the Office of the Superintendent for Public Instruction website). The enrollment figures are reported by schools and compiled by OSPI, and prior to the 2012 administration final results had not been available when the Healthy Youth Survey results were reported in the spring of the following year. In order to provide the "best available" estimates of response rates with the reports, these are calculated using the previous year's enrollment figures. Starting in 2014, fall enrollment figures were available in time to calculate response rates. In previous years, state sample response rates were re-calculated but local response rates were not.

Importance of Participation Rates

Participation or response rates are determined by the number of valid surveys returned divided by the total enrollment (or estimated enrollment before final enrollment figures become available). In general, the following guidance may be used when using county- level Healthy Youth Survey data. If the response rates are:

- 70% or greater: The HYS results are probably representative.
- 40-69%: The HYS results may be representative of students but further examination of other data (by school or district) to identify any important differences between participants and non-participants should be completed before generalizing results to the county.
- Less than 40%: Response rates less than 40% are quite low, and these HYS results should not be interpreted as representative of the county.

Data for grades with less than a 70% response rate should be interpreted cautiously. If important groups of students did not take the survey, there may be limitations even if there is a high response rate.

NOTE: For information on participation rates, Past Participation: http://www.askhys.net/Past

Validity, Reliability and Generalizability

Validity is the degree to which the results are likely to be true, believable, and free of bias to enable generalizing to a larger population. A survey item is valid if it accurately measures the concept it is intended to measure. A number of methods are used to help ensure validity,

including:

- Sampling
- Using items from established youth surveys such as the YRBS and YTS
- Piloting new untested questions with youth
- Data cleaning

Only "valid" surveys are included in the final dataset. The contractor uses a series of quality controls to remove data that were incomplete, obviously inaccurate, or internally inconsistent. On average, about 2% of the returned surveys are culled during the process and are removed from the final dataset. Quality control checks include looking for:

- Inconsistent answers
- Evidence of faking high level of substance use
- Dishonesty
- Wrong grade

Reliability is the extent to which a survey measure, procedure or instrument yields the same result on repeated trials. A survey item is reliable if it consistently produces the same results under the same circumstances. HYS ensures reliability by:

- Using standardized administration procedures (e.g., coordinator training, teacher training, written instructions, teacher stays in room but at desk, single class period to avoid discussion, absent students do not make up).
- Providing a safe and confidential environment
- Informing students about the importance of survey
- Keeping student responses confidential (no collection of student name or other identifying information and students place own answer sheet in envelope)

Confidence Intervals

Confidence intervals are used with the survey data to give an estimate of how accurately you can generalize from samples, such as the state sample, to a larger population, such as students in public schools in Washington, assuming that the data are not biased.

Specifically, the 95% confidence interval gives the range that should contain the true population value 95% of the time.

Bias Analysis

Survey responses are often used to estimate the frequency of behaviors or other characteristics in a population larger than those who actually completed the survey. Thus, while only a portion of students in the state took the Healthy Youth Survey in 2021, we would like to use their responses to characterize all 6th, 8th, 10th, and 12th graders in Washington. This is only possible if those who participated in the Healthy Youth Survey are not different in their behaviors from those who did not participate. If they are different, we say that the survey is biased, and we are then limited in our ability to generalize the results to all students. Bias represents systematic error and is different from the random fluctuation that is measured by confidence intervals.

Comparisons could be done using information from sources such as the census, school achievement test results, or other demographic information.

A bias analysis of the 2021 HYS will be available in July 2022 at: https://www.askhys.net/Reports/BiasAnalysis

Bias analysis of previous HYS administrations have found the HYS may not represent smaller geographic areas and may be subject to bias due to non-participating schools and students. Results may underrepresent students in small schools, alternative schools, schools with lower percent minority enrollment, and secondary schools with higher free/reduced lunch rates and lower on- time graduation rates. They may also underrepresent alternative schools and appeared to be due to the fact that alternative schools were less likely to participate in the Healthy Youth Survey, compared to non-alternative schools.

For past HYS administration, caution was recommended when using questions at the end of the non-removable portion of the questionnaires because younger students were increasingly likely to "drop off" in completing the survey (likely due to slower reading). Results for questions at the end of the survey questionnaires may also underrepresent students getting lower grades, with low socio-economic status, who live in non-English speaking homes, who are Hispanic. For the 2021 HYS, blocks of survey questions were randomized for the E-survey, so students did not answer questions in a specific order.

Getting Access to HYS Data

This section describes HYS data sharing agreements and terms of use.

Data Sharing and Human Research Review Requirements

The ability to share and report data that contains information about geographic levels lower than statewide is limited by protections of confidentiality for participants and by issues of identifiability for schools and school districts. Data sharing agreements provide information about these requirements, as well as other issues important to data users. This information is explained below, and a data sharing agreement is available on request.

Protections of Confidentiality for Participants

Importance of anonymity. Prior to participation, all survey participants are informed "Your answers to these questions are *anonymous*. This means that no one will see your answers or know which answer sheet you completed. There are no codes or information to match a survey to a student." Thus, data sharing procedures are designed to assure anonymity. These procedures are part of the human research review process and are included in the approval by the Washington State Institutional Review Board (WSIRB).

Availability of data with geographic identifiers. Outside of the state agencies participating in the HYS, access to data files containing individual level data (e.g., SAS or STATA files) and geographic identifiers is very limited. Because local health jurisdictions (LHJs) have a long history of ability to handle confidential data and of sharing data with DOH, they have access to the data with a data sharing agreement. Other local organizations wishing information about that geographic area are referred first to the LHJ; DOH acts as backup to the LHJ. Researchers who wish access to the individual-level data with geographic identifiers must submit an Exempt Determination Request to the WSIRB. Although educational institutions such as schools and school districts are important participants in the HYS, educational institutions that might have access to students and information about students drawn from student records do not have easy access to identifiable data because information from the HYS, in combination with additional information available to the educational institutions, might make the students identifiable.

Availability of data without geographic identifiers. Statewide data that does not contain geographic identifiers (i.e., school, school district, ESD, or county identifiers) cannot be used to identify individual students. Thus, a non-identified data (from which all geographic identifiers have been removed) is available to legitimate researchers with a data sharing agreement. Interactive access to aggregate frequencies and crosstab survey results for 2002-2021 are available on www.AskHYS.net. The website includes frequency reports, topic specific fact sheets, and a data query system. HYS data are available at the state, county, and ESD level and with permission from the district superintendents, at the district or school level.

Reporting data while retaining anonymity. LHJs and researchers, prior to receiving HYS data, must

sign a data sharing agreement stating that they will comply with procedures approved by the WSIRB. These include reporting requirements to protect individual identifiability. **These requirements state that for data identified by a geographic level less than statewide, frequencies will only be reported where there are at least 15 valid surveys and crosstabs other than grade level will only be reported where there are at least 10 cases per cell**. At the state level, frequencies in crosstabs can be reported if there are at least 5 cases per cell. They also agree to comply with reporting requirements regarding identifiability of schools, described below.

Identifiability of Schools and School Districts

School and school district level information. The HYS planning committee considers that schools and school districts are the "owners" of their data reports, subject to any state and federal laws pertaining to public access to information. Consistent with this, at the time of registering for participation, schools may "opt out" from receiving a school-level report of results, in which case the report will not be generated. Individuals desiring reports of school or school district results are referred to the school or school district.

Reporting data identifiable by school or school district. If a data user wishes to report data in such a way that the results are identifiable by school or school district, he or she must obtain written permission from the principal or superintendent. Otherwise, data from at least three schools and three school districts must be combined for reporting purposes.

Data Sharing Agreements

Data sharing agreement. Prior to receiving individual-level data, LHJs or researchers must sign a data sharing agreement, which includes the data sharing agreement *per se* and an Attachment A. The agreement must be signed by the individual with authority to sign for the organization. Attachment A must be signed by each of the data users working with the data.

Statutory authority for this data sharing is based on Interlocal Cooperation Act, RCW 39.34, which allows agencies to jointly share their powers and contract with one another, provided the use of the data is for a legally authorized activity and not used in a manner that exceeds the requesting department's jurisdiction. In the data sharing agreement, the receiving agency agrees to (1) not release the data file without the agreement of the agency providing the data, (2) not use the data to identify individual students or report the data in a way that individual students can be identified, and (3) not report the data in ways that identify schools or school districts, unless schools agree in writing and students cannot be identified. It also includes provisions for receiving, storing and destroying the data file. A sample data sharing agreement is available on request.

Receiving the data. Data are sent by a secure means and are available in SAS, STATA, or other formats.

More information about data sharing requirements is available by contacting the HYS Principal Investigator at the Washington State Department of Health, healthy.youth@doh.wa.gov or call

(877) HYS-7111.

More information about the WSIRB is available at http://www.dshs.wa.gov/rda/hrrs/

Getting to Know Your HYS Data

This section describes common variables in the 2021 Healthy Youth Survey dataset. It includes information on:

- Demographic variables
- Current (past 30-day) and lifetime substance use variables
- Calculated and computed variables, including how to code them in STATA
- Risk and protective factors

Most variables consist of a letter such as c, d, f, h, etc. followed by a number. The letter prefixes give you an idea about the variable topic:

C – school climate
D – alcohol, tobacco and other drugs F – family risk and protective factors G – demographics
H – health
L – hope
M – community risk and protective factors
P – peer and individual risk and protective factors
S – school risk and protective factors
V – COVID 19

Computed variables are usually acronyms such as bmi, hopescale, currentasthma, disable, aceflag4, problematicinternet, etc. Computed risk and protective factor scales consist of the word risk followed by a number.

NOTE: For a detailed description of HYS variables since 2002, see the most current version of the HYS Data Dictionary and Crosswalk (XLS) at: http://www.askhys.net/Analyzer

Demographic Variables

coname and conum

Depending on the type of dataset you have, you may or may not have these variables. Each county can be identified with either of the two variables coname and conum.

Coname is a string variable that identifies the county name, e.g., "Adams County." Conum is a unique two-digit numeric code that represents each of the 39 counties in alphabetical order starting with Adams (conum==1) and ending with Yakima (conum==39).

Adams=1, Asotin=2, Benton=3, Chelan=4, Clallam=5, Clark=6, Columbia=7, Cowlitz=8, Douglas=9, Ferry=10, Franklin=11, Garfield=12, Grant=13, Grays Harbor=14, Island=15, Jefferson=16, King=17, Kitsap=18, Kittitas=19, Klickitat=20, Lewis=21, Lincoln=22, Mason=23, Okanogan=24, Pacific=25, Pend Oreille=26, Pierce=27, San Juan=28, Skagit=29, Skamania=30, Snohomish=31, Spokane=32, Stevens=33, Thurston=34, Wahkiakum=35, Walla Walla=36, Whatcom=37, Whitman=38, Yakima=39.

distname, distnum, and codis

Depending on the type of dataset you have, you may or may not have these variables. District level data should never be analyzed or distributed unless you have the written approval from the school district.

Distname is a string variable that identifies the school district name, e.g., "Almira School District." Distnum is a three-digit numeric code for the district. These codes are developed by OSPI (information is available on the OSPI website). The distnum variable is only unique within a county. Codis is a unique five-digit numeric variable for each county – district combination. Codis should be used instead of distnum unless you only have data from a single county.

schname, schnum, schgrd and psu

Again, depending on your dataset you may or may not have these variables. School building data should only be analyzed and distributed with written permission from the school district superintendent.

Schname is a string variable that identifies the school building name. Schnum is a unique four-digit numeric code for the school building. These codes are also developed by OSPI. Most schools have codes between 1500 and 4999. Private schools have numbers between 8000 and 8999. Numbers between 9000 and 9999 are special cases and are not official OSPI codes.

School codes are associated with physical school buildings. Buildings may open, close, move, or change their grade levels over time, making it is important to verify that your school numbers, grades, and names match when comparing data over time.

Schgrd is a six-digit numeric code that combines both the school building code and then the grade level of the respondent. In some 2021 datasets, the schgrd variable is deidentified and replace with the variable "psu". In previous years, the deidentified schgrd variable was called schgnoid.

formtype

The HYS has three main survey forms A, B, and C. All 6th graders take Form C. About half of 8th, 10th and 12th graders take Form A and about half take Form B. In 2021, most schools administered the electronic version of the survey, and a few schools administered a paper and pencil version of the survey. Electronic form types are proceeded by an "E" – AE, BE, and CE. Paper form types are proceeded by a "P" – AP, BP, and CP. During the first week of the survey administration, there was an issue with the form A and B randomization and about 1,400 students received both survey forms. The form type of the responses from students who took both forms A and B is AB.

Some variables cannot be cross-tabulated because they are on different surveys (i.e., one variable is on Form A and the other is on Form B). If you run a crosstab and STATA says there are "no observations" it could mean that you are trying to cross variables on different surveys. Formtype can be useful if you want to investigate which Form your variable is on or if you want

to restrict your analysis to include only respondents from one of the Forms.

Survey Location

For the 2021 HYS, schools could administer the survey at school or remotely. The first question on the 2021 survey for all grades was "Where are you taking this survey?, g28. The response options were "On school property" and "Not on school property". Students who responded that they were not on school property were asked additional questions to determine if they were in an environment where they could answer the HYS safely and honestly.

Age g01, g02

In the HYS dataset there are two different variables for age. Variable g01 is asked on Forms A and B for 8th, 10th and 12th graders, while g02 has less response options and is asked on Form C for 6th graders.

Sex at Birth g05_18

In 2018, the question "Are you female or male" was updated to ask, "What sex/gender were you at birth, even if you are not that gender today?", g05_18.

Gender Identity

A question about gender identity was added in 2018, g26a, g26b, g26c, g26d, g26e, and g26f. The response options include the following and are choose all that apply:

- Male, g26a
- Female, g26b
- Transgender, g26c
- Questioning/not sure of my gender identity, g26d
- Something else fits better, g26e
- I do not know what this question is asking, g26f

A combined response variable, g26g, was also computed for respondents who only selected one responses. In 2018, gender identity was asked on the removable sections of both Forms A and B. In 2021, it was included as part of the main survey on both Forms A and B, but a few schools requested an exemption to not ask the question.

Race/Ethnicity g06, g06a, g06b, g06c, g06d, g06e, g06f, g06g, raceeth

In the HYS dataset, there are three types of race/ethnicity variables. The question asks "How do you describe yourself? Choose all that apply." with responses: American Indian or Alaskan Native, Asian or Asian American, Black or African-American, Hispanic or Latino/Latina, Native Hawaiian or other Pacific Islander, White or Caucasian, and Other.

The question is "choose all that apply" so there is an individual variable for each specific race/ethnicity response that includes students who selected the response alone or in combination (AOIC) with any other race/ethnicities:

- Asian or Asian American, g06a
- American Indian or Alaskan Native, g06b
- Black or African-American, g06c
- Hispanic or Latino/Latina, g06d
- Native Hawaiian or other Pacific Islander, g06e
- White or Caucasian, g06f
- Other, g06g

Choose an individual race variable (g06a-g06g) if you are looking at one particular race and need to capture all of the youth who checked a certain race alone or in combination (AOIC). If a respondent only selected "Asian" and "Black" they would be included in both variables any "Asian" responses in g06a and any "Black" response in g06c.

There is a race/ethnicity variable, **g06**, that includes mutually exclusive responses for each individual race/ethnicity category if a student only one race/ethnicity and a computed response for "More than one race/ethnicity marked" if a student selected two or more responses. For example, if a respondent only selected "Asian" then they are counted as "Asian," but if they selected "Asian" and "Black" they would be counted as "More than one race/ethnicity."

Here's an example of the differences between g06 and g06a-g. In 2021 in the state sample 10th grade, looking at variable g06b, a total of 412 youth checked American Indian or Alaska Native as a response option. In the rolled up g06 variable, there are only 120 American Indian youth listed. That is because 292 of those American Indian youth also checked another race and are included as "More than one race/ethnicity marked" in g06.

There is also a hybrid computed Hispanic/non-Hispanic variable, **raceeth,** that includes mutually exclusive responses for the race/ethnicity categories if a student only one race/ethnicity for Asian or Asian American, American Indian or Alaskan Native, Black or African-American, Native Hawaiian or other Pacific Islander, White or Caucasian, or Other and a response for students who select Hispanic or Latino/Latina alone or in combination with any other race. Use raceeth if you want to present race as Hispanic AOIC, non-Hispanic White, non-Hispanic Black, non- Hispanic American Indian or Alaskan Native, non-Hispanic Asian or Pacific Islander, non-Hispanic Other, or non-Hispanic multiple races.

Asian or Pacific Islander, g21a, g21b, g21c, g21d, g21e, g21f, g21g, g21h, g21i, g21j, g21k, g21l

There is an additional question about Asian and Pacific Islander race groups. The question is "choose all that apply" so there is an individual variable for each specific response that includes students who selected the response alone or in combination (AOIC) with any other Asian and Pacific Islander race groups.

- Not Asian or Pacific Islander, g21a
- Asian Indian, g21b
- Cambodian/Khmer, g21c
- Chinese, g21d

- Filipino, g21e
- Japanese, g21f
- Korean, g21g
- Vietnamese, g21h
- Other Asian, g21i
- Native Hawaiian or other Pacific Islander, g21j

There are also two computed Asian and Pacific Islander race categories:

- Computed - Multiracial Asian, more than one response indicated, g21k
- Computed - Multiracial Asian and Pacific Islander, more than one response indicated, g21l

Language Spoken at Home, g07_06, g08

In the HYS dataset there are two different variables for language spoken at home. Variable g07_06 is asked on Forms A and B for 8th, 10th and 12th graders. Three new response options – Chinese, Korean, and Japanese – were added to g07_06 in 2006. The variable g08 is asked on Form C for 6th graders. It only includes three response options: English, Spanish, or Other. It has not changed over time.

Mother's Education Status, g17 and Socioeconomic Status (SES)

In 2002 and 2004, maternal education was g10 and the question was "What is the highest degree or diploma your mother earned?" In 2006, the question wording and the response options changed and the variable for maternal education became g17. Since 2006, the survey question was, "How far did your mother get in school?"

Some youth do not know their mother's level of education, so this question has a large number of "Don't know" responses and missing values. For example, for maternal education in 2021, only 64% of 8th graders, 73% of 10th graders, and 80% of 12th graders selected an education level for their mother. Because of the large number of missing values, use this question with caution, especially for 8th graders.

From 2002 and 2004 and from 2006 to 2014, the survey included similar questions about paternal education, variables g09 and g18.

Maternal education has been used as a proxy measure for low socio-economic status. See the section on Socioeconomic Status (SES) below.

Socioeconomic status – a measure of an individual or family's relative economic and social ranking - is an important social determinant of health; however, we recognize that youth are not accurately able to report on family income. Maternal education (the level of education that has been completed by the student's mother) is a proxy measure for family SES that has been described in literature. We can use this as a "best guess" for the student's SES. Maternal education can be stratified in a variety of ways; we recommend stratifying as "lower SES" if a mother has no post-high school education and "moderate - higher SES" if a mother has had any

post-high school education.

```
gen lowses=g17
recode lowses 1=1 2=1 3=0 4=0 5=0 6=. 7=.
lab def ses 1"low ses" 0"higher"
lab val lowses ses
```

Free or Reduced Priced Lunch

In 2016, a question about receiving free or reduced priced lunches at school was added, g22. In 2021, many schools provided free lunch to all students due to COVID-19, so this measure may not be useful in determining SES. The question is asked on Form B.

Migrant Status

In 2018, a question was added to identify students from migrant families, g25. The question wording was slightly changed in 2021, g25_20. Migrant status was asked on Forms A and C in 2018 and expanded to all forms in 2021.

Chronic Absenteeism

Chronic absenteeism is when students miss ten percent or more of their school days. Variable g27 asks "During the past 30 days, on how many days have you been absent from school for any reason? Include any day that you missed at least half of the school day, so "3 or more days" absent is considered chronic absenteeism. The absenteeism question is asked on all forms.

Sexual Orientation

The variable for sexual orientation is g20_18. The response options were updated in 2018 and include the following options:

- Heterosexual (straight)
- Gay or lesbian
- Bisexual
- Questioning/not sure
- Something else fits better
- I don't know what this question is asking

In 2014, sexual orientation was asked on the removable section of Form B. In 2016 and 2018, it was asked on the removable sections of both Forms A and B. In 2021, it was included as part of the main survey on both Forms A and B, but a few schools requested an exemption to not ask the question.

Living Situations

Since 2012, HYS has included questions about who students live with and where they live Both of these questions have changed over time.

In 2021, the question about who they live with included the following response options (f34_20):

- Parent(s), step-parent(s), or legal guardian
- Relatives like a grandparent, an aunt, an older brother—but NOT your parents
- Foster care parent(s)
- Adults who are not your parents, relatives, or foster parents
- Friends of yours with no adults present
- On your own
- Other

In 2021, the question about who they live with included the following response options (f35_18):

- In a house or apartment that my family rents or owns
- In a house or apartment that a relative rents or owns
- In a house or apartment with someone who is not a relative
- In a shelter
- In a car or RV, park, or campground
- In a motel/hotel
- On the street
- Moved from place to place
- Other

Use caution when comparing the who they live with question, f35, to 2018. In 2018, there was an error on the Form A questionnaire and the response option "a. In my own house or apartment that my family rents or owns was accidentally excluded – so there are three different variables – f35_18, f35_18original, and f35_18clean. If you are looking at results from the entire question, we recommend using f35_18 clean. If you are looking at a specific response, we recommend different variables for different responses:

- In a house or apartment that my family rents or owns – use f35_18clean
- In a house or apartment that a relative rents or owns – use f35_18clean
- In a house or apartment with someone who is not a relative – use f35_18clean
- In a shelter – use f35
- In a car or RV, park, or campground – use f35
- In a motel/hotel – use f35
- On the street – use f35
- Moved from place to place – use f35
- Other – use f35

Kinship Care

Living in kinship care is defined as living with a relative who is not a parent or step-parent. Available for 2012 to 2021.

```
gen kinship= f34_18
recode kinship 1=0 2=1 3/7=.
lab def kinship 0"with parents" 1"relatives not parents"
lab val kinship kinship
```

Foster Care

Living in foster care is defined as living with foster care parent(s). Available for 2012 to 2021.

```
gen foster= f34_18
recode foster 1/2=0 3=1 4/7=.
lab def foster 0"with parents" 1"in foster care"
lab val foster foster
```

Homeless Situation

In 2008, a question was added to try to identify homeless youth, based on the definition used in the McKinney-Vento act, a complicated legal definition. The question has changed over time but can be used as a surrogate measure for homeless youth. Currently, homelessness includes living in a shelter, a car, a park or campground, or on the street.

```
gen homeless=f35_18
recode homeless 1/3=0 4/5=1 6=0 7=1 8/9=0
lab def homeless 0 "Not homeless" 1"Homeless"
lab val homeless homeless
lab var homeless "Homeless screener"
```

Unstable Housing Situation

Unstable housing could simply be defined as "yes" current living arrangements are the result of losing your home because your family cannot afford housing (f36).

Work for Pay

In 2021, the question about working for pay was rotated back on to the Form B from 2016, g12_21. The question asks "How many hours per week are you currently working for pay, NOT counting chores around your home, yard work, or babysitting?"

Small School Variables

Starting in 2014, school districts with less than 150 students in grades 6, 8, 10, or 12 were allowed to survey additional grade levels – 9th, 11th and 12th grades. Surveying the extra grades allowed the districts and schools in those districts to receive combined grade reports for middle school (grades 6, 7, and 8) and high school (grades 9, 10, 11, and 12). To run small school results, use the following:

Small Results Statewide

```
keep if reportsmallschool==1 svyset[pweight=smallschoolstateweight]
gen middle=1 if grade==6 | grade==7 | grade==8
gen high=1 if grade==9 | grade==10 | grade==11 | grade==12
svy:tab d21use grade, subpop(middle) col se obs per
svy:tab d21use grade, subpop(high) col se obs per
```

Small Results District-Level

```
keep if reportsmallschool==1
keep if codis==X
svyset[pweight=smallschooldistricteweight]
gen middle=1 if grade==6 | grade==7 | grade==8
```

```
gen high=1 if grade==9 | grade==10 | grade==11 | grade==12
svy:tab d21use grade, subpop(middle) col se obs per
svy:tab d21use grade, subpop(high) col se obs per
```

Small Results School Building-Level

```
keep if reportsmallschool==1
keep if schnum==X
svyset[pweight= smallschoolschooleweight
gen middle=1 if grade==6 | grade==7 | grade==8
gen high=1 if grade==9 | grade==10 | grade==11 | grade==12
svy:tab d21use grade, subpop(middle) col se obs per
svy:tab d21use grade, subpop(high) col se obs per
```

Substance Use Variables

Some of the current (past 30-day) and lifetime substance use variables are created from recoded variables or by combinations of variables. The current (past 30-day) use questions ask about the use of a substance in the past 30 days and are available with all of the original responses or in a collapsed version with no days and any days of use. Many of the lifetime substance use variables are recoded from questions that ask about the age of first use.

The following are lists of the current 30-day and lifetime use variables from 2021. Substance use questions have changed and rotated on and off the survey over time. For more information on variables including which survey form they are on and the survey item number, see the HYS Data Dictionary and Crosswalk (XLS) at: http://www.askhys.net/analyzer

Current (past 30-Day) Substance Use Variables in 2021

For each of the substance use questions there are two variables:

One includes all of the responses (e.g., d14 is the current use variable for cigarettes with the response options 0 days, 1-2 days, 3-9 days, 10-29 days, all 30 days).

The other includes collapsed response options of "yes" for use on any days and "no" for use on 0 days (e.g., d14use for cigarettes).

- Cigarettes (2002 to 2021): d14 or collapsed none/any d14use. On all forms.
- Chewing tobacco (2002 to 2021): d15 or collapsed none/any d15use. Only on Forms B and C.
- Cigars (2002 to 2021): d16 or collapsed none/any d16use. Only on Form B. Removable in 2006.
- Electronic cigarettes, e-cigs or vape pens (2012 to 2021): In 2012, d90 or collapsed none/any d90use, only on Form B and did not include "vape pens." In 2014, d90_14 or collapsed none/any d90_14use, only on Form B. In 2016, more response options were added d90_16 or collapsed none/any d90_16use, on Forms B and C. In 2021 it was asked on all forms.
- Hookah (2008, 2012 to 2021): d81 or collapsed none/any d81use. Only on Form B.

- Alcohol (2002 to 2021): d20 or collapsed none/any d20use. On all forms.
- Marijuana (2002 to 2021): from 2002 to 2014, d21 or collapsed none/any d21use. In 2016, more response options were added d21_16 or collapsed none/any d21_16use. On all forms.
- Illegal drug not including alcohol, tobacco or marijuana (2004 to 2021): In 2002, current drug use questions were asked in a different order and are not comparable. From 2004 to 2016, d63 or collapsed none/any d63use. Only on Forms A and B in 2004 to 2008. On all Forms from 2010 to 2014. Only on Forms A and C in 2016 to 2021.
- Illegal drug including marijuana (2004 to 2021): d68 or collapsed none/any d68use. This is a combination of d63 and d21. Only on Forms A and B in 2004 to 2008. On all Forms from 2010 to 2014. Only on Forms A and C in 2016 to 2021.
- Pain killers (2006 to 2021): d75 or collapsed to none/any d75use. On Forms A and B in 2006 and from 2010 to 2021.
- Prescription drugs not prescribed to you (2014 to 2021): d92 or collapsed to none/any d92use. Only on Form A.
- Drugs for non-medical reasons (2018 and 2021), choose all that apply. Only on Form A:
    - Stimulant, like Adderall or Ritalin, d109b
    - Painkiller, like Vicodin, OxyContin, or Percocet, d109c
    - Tranquilizer, like Valium or Xanax, d109d
    - Another kind of prescription drug, d109e
    - Over-the-counter drug, like cough syrup or cold medicine, d109f
    - I took something, but I don't know what it was, d109g (new response option in 2021)
- Flavored tobacco or marijuana products (2021), choose all that apply. Form B:
    - Cigars, little cigars, hookah, or other smoked tobacco, d113b
    - Chewing tobacco, dissolvables, snus or other smokeless tobacco, d113c
    - Joints, bongs, pipes, blunt, or other smoked marijuana products, d113d
    - I do not know., d113e
- E-cig or vaping products (2021), choose all that apply. Forms A and B:
    - Liquid with nicotine in it, d114b
    - Liquid with THC (marijuana) in it, d114c
    - Liquid with nicotine and THC (marijuana) in it, d114d
    - Liquid with neither nicotine nor THC, d114e
    - Don't know, d114f
- E-cig or vaping products that were flavored (2021), choose all that apply. Form B only:
    - Flavored liquid with nicotine in it, d115b
    - Flavored liquid with THC (marijuana) in it, d115c
    - Flavored liquid with nicotine and THC (marijuana) in it, d115d
    - Flavored liquid with neither nicotine nor THC, d115e
    - Don't know, d115f
- Substance in e-cig or vaping products (2018), choose all that apply. Form B only:
    - Liquid with nicotine in it, d102b
    - Liquid with THC (marijuana) in it, d102c

    - Liquid with flavor only (no nicotine or THC) d102d
    - Don't know, d102e
- Marijuana and alcohol use at the same time (2021), d106. Form A only.
- Heated tobacco product (2021), d112. Form B only.

Lifetime Substance Use Variables in 2021

- Cigarette, just a puff (2002 to 2021): d01 - asked as age (p19) but recoded for lifetime. Only on Form A.
- Alcohol, sip (2002 to 2021): d05 - asked as age (p20) but recoded for lifetime on Form A and B. On Form C (p21) – asked as ever yes/no.
- Marijuana (2002 to 2021): d06_14 - asked as age (p17_14) but recoded for lifetime on Form A and B. On Form C (p18_04) – asked as ever yes/no. Smoked marijuana (p17) changed to used marijuana in 2014 (p17_14)
- Methamphetamines (2010 to 2021): d10 - in 2018, response options changed to never (d88_18a), within past year (d88_18b), or over a year ago (d88_18c). From 2010 to 2016 asked as ever yes/no (d88). From 2002 to 2008, asked as age (p46). Only on Form A.
- Heroin (2010 to 2021): d89 - in 2018, response options changed to never (d89_18a), within past year (d89_18b), or over a year ago (d89_18c). From 2010 to 2016, asked ever yes/no. From 2004 to 2008, asked as age (p45). Only on Form A.
- Other illegal drugs (2002 to 2021): d12 – asked as ever yes/no. Only on Form C.
- Electronic cigarette, e-cig or vape pen (2018 to 2021): asked as age (d111) and recoded as lifetime (d110) on Form B.

Age of First Substance Use in 2021

Variables that ask the age of first use for substances can be used to calculate the average age of first:

- Cigarette, just a puff (2002 to 2021): p19 on Form A.
- Alcohol, sip (2002 to 2021): p20 on Forms A and B.
- Marijuana (2002 to 2021): p17_14 on Forms A and B. Smoked marijuana (p17) changed to used marijuana in 2014.
- Electronic cigarette, e-cig or vape pen (2018 to 2021): d111 on Form B.

Prior to running the mean age, you need to recode the respondents who have not used the substances to missing and change the other response options to match the age level they represent. To calculate the age of first sip of alcohol by grade:

```
gen agesip=p20
recode agesip 1=. 2=10 3=11 4=12 5=13 6=14 7=15 8=16 9=17
svy:mean agesip, over(grade)
```

Binge Drinking

Binge drinking (2002 to 2021): d61, having five or more drinks in a row in the past two weeks was asked on all Forms from 2008 to 2021, only on Form A in 2006 and on Forms A and B from 2002 and 2004. A collapsed yes/no variable is computed - d61bool.

Levels of Alcohol Use

The computed levels of alcohol use variable, cdv, is on all Forms A, B and C since 2008. The cdv variable combines current (past 30-day) alcohol drinking and binge drinking to break drinking down into the following levels:

- No drinking
- Experimental drinking – 1-2 days drinking and no binge drinking
- Problem drinking – 3-5 days drinking and/or one binge
- Heavy drinking – 6 or more days drinking and/or two or more binges

Warning: In 2006, the levels of alcohol use variable, cdv, was calculated incorrectly. In 2006, the binge drinking question was only asked on Form A, so the levels of alcohol use should only include respondents who answered Form A. To fix the problem, use the following STATA coding:

```
*Fixing 2006 cdv
replace cdv=. if formtype~="A" & year==2006
```

Sources of Alcohol

HYS asks youth about where they get their alcohol (from 2008 to 2021). Youth were asked to check all sources that applied, so there are multiple variables – d76a, d76b, d76c, d76d, d76e, d76f, d76g, d76h, d76i and d76j. On Form A.

Often we want to recode this variable to look only at the youth who actually got alcohol. The question's first response option is "I did not get alcohol in the past 30 days," so the recommended method for recoding is to create a new variable for each of the sources and replace the "did not get" respondents as missing.

```
gen boughtstore=d76b
replace boughtstore=. if d76a==1
gen friend=d76c
replace friend=. if d76a==1
gen gavemoney=d76d
replace gavemoney=. if d76a==1
gen homewithperm=d76e
replace homewithperm=. if d76a==1
gen homewithout=d76f
replace homewithout=. if d76a==1
gen party=d76g
replace party=. if d76a==1
gen stolestore=d76i
replace stolestore=. if d76a==1
gen other=d76j
replace other=. if d76a==1
```

This variable could also be recoded by restricting it to include only current alcohol users. The method is not recommended because this question does not mention "using" it only mentions "getting alcohol."

Sources of Marijuana

Since 2014, HYS asked youth about where they get their marijuana. Youth were asked to check all sources that applied, so there are multiple variables – d97a, d97b, d97c, d97d, d97e, d97f, d97g, d97h, d97i and d97j. On Form A.

To look only at those who got marijuana, the question's first response option is "I did not get marijuana in the past 30 days is set to missing.

```
gen boughtstore=d97b
replace boughtstore=. if d97a==1
gen friend=d97c
replace friend=. if d97a==1
gen gavemoney=d97d
replace gavemoney=. if d97a==1
gen homewithperm=d97e
replace homewithperm=. if d97a==1
gen homewithout=d97f
replace homewithout=. if d97a==1
gen party=d97g
replace party=. if d97a==1
gen stolestore=d97i
replace stolestore=. if d97a==1
gen other=d97j
replace other=. if d97a==1
```

This variable could also be recoded by restricting it to include only current marijuana users. The method is not recommended because this question does not mention "using" it only mentions "getting" marijuana.

Usual Sources of Tobacco

HYS asks youth about where youth usually get their tobacco, d56 (2002 to 2012, 2016 to 2021). Unlike the questions about the source of alcohol and marijuana, the question about tobacco sources only asks for one response. Often we want to recode this variable to look only at the youth who actually used or got tobacco. The question's first response option is "I did not use tobacco in the past 30 days," so the recommended method for recoding it is to set that response to missing.

```
gen tobsource=d56
recode tobsource 1=. 2=2 3=3 4=4 5=5 6=6 7=7 8=8
lab def tobsource 2"store" 3"vending" 4"gave money" 5"bummed" 6"older person" 7"stole" 8"other"
lab val tobsource tobsource
```

This variable could also be recoded by restricting it to include only current tobacco users. The method is not recommended because there are youth who say "I did not use tobacco in the past 30 days" in the usual source question, but also say that they used a tobacco product in the past 30 days (d14, d15, d16). There are also youth who did not use a tobacco product in the past 30 days, but responded that they usually get their tobacco from one of the options. Unfortunately,

it is difficult to reconcile these differences, as youth may have used tobacco but did not obtain it, or they may have obtained it but not used it in the past 30 days. Asked on Form B.

Usual Sources of Electronic Vapor Products

Starting in 2018, HYS asks about where youth usually get their electronic vapor products in the same way asks about tobacco, d103_20. In 2021, the response option "I bought them in a vape shop" was added. On Form B.

Often, we want to recode this variable to look only at the youth who actually used or got tobacco. The question's first response option is "I did not use vapor products in the past 30 days," so the recommended method for recoding it is to set that response to missing.

```
gen vapesource=d103_20
recode vapesource 1=. 2=2 3=3 4=4 5=5 6=6 7=7 8=8 9=9
lab def vapesource 2"store" 3"vape shop" 4"internet" 5"gave money" 6"bummed" 7"older person" 8"stole" 9"other"
lab val vapesource vapesource
```

This variable could also be recoded by restricting it to include only current vapor product users. The method is not recommended because there are youth who say "I did not use vapor product in the past 30 days" in the usual source question, but also say that they used an electronic cigarette, e-cig, or vape pen in the past 30 days (d90_16) and it is difficult to reconcile these differences. E.g., youth may have used a e-cig or vape pen in the past 30 days, but did not obtain it, or they may have obtained it but not used it in the past 30 days.

Susceptibility to Smoking

The measure of susceptibility to smoking was developed by researchers in California to identify youth who have not made strong commitments to remaining smoke-free. HYS asks youth about susceptibility from 2002 to 2012 and 2016 to 2021. This measure has been found to be predictive of progression to smoking within a longitudinal study of youth behaviors. From 2008 to 2021, d72 and from 2002 to 2012, sus. From 2002 to 2012 on all forms. From 2016 to 2021 on Form B.

You can calculate susceptibility by coding:

```
* Susceptibility to smoking uptake – All respondents
gen sus=.
replace sus=0 if (d29==1 & d30==1)
replace sus=1 if (d29==2 | d29==3 | d29==4 |d30==2 |d30==3 | d30==4)
```

Susceptibility is often only calculated for those youth who are not currently smoking and calculated by coding:

```
* Susceptibility to smoking uptake – Among NON-Smokers
gen nonsmoker=d14use
recode nonsmoker 1=0 0=1
svy:tab sus grade, subpop(nonsmoker) col se
```

Any Tobacco Use

It is possible to combine all types of tobacco for a single "any tobacco use" variable, but its usefulness can be limited by the number of respondents. Moreover, tobacco product questions are not consistently asked on HYS which further limits defining 'any tobacco use' for HYS. For this reason, it is common to describe combinations of specific products (e.g., smoked cigarettes or cigars or used smokeless tobacco) rather than "any tobacco use" as the available combinations could change from year-to-year.

From 2002 to 2021, the cigarette (d14) and smokeless tobacco (d15) questions are core items on Forms A, B and C. The electronic cigarette, cigar, hookah and candy flavored questions are only on Form B. There are a number of ways to calculate an "any tobacco" use variable, depending on the year and the questions you want to include. The calculated "any tobacco use" variable should be restricted to include only respondents who took survey Form B and who answered all of the tobacco product questions. Below is an example of one way to calculate an "any tobacco" for 2021:

```
*Any tobacco use(cigarette, smokeless, cigar, hookah and e-cig)
gen anytob=.
replace anytob=1 if(d14use==1|d15use==1|d16use==1|d81use==1|d90_16use==1)
replace anytob=0 if(d14use==0&d15use==0&d16use==0& d81use ==0&d90_16use==0)
replace anytob=. if(d14use==.|d15use==.|d16use==.| d81use ==.|d90_16use=.)
replace anytob=. if(grade==6 | formtype~="B")
lab def anyuse 1"used any" 0"no use"
lab val anytob anyuse
```

Other Calculated/Computed Variables

There are a number of computed variables in the HYS; some of these were not provided for earlier years of the survey. We are providing the computations so that you can create these variables for datasets where they do not exist and so that you understand where the computed variables come from.

Asthma – recode for "current asthma"

From 2008 to 2021, there were two primary variables used to describe asthma prevalence: "has a doctor or nurse ever told you that you have asthma" (lifetime asthma) h22, and "do you still have asthma", h86. This matches the national Youth Risk Behavioral Survey questions to calculate current asthma. Prior to 2008, HYS used different questions so a comparison of current asthma over time is not available.

For more discussion on this topic, refer to "The Burden of Asthma in Washington State: 2013 Update", available at:
https://www.doh.wa.gov/DataandStatisticalReports/DiseasesandChronicConditions/AsthmaData

```
*Lifetime asthma
gen asdrdiag=h22
recode asdrdiag 1=1 2=0 3=0
```

```
lab val asdrdiag yesno

*current asthma
gen stillasth=h86
recode stillasth 1=0 2=1 3/4=0
lab val stillasth yesno
lab var stillasth "Still have asthma"

gen currentasth=.
replace currentasth=1 if (asdrdiag==1 & stillasth==1)
replace currentasth=0 if (asdrdiag==0 | stillasth==0)
lab val currentasth yesno
lab var currentasth "Current asthma, diagnosed by a dr. and still have"
```

Physical Activity – Recode for Meeting Physical Activity Recommendations

The Centers for Disease Control and Prevention has changed their recommendations for physical activity over time. Currently, they recommend that children and adolescents participate in at least 60 minutes of physical activity daily, and muscle strengthening three days a week. Since 2006, HYS has been asking the physical activity question "In the past 7 days, on how many days were you physically active for a total of at least 60 minutes per day? (Add up all the time you spent in any kind of physical activity that increases your heart rate or makes you breathe hard some of the time.)"

More information about physical activity recommendations is available at

http://www.cdc.gov/nccdphp/dnpa/physical/recommendations/index.htm

```
**physical activity – 60 minutes 7 days a week
tab h63, missing
gen pa_7days=.
recode pa_7days 1=0 2=0 3=0 4=0 5=0 6=0 7=0 8=1
lab def pa_7days 1"daily" 0"< 5 days"
lab val pa_7days pa_60days
```

Nutrition – Recode for Fruits and Vegetables Servings "Five a Day"

The past CDC recommendations for fruit and vegetable consumption for youth is five or more servings per day. The Healthy Youth Survey asks students how often they have eaten several common fruits and vegetables in the past week, and the responses are combined into an estimated daily consumption pattern. Note that the individual responses (for example, fv5, the frequency of carrot consumption alone) are not considered useful and not included in any HYS reports.

It is important to recognize that HYS questions are framed as *times* per day, which is different than *servings*. Also providing serving size information doubles the estimated percent eating "five a day" when number of servings is asked; see Bensley, L., Van Eenwyk, J, and Bruemmer, BA. (2003). Journal of the American Dietetic Association, 103:1530-1532. Thus we can estimate the percent of students who meet past nutrition guidelines ("five a day") using this measure, but it is

likely to be an over-estimate if students eat multiple servings at the time they eat fruits or vegetables.

```
*5 servings of fruits and vegetables daily
gen fiveserve=h07
recode fiveserve 1=0 2=0 3=0 4=1
lab def fiveserve 0"Fewer than 5 a day" 1"5+ fruit-veggies a day"
lab val fiveserve fiveserve
```

You can also look at low fruit or vegetable consumption – fruit less than once a day or vegetables less than once a day

```
* fruits less than once a day
gen numday1=fv1
recode numday1 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday2=fv2
recode numday2 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen fruit=(numday1 +numday2)
recode fruit 0/0.999=1 1/24=0
lab def fruit 1"less than 1" 0"more than 1"
lab val fruit fruit

* vegetable less than once a day
gen numday3=fv3
recode numday3 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday4=fv4
recode numday4 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday5=fv5
recode numday5 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen numday6=fv6
recode numday6 1=0 2=.286 3=.714 4=1 5=2 6=3 7=4
gen veggie=(numday3 +numday4 +numday5 +numday6)
recode veggie 0/0.999=1 1/24=0
lab def veggie 1"less than 1" 0"more than 1"
lab val veggie veggie
```

"Obesity" from Body Mass Index h01_14

Obesity is calculated using BMI based on students' self-reported height and weight. Height is converted to centimeters and weight to kilograms, then BMI is computed using the standard formula:

BMI=(weight in kilograms)/(height in centimeters squared)

The cutpoints for obesity and overweight are based on age and gender-specific growth charts developed by the CDC. Individuals in the top 5 percent for BMI based on age- and gender-specific growth charts are considered obese. Those in the top 15 percent, but not the top 5 percent, are considered overweight. Since 2014, h01_14 was also coded to include the results for underweight in the bottom 5% for BMI, and include the results for "healthy weight", which are respondents above 5% and under 85% for BMI.

Children's Hope Scale

Hope reflects a future orientated mindset and motivational process by which an individual has an expectation toward attaining a desirable goal. Research has linked hope with overall physical, psychological, and social well-being. This section introduces the Children's Hope Scale, an assessment of agency (ability to initiate and sustain action towards goals) and pathways (capacity to find a means to carry out goals). Hope Scale questions were asked in 2018 on Form B and expanded to on all forms in 2021.

```
gen hopescalex=(l14 + l15 + l16+ l17)
recode hopescalex 0/8=1 9/12=2 13/16=3 17/24=4
lab def hopescalex 1"very low hope" 2"slightly hopeful" 3"moderately hopeful" 4"highly hopeful"
lab val hopescalex hopescalex
```

Screen Time

Excessive screen time is defined as watching or playing video games for three or more hours on a school day. The questions for screen time have changed and are not comparable over time. In 2018, the definition of TV watching was expanded to include shows or movies or stream videos (such as YouTube, Netflix, Hulu) on any electronic device (Computer, TV set, tablets or smartphone) and the definition of playing video games was expanded to include tablet or smartphone, social media.

```
gen tvhr=.
replace tvhr=0 if h13_18==1
replace tvhr=.2 h13_18==2
replace tvhr=1 h13_18==3
replace tvhr=2 h13_18==4
replace tvhr=3 h13_18==5
replace tvhr=4 h13_18==6
replace tvhr=5 h13_18==7
gen vidhr=.
replace vidhr=0 h122_18==1
replace vidhr=.2 h122_18==2
replace vidhr=1 h122_18==3 replace vidhr=2 h122_18==4 replace vidhr=3 h122_18==5 replace vidhr=4 h122_18==6 replace vidhr=5 h122_18==7
gen screenttl=tvhr + vidhr
gen scrn3p=.
replace scrn3p=1 if (screenttl>=3 & screenttl<=10)
replace scrn3p=0 if screenttl<3
replace scrn3p=1 if h102==5
replace scrn3p=0 if h102<=4
lab var scrn3p "3+ hours screen time daily"
lab def scrn3p 0"less than 3 hours" 1"3 or more hours"
lab val scrn3p scrn3p
```

Problematic Internet Use, problematicinternet

In 2021, three questions from the Problematic and Risky Internet Use Screening Scale (PRIUSS)-3

were added to HYS on Form B. The PRIUSS-3 includes questions on anxiety when away from the Internet, loss of motivation when on the Internet, and feelings of withdrawal when away from the Internet. The three variable are recoded and added together to create a score 0-12. Scores of 3 or more are considered at risk for problematic internet use.

```
gen h132new=h132
recode h132new 1=0 2=1 3=2 4=3 5=4
gen h133new=h133
recode h133new 1=0 2=1 3=2 4=3 5=4
gen h134new=h134
recode h134new 1=0 2=1 3=2 4=3 5=4
egen problematicinternet_num = rowtotal(h132new h133new h134new)
replace problematicinternet_num =. if h132new==. | h133new==. | h134new==.
gen problematicinternet= problematicinternet_num
recode problematicinternet 1/2=0 3/12=1
lab def problematicinternet 0"no problematic internet use" 1"problematic internet use"
lab var problematicinternet "Problematic and Risky Internet Use Screening Scale"
```

Washington HYS Adverse Childhood Experience Score (WAH-ACEs), aces_count, aceflag4

Adverse Childhood Experience (ACEs) are indicators of severe stressors that occur during a person's first 18 years of life. Research has shown that these adverse experiences can influence physical, mental, social, and behavioral health across the lifespan. The Washington HYS ACEs Score (WAH-ACEs) is computed from 11 HYS questions on Form B. The results are collapsed as binary, (0,1) and to create a WAH-ACEs score of 0-11:

- I feel safe during school (NO!/no).
- During the past 30 days, on how many days did you not go to school because you felt you would be unsafe on your way to and from school?* (Any days)
- Bullying is when one or more students threaten, spread rumors about, hit, shove, or otherwise hurt another student over and over again. It is not bullying when two students of about the same strength or power argue or fight or tease each other in a friendly way. In the last 30 days, how often have you been bullied?* (Any days)
- During the past 12 months, did someone you were dating or going out with ever limit your activities, threaten you, or make you feel unsafe in any other way?** (Yes)
- In the past 12 months, how many times did someone you were dating or going out with physically hurt you on purpose? (Count such things as being hit, slammed into something, or injured with an object or weapon.)** (Any times)
- Have you ever been in a situation where someone made you engage in kissing, sexual touch or intercourse when you did not want to? (Yes)
- Not counting TV, movies, video games, and sporting events, have you seen an adult hit, slap, punch, shove, kick, or otherwise physically hurt another adult more than one time? (Yes)
- Has an adult ever physically hurt you on purpose (like pushed, slapped, hit, kicked or punched you), leaving a mark, bruise or injury? (Yes)
- How often does a parent or adult in your home swear at you, insult you, put you down or

humiliate you? (Sometimes, Often, Very often)
- Are your current living arrangements the result of losing your home because your family cannot afford housing? (Yes)
- How often in the past 12 months did you or your family have to cut meal size or skip meals because there wasn't enough money for food? (Any times)

Some students did not answer all 11 WAH-ACEs questions on the survey. To calculate their individual scores and account for those missing answers, a method called multiple imputation was used. This method also used predictors such as mother's education, sex, and race/ethnicity to estimate students' WAH-ACEs score.

The WAH-ACEs score includes a sexual violence question that is on the removable portion of Form B. Schools that chose not to administer the removable questions will not have results for the WAH-ACEs score. Use the aces_count variable for scores from 0 to 11 and the collapsed aceflag4 for 0, 1, 2, 3, or 4 or more ACEs.

More information is available in the WAH-ACEs Interpretive Guide:
https://www.askhys.net/Docs/HYS_Interpretive-Guide_ACEs_2021_FINAL_1_13_22.pdf

Risk and Protective Factors

Risk factors are characteristics of individuals, families, and communities that make us more vulnerable to ill health. Protective factors are characteristics that "protect" and thus significantly reduce the likelihood of disease, injury, or disability. Health-related risk and protective factors are commonly grouped into three general categories including lifestyle and behavior; environmental exposure, encompassing both the physical and social environments; and biologic and genetic characteristics. Risk and protective factors are often measured as different ends of the same continuum. For example, wearing seatbelts protects against motor vehicle-related injury and death; not using a seatbelt increases risk for these outcomes.

The risk and protective factors in the Healthy Youth Survey focus on lifestyle and behaviors and the social environment. The social environment includes the school, peer, community and home environments, as well as individual assets. The survey includes some factors directly related to health, but most of the risk and protective factors are associated with intermediary behaviors, such as drug and tobacco use, violence, and staying in school. Many of these factors have been compiled into scales following the research of Hawkins and Catalano at the Social Development Research Group (SDRG), University of Washington.

The Hawkins and Catalano theoretical framework of risk and protective factors includes twenty-five factors, the scales for which are part of a survey called Communities That Care (CTC). The presence of multiple risk factors predicts an increased likelihood that an individual will engage in substance use, while the presence of protective factors helps to buffer the effect of risk factors and increase resilience.

For a detailed summary of the history of Risk and Protective Factors Scales used in the HYS see: https://www.askhys.net/Reports/Analytic or look in any of the Frequency Reports available at:

https://www.askhys.net/Reports to find a Reporting Schedule from 2002 to 2021.

Content Changes Over Time

Several Healthy Youth Survey questions have changed over time. A crosswalk and data dictionary of survey questions back to 2002 is available at: http://www.askhys.net/analyzer

Getting to Know STATA

This section includes a table that provides a brief overview of some useful STATA commands.

For more information on the specific commands and the output they generate see Data Analysis sections 4 and 5, type help and the command in STATA, or use the Help drop- down on your STATA tool bar and select the STATA command.

Command	Example	Results
For retrieving and saving data		
use	use "C:\My Documents\2021HYS.dta"	Opens the STATA file
save	save "C:\My Documents\new2021HYSdata.dta"	Saves a modified STATA data file
keep	keep d14 d36 grade g05, or keep if conum==1	Keeps only specific variables, or specified response options. Use caution, keep will permanently delete responses if you save over your old dataset.
drop	drop d14, or drop if conum==2	Drops specific variables, or specified response options. Use caution, drop will permanently delete responses if you save over your old dataset.
For variable exploration		
codebook	codebook c01	Describes the variable c01. Includes the question, the data type (numeric or string), the number of values, the number of missing, the response options and labels.
summarize	summarize c01	The number of observations, the mean, the standard deviation, the minimum value and the max value
summarize, detail	summarize c01, detail	Also includes the percentiles, variance, skewness and kurtosis
histogram	histogram c01	Plots a histogram of the variable responses
Creating and transforming variables		
gen	gen year==2004 gen bully=c01	Creates a new variable, or creates a new variable based on an original variable
recode	recode bully 1=0 2=1 3=1 4=1 5=1	Recodes the variable response options, in this example recodes the response options to be not bullied vs. bullied
replace	gen bully=.	In this example the gen command creates a new variable and the replace commands describe the new variable response options.
	replace bully==0 if c01==1 replace bully==1 if (c01==2 | c01==3 | c01==4 | c01==5)	Replace can also be used to create more complex recodes that combine more than one original variable
For labeling variables		
lab var	lab var bully "bullied, none vs.	Labels the variable with a description of the

Command	Example	Results
	any"	variable
lab def	lab def noneany 0"none" 1"any"	Creates new response option labels that can be applied to a variable
lab val	lab val bully noneany	Applies the response option label
Setup commands for analysis		
svyset	gen fakewt==1	Creates a new variable with a weight of 1
	svyset [pweight=fakewt]	Designates the weighting. In this example the newly created fakewt variable is used, so the weight for all responses is equal to 1. Use for analysis of a census county.
	svyset [pweight=fakewt], psu(schgrd) or svyset [pweight=fakewt], psu(psu)	Sets the weight as 1 and the primary sampling unit as the school building/grade. Use for analysis of the state sample or analysis of a county with a county sample.
Updating STATA		
	update all	Install official updates to STATA and provides new programs or commands.
For computing frequencies		
tab	tab c01 grade	Runs a crosstab of the two variables. Tab does not calculate percentages but just provides the number of observations for each cross
svy:tab	svy:tab c01 grade, col se ci obs	Can be used once the data is set up with the svyset command. Svy:tab runs crosstabs of two variables and provides a percentage by row or column and can include additional information such as the standard error (se), 95% confidence intervals (ci) and the number of observations (obs) if designated
For adding additional datasets		
merge	merge (schgrd) using "C:\My Documents\2021 school demo.dta"	Adds additional data to the respondents. In this example we are adding school building information based on the schgrd, possibly school type or enrollment, or free and reduced lunch rates. Remember if you have a de-identified dataset you will have to use schgnoid or psu variable depending on the HYS year.
	merge 1:1	Newer versions of STATA have additional merge options: For 1 to 1 merge
	merge m:1 merge 1:m merge m:m	Merges many variables one to one Merges one variable one to many Merges many variables one to many
append	append using "C:\My Documents\2018 HYS data.dta"	Adds additional respondents. In this example we are adding an additional year of data from 2018.
A few more useful commands		

Command	Example	Results
if	svy:tab c01 grade if g05_18==1, col se ci obs	Limits the analysis to females. "If" at the end of a command means the command is to use only the data specified. When doing CI, use "if" with caution as it can affect CI. Subpop is preferable.
subpop	tab g05_15, gen(sex) rename sex1 female rename sex2 male svy:tab c01 grade, subpop(female) col se ci obs svy:tab c01 grade, subpop(male) col se ci obs	Use the tab and gen commands to create dummy variables (coded as 0,1). Use the dummy variables with subpop to subset your analysis to a specific group.
&	keep if (conum==1 & grade==6)	And
|	keep if (conum==1 | conum==2)	Or
*		Use in "do files" for notes, * before any statement will not run in STATA
///		Use in "do files" if statements are too long to fit on a page. /// at the end of a statement will make it continue to the next line

HYS Data Analysis in STATA

This section describes how to set up STATA for different types of data, how to explore your data, transform it and run some simple analyses.

For a hands-on experience, a STATA "do file" is provided in the following Appendix:

- Appendix A: Do File ~ HYS State Data Analysis Examples in STATA

The do file follows this section of the manual so that you can run analyses and experience producing similar output. If you are using the state sample data, you should be able to reproduce the outputs in this section. This section is formatted so that STATA commands are highlighted in grey and STATA outputs are highlighted in black boxes

This section covers the following topics:

- Opening your dataset
- Do files
- General setup for survey analysis – state, county, ESD, district and building
- Analysis by Grade
- Frequencies and summaries of statistics
- Creating new variables
- Labeling new variables
- Dichotomizing variables
- Two-way tables and crosstabs
- More options for using "svy"
- Additional tips for formatting
- Stratified analysis and subpopulations

For a table of commonly used STATA commands see the previous section Getting to Know STATA. For short examples of STATA coding see the Data Analysis – Quick Examples.

Results presented in this section are from the 2021 HYS data.

Opening your Dataset and "Do Files"

Open your file by typing "use" and then the file pathway in quotes (see syntax below). Or use the STATA drop down menus by selecting File – Open - then find the dataset you want to open and double click on it:

```
clear
set mem 200m
*use "hys2021 state dataset.dta"
```

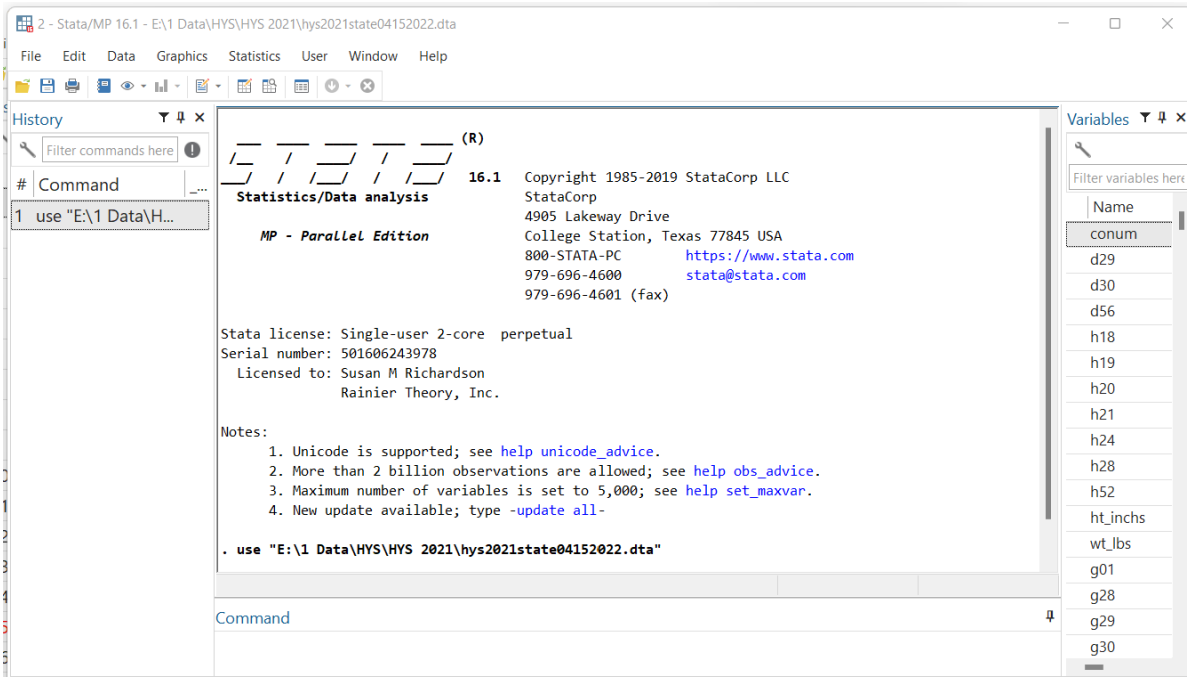
*insert the name of the file path to your state sample data

To open a "do file" click the "New do-file Editor" icon:  on the tool bar, select Do File Editor from the Window drop down menu, or hit Ctrl 9.

Once you have a blank do file open, you can begin writing your commands or open an existing do file by selecting Open from the do file - File drop down menu. “Do files” are handy because you can keep a record of your analysis. They also make it easy to change commands and rerun analysis.

To run individual lines or sections of commands in your “do file,” highlight them and hit the icon  that looks like a page with text with an arrow. To run the complete do file hit the icon that looks like a blank page with an arrow  .

Or you can right click, select all and copy commands you have typed into the History box in STATA (usually on the left) and paste them into a do file.



General Setup for Survey Analysis

Prior to survey analysis you must provide STATA with setup commands to account for weighting, primary sampling units, and strata.

The setup options you use will depend on the type of data you are using and which type of analysis you are conducting. Below are some examples of types of analysis that would influence setup options:

- State sample analysis
- County sample analysis
- County census analysis
- County “mixed sampling” analysis
- ESD analysis
- District analysis

- Building analysis

State Sample Analysis

The state sample was drawn by simple random sample, so there is no weighting or strata required. For survey analysis STATA requires a weight, so you will need to create a fake weight (fakewt) that is equal to 1. The state sample was drawn at the school building level, so the primary sampling unit is the school building (schgrd) if your dataset has identifiers, or schgnoid (2002-2014) or psu (2016 and newer) if the dataset you have has the school buildings de-identified.

Setup command example:

```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
keep if staterec==1
```

County Analysis

You need to be cautious not to report data on counties that should not be reported because either the number of schools/districts or the number of students (less than 15 valid) did not meet the minimum threshold for reporting. It is also recommended to not report counties with survey participation rates below 40%. You can easily exclude specific counties or grades using the "drop" command.

For 2021, the following counties and grades must be dropped and cannot be reported:

```
drop if conum==1 & grade==6
drop if conum==10 & (grade==10 | grade==12)
drop if conum==11
drop if conum==12 & grade==12
drop if conum==26 & grade==10
drop if conum==35
drop if conum==36 & grade==6
drop if conum==38 & grade==6
```

For 2021, the following counties and grades should for participation rates below 40%:

```
drop if conum==2 & (grade==10 | grade==12)
drop if conum==4 & grade==12
drop if conum==8 & grade==12
drop if conum==19 & grade==12
drop if conum==21 & grade==12
drop if conum==23 & (grade==6 | grade==12)
drop if conum==24 & grade==12
drop if conum==27 & grade==12
drop if conum==32 & grade==12
drop if conum==33 & (grade==10 | grade==12)
drop if conum==34 & grade==12
```

To find past year participation, see Appendix A or go to: https://www.askhys.net/Past.

NOTE: The county participation rates vary by year. For more information on who received reports in previous years and who results can be reported for, go to the AskHYS Past Participation webpage: http://www.askhys.net/past

County sample analysis

Random county samples were drawn for counties with more than 30 schools in a grade. The following table describes the county samples from 2002 to 2021.

Year	Clark	King	Kitsap	Pierce	Snohomish	Spokane	Thurston
2002	Grades 6 & 8	All grades	Grade 6	All grades	All grades	Grades 6 & 8	Grade 6
2004	-	All grades	Grade 6	All grades	All grades	Grade 6	-
2006	-	All grades	Grade 6	All grades	All grades	Grades 6 & 8	-
2008	Grades 6 & 8	All grades	-	All grades	All grades	Grade 6	Grade 6
2010	Grades 6 & 8	All grades	-	All grades	All grades	Grades 6 & 8	Grade 6
2012	-	All grades	-	All grades	All grades	Grade 6	Grade 6
2014	Grade 6	All grades	-	All grades	All grades	Grades 6 & 8	Grade 6
2016	Grades 6 & 8	All grades	-	All grades	All grades	Grade 6	-
2018	-	All grades	-	All grades	All grades	Grade 6 & 8	-
2021	-	All grades	-	All grades	All grades	Grade 6 & 8	-

In 2021, county samples were drawn for all grades in King, Pierce, and Snohomish counties. To analyze data from one of these counties, use a similar setup as the state sample.

Setup command example:

```
keep if conum==17
*i.e., conum==17 is King County
keep if corec==1
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd) (or use psu(psu) if you have deidentified data
```

County census analysis

For other counties, all schools in the county are included (a census), so the primary sampling unit is the individual student. You do not need to set a psu.

Setup command example:

```
gen  fakewt=1
keep if conum==2
*i.e., conum==2 is Asotin County
keep if corec==1
svyset [pweight=fakewt]
```

County with "mixed sampling" analysis

In 2021, only one county had a mix of sampling and census, Spokane County. County samples were only drawn for Spokane 6th and 8th grades, but 10th and 12th grades were a census. County sampling changes from year to year. For previous years see the table of Sampled Counties by Year above.

This scenario deserves special attention depending on the grades being analyzed. If you are just

analyzing the 6th grade, then use the setup for county sample analysis noted above. If you are trying to look at all grades in the county, you need to create a new variable for your primary sampling unit. The new variable needs to simultaneously take into account 1) the primary sampling unit for grade six as the school building and 2) the primary sampling unit for the other grades as the individual student.

Setup command for Spokane example:

```
keep if conum==32
keep if corec==1
gen fakewt=1
gen id = _n
gen psu=id +10000
replace psu=schgrd if (grade==6 | grade==8)
svyset [pweight=fakewt], psu(psu)
```

All or multiple county analysis

The following commands can be used if you are running analysis on all counties, some sampled and some census. You need to have a complete state census dataset to run all counties.

You will also need to create a new primary sampling unit variable that takes into account the different sampling schemes, school building for counties and grades with samples and individual students for census counties.

For 2021, the following code is needed to create a psu to account for county sampling and set up data for analyzing data from multiple counties:

```
keep if corec==1
gen fakewt=1
gen id=_n
gen psu=id +10000
replace psu=schgrd if conum==17
replace psu=schgrd if conum==27
replace psu=schgrd if conum==31
replace psu=schgrd if conum==32 & (grade==6 | grade==8)
svyset [pweight=fakewt], psu(psu)
```

The command "gen id=_n" creates a unique identifier for each respondent. When we create our new "psu" variable we add 10,000 to the "id" variable to make sure the new "psu" variable is also unique. Then we replace the individual "id" with the school identifier "schgrd" (or schgnoid) in the counties that were sampled.

Also, drop any county/grades that should not be reported due to participation. For 2021, the following counties and grades should be dropped (because they cannot be reported and less than 40% response):

```
drop if conum==1 & grade==6
drop if conum==10 & (grade==10 | grade==12)
drop if conum==11
drop if conum==12 & grade==12
drop if conum==26 & grade==10
```

```
drop if conum==35
drop if conum==36 & grade==6
drop if conum==38 & grade==6
drop if conum==2 & (grade==10 | grade==12)
drop if conum==4 & grade==12
drop if conum==8 & grade==12
drop if conum==19 & grade==12
drop if conum==21 & grade==12
drop if conum==23 & (grade==6 | grade==12)
drop if conum==24 & grade==12
drop if conum==27 & grade==12
drop if conum==32 & grade==12
drop if conum==33 & (grade==10 | grade==12)
drop if conum==34 & grade==12
```

To find out which counties and grades that need to be dropped for previous years, please see Appendix A: County-level Analysis Coding by Year.

ESD Analysis

ESDs are made up of counties or sections of counties. Some ESDs are made up of counties with samples and some with census. To account for the different sampling schemes, a weight needs to be used that takes the enrollment of schools in the sampled counties. The different sampling schemes also affect the primary sampling units, so a new primary sampling unit variable needs to be created. Also, because county is another layer of sampling, it needs to be accounted for by being designated as strata.

As with counties, you should make sure that the ESD had a 40% response rate and enough school participation to receive results. For 2021, the following ESDs and grade should be dropped for low response:

```
drop if esdnum==101 & grade==12
drop if esdnum==113 & grade==12
drop if esdnum==123 & grade==12
```

Some datasets may already have a "esdpsu" variable. Please do not use this variable, it has an error where the stratum has only a single sampling unit for census counties. Please use the codes below to compute new "esdpsu" variables.

ESD without sampled counties analysis

In 2018, ESDs 105, 112, 113, 114, 123 and 171 did not have any sampled counties, so no special weight or psu needs to be applied, but county does need to be included as strata.

Setup command example for ESD with no sampled counties:

```
keep if esdum==105
keep if esdrec==1
gen fakewt=1
svyset [pweight=fakewt], strata(conum)
```

ESD with sampled counties analysis

In 2021, ESD 101, 121 and 189 had some counties with samples. To analyze data from one of these ESDs, we need to apply weighting and a psu that takes into account the different county sampling.

Setup command examples for ESDs with some sampled counties:

```
keep if esdum==101
keep if esdrec==1
gen id=_n
gen esdpsu=id +10000
replace esdpsu=schgrd if conum==32 & (grade==6 | grade==8)
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

```
keep if esdum==121
keep if esdrec==1
gen id=_n
gen esdpsu=id +10000
replace esdpsu=schgrd if conum==17
replace esdpsu=schgrd if conum==27
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

```
keep if esdum==189
keep if esdrec==1
gen id=_n
gen esdpsu=id +10000
replace esdpsu=schgrd if conum==31
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

All or multiple ESD analysis

Similar to the counties, the following code is needed to create a psu to account for county sampling and set up data for analyzing data from multiple ESDs:

```
keep if esdrec==1
gen id = _n
gen esdpsu=id + 10000
replace esdpsu=schgrd if conum==17
replace esdpsu=schgrd if conum==27
replace esdpsu=schgrd if conum==31
replace esdpsu=schgrd if conum==32 & (grade==6 | grade==8)
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

Drop ESD grades with less than 40% participation:

```
drop if esdnum==101 & grade==12
drop if esdnum==113 & grade==12
drop if esdnum==123 & grade==12
```

District and Building Analysis

For district analysis, all school buildings are to be included because all buildings were eligible to participate, so the primary sampling unit is the student. The variable distnum is not a unique number, i.e., more than one district have the distnum 100. District numbers are only unique within counties, so for district analysis always use the codis variable (a number that includes the county number and the district number).

Setup command example for district:

```
keep if codis==15204
*i.e., codis=15204 is Coupeville School District in Island County
keep if distrec==1
gen fakewt=1
svyset [pweight=fakewt]
```

For building analysis all students were eligible, so students are the primary sampling unit.

Setup command example for building:

```
keep if schnum==4460
*i.e., schnum=4460 is Beaver Lake Middle School in Issaquah
gen fakewt=1
svyset [pweight=fakewt]
```

Special Regions Analysis

Accountable Communities of Health (ACH), Behavioral Health Organizations (BHO), and Regional Service Area (RSA)

The 2021 HYS datasets include variables to help run analysis by special regions for ACH's, BHO's, and RSA's.

Setup command example for Cascade Pacific Alliance Region ACH:

```
tab achid
keep if achid==2
*i.e., achid=2 is Cascade Pacific Alliance Region
keep if distrec==1
gen fakewt=1
svyset [pweight=fakewt]
```

Setup command example for Greater Columbia BHO:

```
tab bhoid
keep if bhoid ==1
*i.e., achid=1 is Greater Columbia BHO
keep if distrec==1
gen fakewt=1
svyset [pweight=fakewt]
```

Setup command example for Thurston-Mason RSA:

```
tab rsaid
keep if rsaid==3
*i.e., achid=3 is Thurston-Mason RSA
keep if distrec==1
```

```
gen fakewt=1
svyset [pweight=fakewt]
```

Analysis by Grade

The variable for a student's grade level is "grade."

We recommend that all analyses be done stratified by grade, because of the sampling procedure by grade and since responses are often variable according to the student grade level. We also recommend that you use the "svy" option in STATA. "Svy" is a prefix used with STATA commands when you are analyzing survey data. "Svy" takes your weights, psu, strata, etc. into account when you are running estimation commands.

NOTE: There may be some exceptions to this, see the Combining Grade Levels section.

You may choose to look at grade differently depending on the types of analysis you are doing and the variables you are looking at. Some variables such as the ones that measure substance abuse, vary greatly by grade level. Others such as the prevalence of asthma are more stable across grade levels.

To simply look at the results for one variable by grade, use "svy:tab":

```
svy:tab d20use grade, col obs per
```

Drank alcohol - Any use in the past 30 days	Grade				
	6	8	10	12	Total
No	97.82	96.4	91.61	80.02	92.39
	7420	6749	7702	4065	2.6e+04
Yes	2.175	3.599	8.386	19.98	7.612
	165	252	705	1015	2137
Total	100	100	100	100	100
	7585	7001	8407	5080	2.8e+04

```
  Key:  column percentage
        number of observations

  Pearson:
    Uncorrected   chi2(3)         = 1591.1928
    Design-based  F(2.05, 465.06) =  116.6508     P = 0.0000
```

Interpretation: 2% of 6th graders, 8% of 8th graders, 18% of 10th graders, and 28% of 12th graders statewide drank alcohol in the past 30 days in 2018.

Frequencies and Summaries of Statistics

Even before you use your setup commands you can run basic frequencies using the "tab" command.

Example in STATA using variable d14, 30-day current cigarette use:

tab d20

During the past 30 days, on how many days did you: Drink a glass, can or bottle	Freq.	Percent	Cum.
a. None	25,936	92.39	92.39
b. 1-2 days	1,429	5.09	97.48
c. 3-5 days	430	1.53	99.01
d. 6-9 days	133	0.47	99.48
e. 10 or more days	145	0.52	100.00
Total	28,073	100.00	

For initial variable exploration, you can use the summarize command to find out the number of observations, mean, standard deviation, min and max type:

summarize d14

Variable	Obs	Mean	Std. Dev.	Min	Max
d20	28,073	1.116411	.4731901	1	5

For more information, including the percentile breakdowns, variance, skewness and kurtosis:

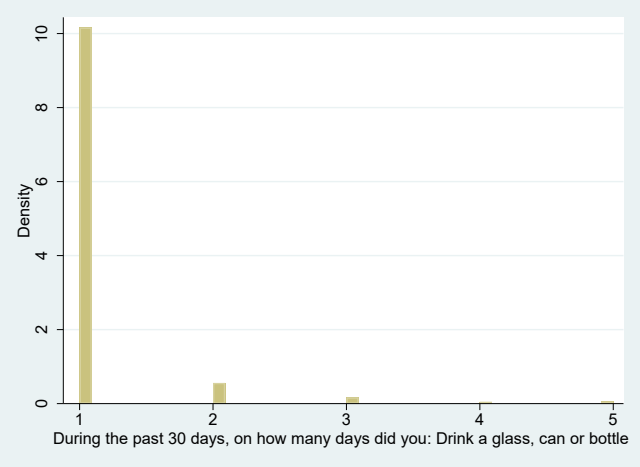
summarize d14, detail

During the past 30 days, on how many days did you: Drink a glass, can or bottle

Percentiles	Smallest		
1%	1	1	
5%	1	1	
10%	1	1	Obs 28,073
25%	1	1	Sum of Wgt. 28,073
50%	1		Mean 1.116411
	Largest		Std. Dev. .4731901
75%	1	5	
90%	1	5	Variance .2239089
95%	2	5	Skewness 5.211655
99%	3	5	Kurtosis 34.43935

Using histograms can also be helpful in getting a quick view of the distribution:

histogram d14



You can also explore your variables by demographics such as grade to find out the number of

observations for each category. Example, current alcohol drinking on any days by grade:
tab d20use grade

Drank alcohol - Any use in the past 30 days	Grade 6	8	10	12	Total
No	7,420	6,749	7,702	4,065	25,936
Yes	165	252	705	1,015	2,137
Total	7,585	7,001	8,407	5,080	28,073

Notice that the proportion of 12th graders who drank alcohol on any days in the past 30 days is much higher than 6th graders.

You can also get a visual look with a histogram by grade:
histogram d20 by(grade)

Creating New Variables

There are many ways to create new variables in STATA - below are a few commands you can use.

Generating

The command for creating a new variable is "generate" or "gen" for short. Below are a few examples of how you can use the "gen" command:

- gen alc30=d20 ~ creates a new variable that is the same as the original variable
- gen alcmarij30 = d20use + d21_16use ~ creates a variable that adds the responses from one variable to another for each respondent
- gen new=. ~ creates a variable with all missing values
- tab grade, gen(gradecat) ~ creates a new dummy variable for each of the original variable response options – with "gradecat" as the prefix followed by the numbers 1,2,3,

etc. depending on the number of response options. In this case "gradecat1", "gradecat2," etc

NOTE: For more information on generating variables, type the command "help gen" in STATA

Recoding

Often during analysis, you want to collapse or drop response options. The simplest way to do this is to create a new variable using the "gen" command and reorder the response options using the "recode" command. It is always a good idea to create a new variable before recoding because you may want to go back and use the original response options sometime during your analysis or recode the variable in a different way.

Before recoding, look at the numerical values assigned to each response option using the "codebook" command:

```
codebook d20
```

```
d20                During the past 30 days, on how many days did you: Drink a glass, can or bottle
-------------------------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  HYSD20FMT

                 range:  [1,5]                        units:  1
         unique values:  5                          missing .:  3,094/31,167

            tabulation:  Freq.   Numeric  Label
                        25,936         1  a. None
                         1,429         2  b. 1-2 days
                           430         3  c. 3-5 days
                           133         4  d. 6-9 days
                           145         5  e. 10 or more days
                         3,094         .
```

Now you know that the variable has 6 response options. If you wanted to recode the 30- day smoking response options into none or any, you need to change the "none" response to 0 and all of the other responses to 1 "any." After recoding your new variable, run a "tab" to make sure your new response options are the way you want them.

```
gen alc30 = d20
recode alc30 1=0 2=1 3=1 4=1 5=1
tab alc30 grade
```

	Grade				
alc30	6	8	10	12	Total
0	7,420	6,749	7,702	4,065	25,936
1	165	252	705	1,015	2,137
Total	7,585	7,001	8,407	5,080	28,073

You can also recode the above variable like this:

```
recode alc30 1=0 2/5=1
```

After recoding it is always a good idea to check your new results to make sure they make sense when compared to your pre-collapsed variable results. In this case, you can check your recode by using the pre-collapsed variable d20use.

```
tab d20use alc30
```

NOTE: For more information on "recode," type the command "help recode" in STATA

Replacing

To combine more than one variable and do more complex recoding, you can use the "replace" command. For example, to calculate if someone has either seen a doctor or a dentist in the past 24 months, you need to combine two different variables, h24 visiting a doctor and h25 visiting a dentist.

Before starting to replace, it's always a good idea to run the codebook command on any variables that you will be using to make sure you know which numeric value is given to each response option.

```
codebook h24 h25
```

```
h24                 When was the last time you saw a doctor or health care provider for a check-up o
-------------------------------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  HYSH24FMT

                 range:  [1,5]                        units:  1
          unique values:  5                         missing .:  20,707/31,167

             tabulation:  Freq.   Numeric  Label
                          6,959         1  a. During the past 12 months
                          1,547         2  b. Between 12 and 24 months ago
                            602         3  c. More than 24 months ago
                            211         4  d. Never
                          1,141         5  e. Not sure
                         20,707         .

h25                 When was the last time you saw a dentist for a check-up, exam, teeth cleaning, o
-------------------------------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  HYSH25FMT

                 range:  [1,5]                        units:  1
          unique values:  5                         missing .:  20,907/31,167

             tabulation:  Freq.   Numeric  Label
                          7,675         1  a. During the past 12 months
                          1,172         2  b. Between 12 and 24 months ago
                            559         3  c. More than 24 months ago
                            107         4  d. Never
                            747         5  e. Not sure
                         20,907         .
```

If you wanted to determine who visited both a doctor and a dentist, you can create a new variable "visitboth" with all values designated as missing. To do this type "gen visitboth=." This ensures that you will only add in the respondents you want.

```
gen visitboth=.
```

For those who visited both a doctor and a dentist in the past 12 months, we want respondents who answered "during the past 12 months" for both of the questions. The following symbols are needed to tell STATA what to do:

Use "=" to assign the numeric value to the response option for the new variable Use "==" to designate which variable response options you are using

Use "&" to symbolize the word "and"

Below is an example of how you would use the symbols mentioned above to tell STATA the conditions for designating those who visited both as one:

```
replace visitboth=1 if (h24==1 & h25==1)
```

For those who didn't visit either a doctor or a dentist in the past 12 months, we want respondents who answered "between 12 and 24 months ago" or "more than 24 months ago" or "never." To tell STATA what to do, use "|" to symbolize the word "or" (use shift and hit "\").
Below is an example of how you would use this symbol above to tell STATA the multiple conditions for designating those who did not visit both as zero:

```
replace visitboth=0 if (h24==2 | h24==3 | h24==4 | h25==2 | h25==3 | h25==4)
```

When generating variables with "replace", make sure respondents who didn't answer both questions are excluded and tell STATA to set them to missing:

```
replace visitboth=. if (h24==. & h25==.)
tab visitboth grade
```

```
. gen visitboth=.
(31,167 missing values generated)

. replace visitboth=1 if (h24==1 & h25==1)
(5,816 real changes made)

. replace visitboth=0 if (h24==2 | h24==3 | h24==4 | h25==2 | h25==3 | h25==4)
(3,354 real changes made)

. replace visitboth=. if (h24==. & h25==.)
(0 real changes made)

. tab visitboth grade
```

visitboth	Grade 8	10	12	Total
0	968	1,420	966	3,354
1	1,951	2,486	1,379	5,816
Total	2,919	3,906	2,345	9,170

Notice that there are no results for 6th grade because these questions were not asked of 6th graders.

Recoding can be tricky because it is not just one-sided coding. You need to include exactly the respondents you want and exclude the respondents you don't want.

Labeling New Variables

Once you have created a new variable or recoded response options, you may want to create labels for them. Use the following commands to create labels:

- "lab var" or "label variable" ~ adds a description to your variable
- "lab def" or "label define" ~ creates response option labels (once you create a response option label, you can reuse it over and over with other variables)
- "lab val" or "label value" ~ applies response option labels to your variable

```
lab var visitboth "visited both a doctor and a dentist in the past year"
lab def visit 1"both" 0"one or none"
lab val visitboth visit
```

tab visitboth

visited both a doctor and a dentist in the past year	Freq.	Percent	Cum.
one or none	3,354	36.58	36.58
both	5,816	63.42	100.00
Total	9,170	100.00	

NOTE: For more information on labeling, type the command “help label” in STATA

General Rules on Creating Dichotomous or Binary Variables

When creating dichotomous or binary variables, HYS uses the guidelines from the CDC's Youth Risk Behavior Survey when possible so we can compare ourselves to national data.

When calculating dichotomous or binary variables, in general, the numerator is the percent saying "Yes." The denominator is either all students or a subset of students who have indicated in the current survey they participate in a selected activity or behavior. Students must have provided valid data to be included in any dichotomous variable calculations.

Therefore, students with missing responses are not included. Some examples are included below.

Example 1:

Question: Has a doctor or nurse ever told you that you have asthma?

a. Yes
b. No
c. Not sure

Summary text: Percentage of students who had ever been told by a doctor or nurse that they had asthma

Numerator: Students who answered a. Yes
Denominator: Students who answered a. Yes, b. No, or c. Not sure
gen asthmalifetime=h22
recode asthmalifetime 1=1 2/3=0

Example 2:

Question: During the past 30 days, if you used alcohol, what type of alcohol did you usually drink?

a. I did not drink alcohol during the past 30 days.
b. I do not have a usual type.
c. Beer
d. Flavored malt beverages, such as Smirnoff Ice, Bacardi Silver, or Hard Lemonade
e. Wine coolers, such as Bartles & Jaymes or Seagrams
f. Wine
g. Liquor, such as vodka, rum, scotch, bourbon or whiskey
h. Some other type

Summary text: Among students who said they drank alcohol in the past 30 days in this question, the percentage who drank beer

Numerator: Students who answered c. Beer
Denominator: Students who answered b, c, d, e f, g, or h
Students who answered a. I did not drink alcohol during the past 30 days are set to missing
gen drankbeer=d94
recode drankbeer 1=. 2=0 3=1 4/8=0

Example 3:

Question: How much do you think people risk harming themselves if they smoke one or more

packs of cigarettes per day?

a. No risk
b. Slight risk
c. Moderate risk
d. Great risk
e. Not sure

Summary text: Percentage of students who said they "great risk" from pack a day smoking

Numerator: Students who answered d. Great risk
Denominator: Students who answered a, b, c, d, or e
gen ciggreatrisk=p01
recode ciggreatrisk 1/3=0 4=1 5=0

OR if you don't want to count the students who said "Not sure"
Numerator: Students who answered d. Great risk
Denominator: Students who answered a, b, c, d
gen ciggreatrisk=p01
recode ciggreatrisk 1/3=0 4=1 5=.

NOTE: If a question has a "Not sure" response, you need to decide if the "Not sure" respondents are included in the denominator. If "Not sure" means "No" to you because they didn't answer "Yes," then combine "No" and "Not sure" together. If "Not sure" means the respondent didn't have a response or didn't understand the question, then you should not include them in the denominator. This is different from the Behavioral Risk Factor Surveillance Survey (BRFSS), a telephone survey of adults that allows the caller to keep probing for a Yes/No response.

Two-Way Tables or Crosstabs

"Svy" is a prefix used with STATA commands when you are analyzing survey data. "Svy" takes your weighting, psu, strata, etc. into account when you are running estimation commands. "Svy:tab" is a tabulation command. It also provides you with a test of independence.

Example of crosstab using variables:

h53: During the past 12 months, did you ever feel so sad or hopeless almost every day for two weeks or more in a row that you stopped doing some usual activities? (no/yes)

g05_18: What sex/gender were you at birth, even if you are not that gender today?

```
svy:tab h53 g05_18
```

```
During the past 12 months, did you ever feel so sad or hopeless almost |          Are you:
every day                              | a. Femal   b. Male     Total
---------------------------------------+----------------------------------
                                   b. No |     .2541     .3597     .6138
                                  a. Yes |     .2541     .1321     .3862
                                         |
                                   Total |     .5082     .4918         1
---------------------------------------+----------------------------------
  Key:  cell proportion

  Pearson:
    Uncorrected   chi2(1)         = 1152.4647
    Design-based  F(1, 156)       =  861.2373     P = 0.0000
```

Interpretation: There are four cells in the two-way table. The results in the four cells add up to 100%: female no (25%) + female yes (25%) + male no (36%) + male yes (13%) = 100%

The key below the total row of the table reminds you that the results are displayed as cell proportions.

Underneath the key, the output gives you the results of a Pearson correlation test. If the P (p value) is less than 0.05, then one of the cells is significantly different than the others at a 95% confidence level.

Additional Options with "Svy"

There are a number of additional options that can be added to a "svy:tab" to change the way the data is displayed or to provide you with more information. To use the additional options, type a comma (,) after the variables.

Col and Row

"col": gives you column percents. In this example, results are displayed for females no/yes in the first column and for males yes/no in the second column. **Each column adds up to 100%.**

```
svy:tab h53 g05, col
```

```
ever feel so sad or hopeless almost |          Are you:
every day                           | a. Femal   b. Male     Total
------------------------------------+----------------------------------
                                b. No |        .5     .7314     .6138
                               a. Yes |        .5     .2686     .3862
                                      |
                                Total |         1         1         1
------------------------------------+----------------------------------
```

Interpretation: 50% of females and 27% of males have experienced depressive feelings in the past year.

"row": gives you row percents. In this example, results are displayed for no female/male in the first row and yes female/male in the second row. **Each row adds up to 100%.**

svy:tab h53 g05, row

ever feel so sad or hopeless almost every day	Are you: a. Femal	b. Male	Total
b. No	.414	.586	1
a. Yes	.658	.342	1
Total	.5082	.4918	1

Interpretation: Of those students who experienced depressive feelings in the past year, 66% were female and 34% were male.

NOTE: Remember if "col" or "row" are not specified, the cells in the entire table add up to 100%.

Obs

Adding "obs" at the end of the "svy:tab" command will give you the number of observations in each cell, each column, each row, and the total observations.

SE and CI

You can also add options at the end of "svy:tab" to give you the standard error (se) and 95% confidence intervals (ci).

Percentages

The "per" or "percent" command allows you to display the point estimates as percentage points.

svy:tab h53 g05, col obs se ci per

ever feel so sad or hopeless almost every day	Are you: a. Femal	b. Male	Total
b. No	50	73.14	61.38
	(.7625)	(.7641)	(.6591)
	[48.49,51.5]	[71.61,74.62]	[60.07,62.67]
	5184	7339	1.3e+04
a. Yes	50	26.86	38.62
	(.7625)	(.7641)	(.6591)
	[48.5,51.51]	[25.38,28.39]	[37.33,39.93]
	5185	2695	7880
Total	100	100	100
	1.0e+04	1.0e+04	2.0e+04

Interpretation:
Among females, percent who experienced depressive feelings in the past year: 50.0% = point estimate

- 50.0% = point estimate
- ±1.4% = symmetric 95% confidence interval (calculated by multiplying the standard error 0.7625 * 1.96 = 1.4%)
- [48.5%, 51.5%] = non- symmetric 95% percent upper and lower bound confidence intervals
- 5185 respondents

Among males, percent who experienced depressive feelings in the past year:

- 26.7% = point estimate
- ±1.5% = symmetric 95% confidence interval (calculated by multiplying the standard error 0.7641 * 1.96 = 1.5%)
- [25.4%, 28.4%] = non- symmetric 95% percent upper and lower bound confidence intervals
- 2695 respondents

Additional Tips for Formatting Data

The following commands can be used to format your output into a more understandable and readable format:

Widening table columns

Use stubwidth and cellwidth to change the size of your columns so that all of the label text can be displayed:

```
svy:tab s01 g05, row ci stubwidth (20) cellwidth (15)
```

How often do you feel the schoolwork you are assigned is meaningful and importan	What sex/gender were you at birth, even if you are not that gender today? a. Female	b. Male	Total
a. Almost always	.5174 [.4957,.5391]	.4826 [.4609,.5043]	1
b. Often	.4945 [.4794,.5097]	.5055 [.4903,.5206]	1
c. Sometimes	.5255 [.5137,.5373]	.4745 [.4627,.4863]	1
d. Seldom	.5031 [.4806,.5256]	.4969 [.4744,.5194]	1
e. Never	.4233 [.3972,.4499]	.5767 [.5501,.6028]	1
Total	.5039 [.4955,.5123]	.4961 [.4877,.5045]	1

Rounding

The "format" command allows you to specify the display format for variables. When used as below, the number after the period allows you to indicate how many decimal points you want to show (thus 0 means to round the results to a whole number).

If you need the exact number of observations, use the format command to tell STATA how many numbers you would like it to display before and after the decimal point.

The rounding used to produce HYS results has changed over time. For consistency, it is recommended that analysis be produced to the 100th and rounded to the 10th or a whole number. E.g., use format(%3.2f) to produce the results 52.45% and report as 52.5% or 52%.

```
svy:tab grade g05, per row ci format(%3.2f)
```

Grade	What sex/gender were you at birth, even if you are not that gender today? a. Femal	b. Male	Total
6	49.48 [48.19,50.76]	50.52 [49.24,51.81]	100.00
8	50.11 [49.12,51.11]	49.89 [48.89,50.88]	100.00
10	50.63 [49.63,51.63]	49.37 [48.37,50.37]	100.00
12	49.53 [47.76,51.30]	50.47 [48.70,52.24]	100.00
Total	49.99 [49.36,50.61]	50.01 [49.39,50.64]	100.00

Removing Scientific Notation

Rounding can also be useful if you have large numbers of observations and your results come out in scientific notation.

svy:tab grade g05, row per obs format(%9.3f)

Grade	What sex/gender were you at birth, even if you are not that gender today? a. Femal	b. Male	Total
6	49.476 4109.000	50.524 4196.000	100.000 8305.000
8	50.113 3776.000	49.887 3759.000	100.000 7535.000
10	50.632 4603.000	49.368 4488.000	100.000 9091.000
12	49.525 2764.000	50.475 2817.000	100.000 5581.000
Total	49.987 15252.000	50.013 15260.000	100.000 30512.000

In this example of the option "format(%9.3f)", the 9 tells STATA to display up to 9 digits before the decimal point and the .3 tells it to display 3 digits after the decimal point. You can see how this affects both the point estimate (in the previous example when format was not specified, 4 digits were displayed after the decimal point) and how it affects the observations. Experiment with the numbers in the format command to get your ideal display.

Vertical Alignment

The "vert" or "vertical" command will display your upper and lower bound confidence intervals (ci) on top of each other and without the bracket and comma. This can be useful if you are copying your results into Excel.

svy:tab grade g05, row ci per vert

Grade	What sex/gender were you at birth, even if you are not that gender today? a. Femal	b. Male	Total
6	49.48	50.52	100
	48.19	49.24	
	50.76	51.81	
8	50.11	49.89	100
	49.12	48.89	
	51.11	50.88	
10	50.63	49.37	100
	49.63	48.37	
	51.63	50.37	
12	49.53	50.47	100
	47.76	48.7	
	51.3	52.24	
Total	49.99	50.01	100
	49.36	49.39	
	50.61	50.64	

For more information on format, type the command help format or see the Additional Tips for Formatting Data section in this manual.

Stratified Analysis and Subpopulations

Often you want to look at crosstab results among specific subpopulations, i.e. among certain grade levels, races, etc. One simple way is to use "drop" or "keep" commands to limit your dataset to only the subgroup you are interested in. For example if you are only looking at results among 8th grade students:
keep if grade==8 will remove students from all of the other grades.
drop if grade==8 will remove 8th grade students, but keep other graders.

NOTE: Make sure you do not save over your data file after using a keep or drop command. Doing so will overwrite your file and you will lose the records that were there previously.

If you are only looking at results among students who used marijuana in the past 30 days:
keep if d21_16use==1 will only keep the current smokers in your dataset.

Another option is to use the subpop command. Any binary variable that is coded as 0, 1 can be used as a subpopulation. Examples for making subpop variables:

Creates a subpop of only current smokers
```
gen marij30=d20_16use
recode marij30=1=1 0=0
```

Creates a subpop of only Black-African Americans
```
gen black=g06
recode black 1=0 2=0 3=1 4=0 5=0 6=0 7=0 8=0
```

Creates a subpop of only 8th grade students (ok to use replace since there are no missing respondents in the grade variable, but check the number of missing before using this command

for any other variable as missing values will be coded 0)

```
tab grade, missing
gen eight=1 if grade==8
replace eight=0 if grade~=8
```

You can also create new combined variables for subpops, for example, this creates a subpop of only 8th grade Black-African American students:

```
gen black8=g06
recode black8 1=0 2=0 3=1 4=0 5=0 6=0 7=0 8=0
replace black8=. if grade~=8
```

The best way to create subpops is to create "dummy" variables. This command will generate a new variable for each response option:

```
tab grade, gen(gradecat)
```

Creates four new variables:

- gradecat1 (for 6th grade)
- gradecat2 (for 8th grade)
- gradecat3 (for 10th grade)
- gradecat4 (for 12th grade)

NOTE: Four dummy variables will be created if you still have all four grades left in your dataset. If for example you have dropped 6th grade, then you will only get three dummy variables and gradecat1 will be 8th grade.

Once you have your subpop variables created, you can use them with svy:tab.

For example, to look at marijuana use in the home by current marijuana use among 8th graders:

```
svy:tab d21_16use d99, subpop(gradecat2) col per
```

Used marijuana - Any use in the past 30 days	Does anyone who lives with you now use marijuana?		
	a. No	b. Yes	Total
no	86.49	13.51	100
yes	47.52	52.48	100
Total	85.33	14.67	100

```
  Key:  row percentage

  Pearson:
    Uncorrected   chi2(1)         =  944.2067
    Design-based  F(1, 227)       =  125.6666     P = 0.0000
```

Interpretation: Among 8th graders who use marijuana, 52% live with someone who uses marijuana. Among 8th graders who do not use marijuana, 14% live with a marijuana user. The p-value is 0.0000, so 8th graders who use marijuana are more likely to live with someone who uses marijuana compared to 8th graders who don't use marijuana.

Note that the p-value does not actually equal 0, but the value is smaller than the number of digits the STATA output shows. In this case, for example, the p-value is less than 0.0000.

You could also look at this the other way, by switching the variable order, i.e., to look at current marijuana use by marijuana use in the home:

```
svy:tab d99 d21_16use, subpop(gradecat2) col per
```

```
Does anyone who lives with you now use marijuana ?  | Used marijuana - Any use in the past 30 days
                                                    |        no        yes      Total
----------------------------------------------------+----------------------------------
                                              a. No |     98.34      1.664        100
                                             b. Yes |     89.31      10.69        100
                                                    |
                                              Total |     97.01      2.987        100
----------------------------------------------------+----------------------------------
  Key:  row percentage

  Pearson:
    Uncorrected   chi2(1)         =  944.2067
    Design-based  F(1, 227)       =  125.6666     P = 0.0000
```

Interpretation: Among 8th graders who live with a marijuana user, 11% use marijuana. Among 8th graders who do not live with a marijuana user, 2% use marijuana. The p-value is 0.000, so 8th graders who live with a marijuana user are more likely to use marijuana compared to 8th graders who do not live with a marijuana user.

Another way to conduct stratified analyses is to use the "over" command. The "over" command replaces the "by" command used in previous versions of STATA (version 8 and earlier). The variable or variables in parentheses after the over command define your subpopulations, e.g., to look at current marijuana use by grade and gender. Since you are looking at the mean, make sure that the response you are interested in is equal to 1 and the other responses are equal to 0.

```
svy:mean d21_16use, over(grade g05_18)
```

```
                                |             Linearized
                                |       Mean   Std. Err.     [95% Conf. Interval]
--------------------------------+------------------------------------------------
c.d21_16use@grade#g05_18        |
                     6#a. Female |   .0093333   .0019106      .0055686    .0130981
                       6#b. Male |    .008291   .0017548      .0048332    .0117488
                     8#a. Female |   .0340619   .0040764      .0260295    .0420942
                       8#b. Male |   .0226404   .0036666      .0154154    .0298654
                    10#a. Female |   .0791823   .0077533      .0639047    .0944599
                      10#b. Male |   .0641026   .0052315       .053794    .0744111
                    12#a. Female |   .1722507   .0164905      .1397567    .2047447
                      12#b. Male |   .1460406   .0136912      .1190625    .1730187
--------------------------------+------------------------------------------------
```

Interpretation: Current marijuana use for 12th grade females is 17.2%. Current smoking for 12th grade males is 14.6%.

You can also use this with a continuous variable like BMI to get a mean by grade and gender:

```
svy:mean bmi, over(grade g05_18)
```

	Mean	Linearized Std. Err.	[95% Conf.	Interval]
c.bmi@grade#g05_18				
8#a. Female	21.85369	.3072676	21.24668	22.46069
8#b. Male	22.37541	.3845874	21.61566	23.13516
10#a. Female	23.11699	.405341	22.31624	23.91773
10#b. Male	23.75519	.3652627	23.03362	24.47677
12#a. Female	24.22337	.3213671	23.58851	24.85823
12#b. Male	25.17303	.4184388	24.34641	25.99965

Interpretation: The mean BMI for 8th grade females is 21.9. The mean BMI for 8th grade males is 22.4.

HYS Data Analysis – Quick Examples

This section provides a few examples of how to run crosstab analyses in STATA with:

STATA setup commands for analysis:

- State sample data
- State census data
- County sample, census, or mixed sampling data
- ESD level analysis

STATA commands for simple crosstabs:

- One variable by grade
- One variable by grade and gender
- One recoded variable by race and grade
- Two variables by grade
- Two variables by race
- Two variables by grade and gender

For a hands on experience, a STATA "do file" was provided with this manual. The do file is available in the following Appendix:

- Appendix B: Do File ~ Quick Examples of HYS Data Analysis in STATA

Setup for Survey Analysis

The following STATA commands can be used to set up each different level of analysis. For more information on setup commands see the section General Setup for Survey Analysis.

NOTE: Some datasets do not have the variable "schgrd," but instead have the variable "schgnoid" from 2002-2014 and "psu" in 2016-2018If your dataset has "psu" or "schgnoid", use it in place of schgrd in these STATA commands.

State Sample Data

```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

State Census Data

```
gen fakewt=1
svyset [pweight=fakewt]
```

County Sample Data ~ for Counties without Samples (Census)

In 2021, these counties included: Adams, Asotin, Benton, Clark, Chelan, Clallam, Columbia, Cowlitz, Douglas, Ferry, Franklin, Garfield, Grant, Grays Harbor, Island, Jefferson, Kitsap, Kittitas, Klickitat, Lewis, Lincoln, Mason, Okanogan, Pacific, Pend Oreille, San Juan, Skagit, Skamania, Stevens, Thurston, Wahkiakum, Walla Walla, Whatcom, Whitman, Yakima.

```
*keep if conum==X
*Insert your county number (conum) for X (see Demographic variables, conum)
keep if corec==1
gen fakewt=1
svyset [pweight=fakewt]
```

County Sample Data ~ for Sampled Counties

In 2021, these counties included: Pierce, King, Snohomish.

```
*keep if conum==X
*Insert your county number (conum) for X (see Demographic variables, conum)
keep if corec==1
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

County Sample Data ~ for Counties with Mixed Sampling

In 2021, counties with mixed sampling (some grades census and some grades sampled) included:

Spokane: 6th and 8th grades were sampled - 10th and 12th grades were census

```
keep if conum==32
keep if corec==1
gen fakewt=1
gen id=_n
gen psu=id +10000
replace psu=schgrd if (grade==6 | grade==8)
svyset [pweight=fakewt], psu(psu)
```

NOTE: Make sure not to run analysis for any county/grade levels that didn't meet county-level reporting requirements

Regional ESD Data

ESD regions are made up of counties or parts of counties, some which are sampled and some which are census. The following coding will set up analysis of any ESD.

```
*keep if esdnum==X
*Insert your ESD number (esdnum) for X
keep if esdrec==1
gen id = _n
gen psu=id + 10000
replace psu=schgrd if (conum==17)
replace psu=schgrd if (conum==27)
replace psu=schgrd if (conum==31)
replace psu=schgrd if conum==32 & (grade==6 | grade==8)
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

Data Analysis Example

*Current marijuana use by grade

To run one variable, d21_16use (current marijuana – already coded as no use or any use 0,1) by grade and include the following formatting options after the comma.

- col for column percentages
- per for results displayed in %
- se for standard error (to convert se to 95% ci, you need to *1.96)
- ci for upper/lower confidence intervals
- obs for "n"
- format(%3.2f) to designate the numbers before and after the decimal

```
svy:tab d21_16use grade, col per se ci obs format(%3.2f)
```

*Current marijuana use by grade and sex assigned at birth

Generate a binary (1,0) sex dummy variable for subpopulations for marijuana use by grade among a specific sex.

```
tab g05_18, gen(sex)
rename sex1 female
rename sex2 male
svy:tab d21_16use grade, subpop(girl) col per se ci obs format(%3.2f)
svy:tab d21_16use grade, subpop(boy) col per se ci obs format(%3.2f)
```

Generate a binary (1,0) grade dummy variable for subpopulations for marijuana use by sex at birth for a specific grade.

```
tab grade, gen (gradecat)
svy:tab d21_16use g05_18, subpop(gradecat1) col per se ci obs format(%3.2f)
svy:tab d21_16use g05_18, subpop(gradecat2) col per se ci obs format(%3.2f)
svy:tab d21_16use g05_18, subpop(gradecat3) col per se ci obs format(%3.2f)
svy:tab d21_16use g05_18, subpop(gradecat4) col per se ci obs format(%3.2f)
```

*Current marijuana by grade and chronic absenteeism

Create a chronic absenteeism variable (absent 3 or more days) with recode, then define and attach the new response option labels and add the new variable with a description. Crosstab marijuana by race and grade.

```
codebook g27
gen chronicabsent= g27
recode chronicabsent 1/2=0 3=1
lab def chronicabsent 1 "Yes-3 or more days" 0"No"
lab val chronicabsent chronicabsent
lab var chronicabsent "Absent from school on 3 or more days in past month for any reason"
svy:tab d21_16use chronicabsent, subpop(gradecat2) col per se ci obs format(%3.2f)
svy:tab d21_16use chronicabsent, subpop(gradecat3) col per se ci obs format(%3.2f)
svy:tab d21_16use chronicabsent, subpop(gradecat4) col per se ci obs format(%3.2f)
```

*Current marijuana use by depressive feelings

Crosstab marijuana by depressive feelings and grade.

```
svy:tab d21_16use h53, subpop(gradecat2) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(gradecat3) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(gradecat4) col per se ci obs format(%3.2f)
```

Current marijuana use by depressive among chronic absentee students

Generate binary (0,1) for chronic absenteeism and grade subpopulations.

```
gen absent8=1 if chronicabsent==1 & grade==8
replace absent8=0 if chronicabsent==0 & (grade==10 | grade==12)
gen absent10=1 if chronicabsent==1 & grade==10
replace absent10=0 if chronicabsent==0 & (grade==8 | grade==12)
gen absent12=1 if chronicabsent==1 & grade==12
replace absent12=0 if chronicabsent==0 & (grade==8 | grade==10)
svy:tab d21_16use h53, subpop(absent8) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(absent10) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(absent12) col per se ci obs format(%3.2f)
```

Current marijuana use by depressive feelings among boys

```
gen boy8=1 if g05_18==2 & grade==8
replace boy8=0 if g05_18==1 &  (grade==10 | grade==12)
gen boy10=1 if g05_18==2 & grade==10
replace boy10=0 if g05_18==1& (grade==8 | grade==12)
gen boy12=1 if g05_18==2 & grade==12
replace boy12=0 if g05_18==1 & (grade==8 | grade==10)
svy:tab d21_16use h53, subpop(boy8) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(boy10) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(boy12) col per se ci obs format(%3.2f)
```

NOTE: Use caution with crosstabs of variables with low prevalence, or when you are using small subpopulations. Do NOT report results if there are less than 5 observations per cell when running state level data or less than 10 observations per cell when running sub-state-level analysis.

Comparing State and Local Data

This section describes how to compare local data to the state. How you compare data depends on the type of data you have and the types of comparisons you want to make.

The easiest way to compare state and local data is to use the HYS Reports of Results. Reports of Results were generated by the HYS survey contractor, Looking Glass Analytics for school buildings, districts, ESD regions, and counties that participated in the survey at the minimum level. Reports of results for the state sample, state sample subpopulations (gender and race), and counties are available on the AskHYS.net website. Building, district, ESD, and county reports also include the state sample results so comparisons can be made by using the confidence intervals to determine differences. (If confidence intervals do not overlap then the difference is statistically significant.) This is always a good first step, even if you go on to run your comparisons in STATA. You can confirm your results with the produced reports.

Now that you have these results, you can also use an "Excel Tool for Determining Statistical Significance" available at: https://www.askhys.net/Training. For more information about this tool, see Checking Findings and Significance Online.

You can also do formal comparisons for statistical testing with STATA. How you make comparisons depends on the datasets you have and the comparison you want to run.

There are two ways to make state and local comparisons:

1. Comparing local to state results that don't include the local results (the rest of the state)
2. Comparing local to the complete state sample

We recommend that when determining statistically significant differences between your local data and state data, that you compare local to the rest of the state sample (i.e., the state sample minus your local results).

When you report percentage point estimates and confidence intervals for the state sample, you may want to use the full state sample results so that you do not contradict previously published state results. If you do this, you should note in your methods or under your results where your results came from and how your comparisons were conducted.

Appending

Use the "append" command if you want to add datasets with similar variables. For example, if you wanted to combine your local and state sample Healthy Youth Survey results, use the "append" command. Appending simply adds the additional data respondents below to the original respondents matching up the responses to the variable names.
Note: STATA defines your original data (the one you open first) as the "master data" and the new data you are appending on as your "using data."

Data Preparation:

Create a new variable that will differentiate the respondents from each dataset. Open your 2021 state sample dataset and create a new variable for location:

```
use "C:2021 state.dta"
gen location=0
```

If you are comparing your results to the rest of the state, you also need to drop any of your local schools from the state sample. Double check your conum variable to see if it was dropped and then save the file under a new name.

```
*drop if conum==X
*Insert your county number (conum) for X (see Demographic variables, conum)
tab conum
```

Be careful to save your new dataset under a different name. Don't save over your original state sample dataset.

```
save "C:2021 state location.dta"
```

Open your 2018 local dataset and create the same location variable with a different value:

```
use "C:2021 local.dta"
gen location=1
keep if corec==1
save "C:2021 local location.dta"
```

Sometimes it is useful to include only the variables that you will need for your analysis. Use the "drop" or "keep" command to get rid of any unnecessary variables in both datasets before you append. This can speed up analysis and decrease the chance that STATA may become confused during the append.

Append the data:

Open your new 2021 state dataset and append using with the 2021 local dataset:

```
use "C: 2021 state location.dta"
append using "C:2018 local location.dta"
```

You can also use the dropdown menus in more recent versions of STATA to append. Open your 2021 state location dataset, then select Data, Combine Datasets, Append Datasets. Browse to find your 2021 local dataset, and hit Submit.

Label your new location variable:

```
lab var location "state and local identifier"
lab def location 0"state" 1"county"
lab val location location
```

Append Investigation:

It is important to verify that your append came out correctly. To make sure that all of the data is there, run a tab by location to see if you have the same number of respondents as you did in both of the original datasets (you should not have any missing data):

```
tab location, missing
```

If everything looks good, save your new combined dataset with a new file name:

```
save "C: 2021 state and local combo.dta"
```

Comparing Local vs. the Rest of the State Sample

Now that you have a combined state and local dataset, you need to open it and set it up for survey analysis:

```
use "C: 2021 state and local combo.dta"
```

If your local data is a sample (such as King, Pierce, Snohomish counties) then set up for analysis with:

```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

If your local data is census data (most counties, all districts and schools) then set up for analysis with:

```
gen fakewt=1
gen id = _n
gen psu = id + 5000
replace psu = schgrd if location==0
svyset [pweight=fakewt], psu(psu)
```

This creates a psu with individual responses for your local census and groups school building responses for the state sample.

Now you are ready to run a svy:tab by your group variable. You will need to first create a subpopulation to run your variable by a specific grade:

```
tab gradecat
svy:tab d14use location, subpop(eight) col se obs
```

Comparing Local vs. the Complete State Sample

Instead of dropping your local results from the state sample, you can also compare your local results to the complete state sample.

Append the data:

Open your 2021 state dataset and append using with the 2021 local dataset:

```
use "C: 2021 state.dta"
gen location=0
lab var location "state and local identifier"
lab def location 0"state" 1"county"
lab val location location
append using "C:2021 local.dta"
```

Then to label your local results location, use:

```
replace location=1 if location==.
```

If your local data is a sample (such as King, Pierce, Snohomish counties) then set up for analysis with:

```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

If your local data is census data (most counties, all districts and schools) then set up for analysis with:

```
gen fakewt=1
gen id = _n
gen psu = id + 5000
replace psu = schgrd if location==0
svyset [pweight=fakewt], psu(psu)
```

This creates a psu with individual responses for your local census and groups school building responses for the state sample.

Now you are ready to run a svy:tab by your group variable. You will need to first create a subpopulation to run your variable by a specific grade:

```
svy:tab d14use location, subpop(eight) col se obs
tab grade, gen(gradecat)
```

Comparing Years of Data

This section describes how to combine multiple years of HYS data. It includes information about how to use the append command. Append allows you to add more respondents to your data.

Note: STATA defines your original data as the "master data" and the new data you are appending on as your "using data."

Appending

Use the "append" command if you want to add datasets with similar variables. For example, if you wanted to combine your 2021 and 2018 Healthy Youth Survey results you can use the "append" command. Appending simply adds the additional data respondents below to the original respondents matching up the responses to the variable names.

Data Preparation:

You need to create a new variable that will differentiate the respondents from each datasets. Open your 2021 dataset and create a new variable for year:

```
use "C:2021 data.dta"
gen year=2021
save "C: 2021 data year.dta"
```

Open your 2018 dataset and create the year variable:

```
use "C: 2018 data.dta"
gen year=2018
save "C: 2018 data year.dta"
```

Sometimes it is useful to include only the variables that you will need for your analysis. Use the "drop" or "keep" command to get rid of any unnecessary variables in both datasets before you append. This can speed up analysis and decrease the chance that STATA may become confused during the append.

Append the data:

Open your 2018 dataset and append using:

```
use "C:2018 data year.dta"
append using "C:2016 data year.dta"
```

It's always best to append older data onto the new data, that way the newest variable labels and formats will be included in your appended dataset.

Append Investigation:

It is important to verify that your append came out correctly. In general, the 2021 variables will stay in the same order and any variables that were unique to the 2018 data will now be at the bottom of your variable list. To make sure that all of the data is there, run a tab by year to see if you have the same number of respondents as you did in both of the original datasets:

```
tab year, missing
```

You should not have any missing data. You may also want to run some frequencies to verify that you are getting the same results as you were before your append.

If everything looks correct, save your new combined dataset with a new file name:
```
save "C:2018 and 2016 combo.dta"
```

Analysis Stratified by Year

At this time, we are not recommending that you use STATA to determine significant trends over time. For trend analysis, we recommend that you have at least 5 data points and use a regression analysis program like Joinpoint.

Joinpoint is available at: https://surveillance.cancer.gov/joinpoint/

You can use STATA to determine changes from a single survey administration to another, e.g., a change from 2018 to 2021.

Now that you have a combined year dataset, you need to open it and set it up for survey analysis. The following is a comparison of current alcohol use for 8th and 10th graders from 2018 to 2021 using the state sample:
```
use "C:2021 and 2018 combo.dta"
```

If you are comparing 2018 to 2021 state sample data, or local sample data (such as King, Pierce, Snohomish counties) then set up for analysis with:
```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

If you are comparing 2018 to 2021 local census data (most counties, all districts and schools) then set up for analysis with:
```
gen fakewt=1
svyset [pweight=fakewt]
```

If you are comparing local census data that is sampled one year and not the other, or in one grade or not the other, see HYS Data Analysis in STATA – the section General Setup for Survey Analysis and the sub-section on County with mixed sampling analysis.
Then you will need to create subpopulations to run your variable by a specific grade:
```
tab grade, gen(gradecat)
rename gradecat1 six
rename gradecat2 eight
rename gradecat3 ten
rename gradecat4 twelve
```

Compare current e-cig/vaping in the past two years among 8th graders.
```
svy:tab d90_16use year, subpop(eight) col se obs per format(%9.2f)
```

Electronic cigarettes, e-cigs, or vape pens - Any use in the past 30 days	Survey year		
	2018	2021	Total
no	89.54	95.15	93.00
	(0.71)	(0.49)	(0.49)
	3911.00	6674.00	10585.00
yes	10.46	4.85	7.00
	(0.71)	(0.49)	(0.49)
	457.00	340.00	797.00
Total	100.00	100.00	100.00
	4368.00	7014.00	11382.00

```
  Key:  column percentage
        (linearized standard error of column percentage)
        number of observations

  Pearson:
    Uncorrected   chi2(1)         =  666.7704
    Design-based  F(1, 314)       =   44.5965     P = 0.0000
```

Interpretation: 8th grade current e-cig/vaping use was 10.5% in 2018 and 4.9% in 2021. There was a significant decrease in current e-cig/vaping among 8th graders from 2018 to 2021 (p-value is 0.0000, less than 0.05).

Then compare current e-cig/vaping in the past two years among 6th graders.

```
svy:tab d20use year, subpop(six) col se obs per format(%9.2f)
```

Electronic cigarettes, e-cigs, or vape pens - Any use in the past 30 days	Survey year		
	2018	2021	Total
no	96.97	96.99	96.98
	(0.27)	(0.33)	(0.22)
	8571.00	7380.00	15951.00
yes	3.03	3.01	3.02
	(0.27)	(0.33)	(0.22)
	268.00	229.00	497.00
Total	100.00	100.00	100.00
	8839.00	7609.00	16448.00

```
  Key:  column percentage
        (linearized standard error of column percentage)
        number of observations

  Pearson:
    Uncorrected   chi2(1)         =    0.0264
    Design-based  F(1, 314)       =    0.0029     P = 0.9568
```

Interpretation: 6th grade current alcohol use was 3.0% in 2018 and 3.0% in 2021. There was no change in current alcohol among 6th graders from 2018 to 2021 (p-value is 0.957, greater than 0.05).

When to Combine Multiple Years of Data

We generally recommend that all analyses be done stratified by year.

However, under certain conditions, you may want to consider combining years of data. Some

possible reasons to combine years include:

1. If your crosstabs don't meet the minimum cell requirements (5 per cell for state and 10 per cell for local).
2. If you have small number of respondents, like in smaller counties, or when analyzing non-core items located toward the end of the survey form.
3. If the survey question has been asked consistently over the years and results are similar from year to year.
4. If you want to analyze variables that are only applicable to a small group, such as trying to find out how many students with current asthma visited an emergency room in the past year.

Please give extra attention to dot point 3 for 2021. Many questions were altered for 2021 to allow students to respond appropriately if they were attending school remotely. Results for 2021 may also be different from previous years due to COVID-19. Use caution when combining 2021 data with previous years.

Methods for Combining Years

We recommend the following decision rules for year-adjusted when you are considering using year-combined single grade estimates:

1. Crude: If there is no substantial difference in a factor across years, you could report a "crude" estimate and note that the results are from multiple years. If you run a factor by year in STATA, the Total column will give you this "crude" estimate, or you can run an analysis without year as a variable or subpopulation (e.g., not stratifying by grade). In this case, you want to set up your analysis to include "year" as strata, e.g., for state sample analysis:
   ```
   svy:set [pweight=fakewt], psu(schgrd) strata(year)
   ```
2. Average: If there is a significant difference in a factor by year, but the purpose of your analysis is simply to express the burden of a condition, then you can use an average of the year-specific results. Averaging gives equal weight to the results for each year, instead of giving equal weight to each respondent.

For example, if you wanted to estimate the percent of 10th graders who seriously considered suicide, run the factor by year then add the estimates for each year together and divide by the number of years, e.g., in 2018 and 2021 the statewide average seriously considering suicide was: 21.3%, calculated by (23.0+19.6)/2.

Unfortunately, this method does not give you a confidence interval.

3. Adjusted: If there is significant difference in a factor by year and the purpose of your analysis is to present an assessment of underlying factors that may lead to a condition, then it would be appropriate to use a year-adjusted estimate.

For example, if you are displaying the percent of youth smokers by gender who say that tobacco is easy to get and want to illustrate that it is different for males and females to inform planning, then you should use a year-adjusted estimate.

Year-Adjusted Estimates

Year-adjustment ensures that each year contributes equally to the overall percent estimate. This can be especially useful if the number of respondents differ by year, e.g., there was greater participation in 2018 than in 2021. To generate a year-adjusted estimate, you must weight the data.

Steps for Creating Year-Adjusted Estimates

Using the 2018 and 2021 state samples, here is the methodology for weighting the data to create year-adjusted estimates (i.e., combining both years 2018 and 2021).

According to the 2018-2019 and the 2021-2022 OSPI enrollment data for the state (available on their website: https://www.k12.wa.us/enrollment-reports), there are:

Grade	2018	2021	Combined
6th graders	87,276	80,738	168,014
8th graders	83,025	85,819	168,844
10th graders	83,460	84,871	168,331
12th graders	89,963	91,763	181,726

Looking at the number of valid respondents in the 2018 and 2021 state sample:

```
tab grade
```

Grade	2018	2021
6th graders	9,604	8,426
8th graders	8,895	7,691
10th graders	8,096	9,378
12th graders	5,676	5,672

For each grade, the enrollments for each year are added together to produce a combined enrollment number for the years. Then for each grade and year, the combined enrollment for that grade is divided by the total number of respondents for that specific grade and year.

```
gen yearwt=.
replace yearwt = 168014/9604 if (grade==6 & year==2018)
replace yearwt = 168014/8426 if (grade==6 & year==2021)

replace yearwt = 168844/8895 if (grade==8 & year==2018)
replace yearwt = 168844/7691 if (grade==8 & year==2021)

replace yearwt = 168331/8096 if (grade==10 & year==2018)
replace yearwt = 168331/9378 if (grade==10 & year==2021)

replace yearwt = 181726/5676 if (grade==12 & year==2018)
replace yearwt = 181726/5672 if (grade==12 & year==2021)
```

Year-Standardized Example

The following is an example looking at the prevalence of 10th graders missing school because they felt unsafe and carrying a weapon at school in the 2021 state sample:

```
gen unsafe=s20_21
```

```
recode unsafe 1=0 2/5=1 6=.
lab def days 1"any days" 0"no days"
lab val unsafe days
lab var unsafe "did not go to school because felt unsafe in past month"

gen carriedweapon=h39_21
recode carriedweapon 1=0 2/3=1 4=.
lab val carriedweapon days
lab var carriedweapon "carried weapon at school in past month"

tab grade, gen(gr)
rename gr3 ten
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)

svy:tab carriedweapon unsafe, subpop(ten) col se obs per format(%3.2f)
```

carried weapon at school in past month	did not go to school because felt unsafe in past month: no days	any days	Total
no days	98.25	95.54	98.03
	(0.27)	(1.21)	(0.25)
	3819.00	321.00	4140.00
any days	1.75	4.46	1.97
	(0.27)	(1.21)	(0.25)
	68.00	15.00	83.00
Total	100.00	100.00	100.00
	3887.00	336.00	4223.00

```
  Key:  column percentage
        (linearized standard error of column percentage)
        number of observations

  Pearson:
    Uncorrected   chi2(1)         =   72.8694
    Design-based  F(1, 168)       =    8.0269     P = 0.0052
```

Interpretation: Looking at 10th graders, it appears that carrying a weapon on any days is higher among those who didn't go to school on any days because they felt unsafe (4.5%) compared to those who didn't miss school due to safety (1.8%). Notice that the number of respondents is fairly small (n=15).

Using a combined 2018 and 2021 dataset (see the previous Appending section), first look at your results by year to see if combining them makes sense. Use extra caution when combining 2021 results with previous years, as there may be differences in results due to COVID-19.

```
use "C:2018 and 2021 combo.dta"

gen unsafe=s20_21
recode unsafe 1=0 2/5=1 6=.
replace unsafe=0 if s20==1
replace unsafe=1 if s20==2 | s20==3 | s20==4 | s20==5
lab def days 1"any days" 0"no days"
lab val unsafe days
lab var unsafe "did not go to school because felt unsafe in past month"

gen carriedweapon=h39_21
```

```
recode carriedweapon 1=0 2/3=1 4=.
replace carriedweapon =0 if h39_06==1
replace carriedweapon =1 if h39_06==2 | h39_06==3
lab val carriedweapon days
lab var carriedweapon "carried weapon at school in past month"

tab grade, gen(gr)
rename gr3 ten
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)

svy:tab unsafe year, subpop(ten) col se obs per format(%3.2f)
svy:tab carriedweapon year, subpop(ten) col se obs per format(%3.2f)
```

For 2021, recommend using caution when combining didn't go to school for feeling unsafe or carrying a weapon at school. Both of the questions changed in 2021, adding a response option for not going to school and both of the results significantly decreased from 2018 to 2021.

If you decide your results are appropriate to combine, they you can create the "yearwt" that we calculated previously by combining the enrollments from each year by grade and dividing them by the number of respondents for each year and grade. Then use that yearwt in your svyset command and set strata to year.

```
use "C:2018 and 2021 combo.dta"
gen yearwt=.
replace yearwt = 171494/9604 if (grade==6 & year==2018)
replace yearwt = 171494/8426 if (grade==6 & year==2021)
replace yearwt = 164180/8895 if (grade==8 & year==2018)
replace yearwt = 164180/7691 if (grade==8 & year==2021)
replace yearwt = 167712/8096 if (grade==10 & year==2018)
replace yearwt = 167712/9378 if (grade==10 & year==2021)
replace yearwt = 174215/5676 if (grade==12 & year==2018)
replace yearwt = 174215/5672 if (grade==12 & year==2021)
tab grade, gen(gr)
rename gr3 ten
svyset [pweight=yearwt], psu(schgrd) strata(year)
svy:tab unsafe carriedweapon, subpop(ten) col se obs per format(%3.2f)
```

carried weapon at school in past month	did not go to school because felt unsafe in past month		
	no days	any days	Total
no days	97.55	93.98	97.21
	(0.27)	(0.91)	(0.27)
	7004.00	703.00	7707.00
any days	2.45	6.02	2.79
	(0.27)	(0.91)	(0.27)
	176.00	45.00	221.00
Total	100.00	100.00	100.00
	7180.00	748.00	7928.00

Key: column percentage
(linearized standard error of column percentage)
number of observations

Pearson:
Uncorrected chi2(1) = 215.9437
Design-based F(1, 436) = 24.2171 P = 0.0000

Interpretation: Looking at 10th graders, it appears that carrying a weapon on any days is higher among those who didn't go to school on any days because they felt unsafe (6.0%) compared to those who didn't miss school due to safety (2.5%). Notice that the number of respondents who carried a weapon and missed school due to safety tripled (from 15 to 45) with two years of data. You also want to look to make sure the results between the two years are fairly similar – youth carrying weapons feeling unsafe was 4.5% in 2021 and is now 6.0% in with 2018/2021 combined, and youth carrying weapons and not feeling unsafe was 1.8% in 2021 and is now 2.5% in 2018/2021 combined.

Instead of combining years, you could consider combining grades, depending on how you want to present the results. For missing school due to safety, combining grades isn't recommended because the prevalence increases as students get older (p-value=0.0006). In this case and when analyzing the use of a substances where the prevalence usually increases by grade – it is better to combine years.

If results aren't different by grade, you can explore the next section on Combining Grade Levels.

Combining Grade Levels

This section describes when it is acceptable to combine grade levels and how to create grade-adjusted and high school estimates.

When to Combine Grades

We generally recommend that all analyses be done stratified by grade (see Analysis by Grade in Section 5). However, under certain conditions it may be desirable to combine the results from different grade levels. Some possible reasons to combine grades include:

1. If your crosstabs don't meet the minimum cell requirements (5 per cell for state and 10 per cell for local).
2. If you have a small number of respondents, like in smaller counties, or when analyzing non-core items located toward the end of the survey form.
3. If you want to analyze variables that are only applicable to a small group, such as trying to find out how many students with current asthma visited an emergency room in the past year.
4. If you need to produce a high school estimate for comparison with the YRBS – see Synthetic High School Estimates.
5. If you need to replicate combined grade results to match estimates in CPWI Data Books – see Data Book Combined Grade Estimates.

Methods for Combining Grades

We recommend the following decision rules for grade-adjustment when you are considering using grade-combined estimates for a single year of data to determine if you should report crude, average, or adjusted results.

Crude

If there is no substantial difference in a factor across grades, you could report a "crude" estimate and note that the results are from multiple grades. If you run a factor by grade in STATA, the Total column will give you this "crude" estimate, or you can run an analysis without grade as a variable or subpopulation (i.e., not stratifying by grade).

Average

If there is a significant difference in a factor by grade, but the purpose of your analysis is to simply express the burden of a condition, then you can use an average of the grade specific results. Averaging gives equal weight to the results for each grade, instead of giving equal weight to each respondent.

For example, if you wanted to estimate the percent of youth who seriously considered suicide, run the factor by grade then add the estimates for each grade together and divide by the number of grades, e.g., in 2021 the statewide average seriously considering suicide was: 19.7%,

calculated by (19.0% +19.6% + 20.4%)/3. Unfortunately, this method does not give you a confidence interval.

Adjusted

If there is significant difference in a factor by grade and the purpose of your analysis is to present an assessment of underlying factors that may lead to a condition, then it would be appropriate to use a grade-adjusted estimate.

For example, if you are displaying the percent of youth smokers by gender who say that tobacco is easy to get and want to illustrate that it is different for males and females in order to inform planning, then you should use a grade-adjusted estimate.

Grade-Adjusted Estimates

Grade-adjusted estimates ensure that each grade group contributes equally to the overall percent estimate, instead of giving equal "weight" to each respondent. They can be especially useful if the number of respondents differ by grade, e.g., there are more 8th grade respondents than 10th grade respondents. To generate a grade-adjusted estimate, you must weight the data.

This is similar to "age-adjusted" analyses often used in Healthy People 2020 or other national measures where population demographics change over time and may influence the factor you are trying to measure.

Steps for Creating Grade-Adjusted Estimates

Using the 2021 state sample, here is the methodology for weighting the data to create grade-adjusted estimates (i.e., combining all grades together 6, 8, 10 and 12):

Looking at the number of valid respondents in the 2018 state sample, there are:

```
tab grade
```

Grade	2021
6th graders	8,426
8th graders	7,691
10th graders	9,378
12th graders	5,672
Total to use for questions asked of all grades (6,8,10,12)	31,167
Total for questions asked only of secondary students (8,10,12)	25,495

The enrollments for each grade are added together to produce a combined enrollment number for the grades. Then for each grade, the combined enrollment is divided by the total number of respondents for that specific grade:

```
gen gradewt=.
replace gradewt = 31167/8426 if grade==6
replace gradewt = 31167/7691 if grade==8
replace gradewt = 31167/9378 if grade==10
replace gradewt = 31167/5672 if grade==12
svyset [pweight=gradewt], psu(schgrd)
```

Grade-Adjusted Example

The following is an example looking at missing school on five or more days because of toothache by whether or not they've been to a dentist in the past 2 years for 8th graders in the 2021 state sample. First, look at each of your variables by grade.

```
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
tab grade, gen(gr)
rename gr2 eight
rename gr3 ten
rename gr4 twelve
svy:tab misstooth dentistno, subpop(eight) col se obs per format(%3.2f)
```

missed school due to toothache on 5 or more days	have not been to a dentist in past 2 years		
	0	2 or mor	Total
5 or mor	98.86		
	(0.22)	(0.79)	(0.20)
	2851.00	174.00	3025.00
none-few	1.14		
	(0.22)	(0.79)	(0.20)
	33.00	2.00	35.00
Total	100.00	100.00	100.00
	2884.00	176.00	3060.00

```
svy:tab dentistno misstooth, subpop(ten)col se obs per format(%3.2f)
```

have not been to a dentist in past 2 years	missed school due to toothache on 5 or more days		
	5 or mor	none-few	Total
0	93.20		
	(0.64)	(3.97)	(0.64)
	3727.00	29.00	3756.00
2 or mor	6.80		
	(0.64)	(3.97)	(0.64)
	272.00	2.00	274.00
Total	100.00	100.00	100.00
	3999.00	31.00	4030.00

```
svy:tab dentistno misstooth, subpop(twelve)col se obs per format(%3.2f)
```

have not been to a dentist in past 2 years	missed school due to toothache on 5 or more days		
	5 or mor	none-few	Total
0	91.27	75.00	91.13
	(0.72)	(9.86)	(0.73)
	2163.00	15.00	2178.00
2 or mor	8.73	25.00	8.87
	(0.72)	(9.86)	(0.73)
	207.00	5.00	212.00
Total	100.00	100.00	100.00
	2370.00	20.00	2390.00

- It looks like missing school due to a toothache is higher among those who didn't go to the dentist in the past two years for all grades, but the number of respondents is small and unreportable (less than 5 per cell) for 8th and 10th grade: two 8th graders, two 10th graders, and five 12th graders.

To calculate grade-adjusted estimates we can create the "gradewt" that we calculated previously by combining the grade enrollments and dividing it by the number of respondents for each grade.

```
use "C:2021 data year.dta"
gen gradewt=.
replace gradewt = 25495/7691 if grade==8
replace gradewt = 25495/9378 if grade==10
replace gradewt = 25495/5672 if grade==12
svyset [pweight=gradewt], psu(schgrd)
svy:tab dentistno misstooth,col se obs per format(%3.2f)
```

have not been to a dentist in past 2 years	missed school due to toothache on 5 or more days		
	5 or mor	none-few	Total
0	92.88	88.10	92.84
	(0.38)	(3.74)	(0.38)
	8741.00	77.00	8818.00
2 or mor	7.12	11.90	7.16
	(0.38)	(3.74)	(0.38)
	653.00	9.00	662.00
Total	100.00	100.00	100.00
	9394.00	86.00	9480.00

```
  Key:  column percentage
        (linearized standard error of column percentage)
        number of observations

  Pearson:
    Uncorrected   chi2(1)         =    2.9437
    Design-based  F(1, 156)       =    2.5752     P = 0.1106
```

- Interpretation: In 2021, among 8th, 10th, and 12th graders combined who miss 5 or more days of school due to toothache (12%) were just as likely to have not seen a dentist in the past year as students who did not miss 5 or more days of school due to toothache. (7%).

Synthetic High School Estimates

The Centers for Disease Control and Prevention's Youth Risk Behavior Survey (YRBS) measures health behaviors of students in grades 9, 10, 11, and 12. They report "high school" estimates that combine all four grades. YRBS high school estimates are often used for setting bench marks, like the Healthy People 2020. In order to compare HYS results to national measures, we can create a synthetic high school estimate by following the steps for grade-adjusted weighting (described above) and applying an additional weight for the non-surveyed grades 9th and 11th.

According to the 2021-2022 OSPI enrollment data for the state, there are (available at: https://www.k12.wa.us/data-reporting/data-portal):

Grade	Enrolled	% High School
9th	87795	0.2526
10th	84871	0.2442
11th	83110	0.2391
12th	91763	0.2640
Total	347539	1.0000

To create a weight for each grade, we include the proportion that each grade contributes towards the high school enrollment and a proportion that takes into account how much the grade level should contribute to the overall estimate. For example, ½ of the 8th grade estimate and ½ of the 10th grade estimate should be used to create a 9th grade estimate.

Grade	Weight Formula	Reasoning
8th	0. 2526* 0.5 = **0.1263**	contributes to ½ of 9th grade
10th	0. 2526* 0.5 = 0.1263	contributes to ½ of 9th grade for the 10th grade
	0. 2442* 1 = 0.2442	
	0. 2391* 0.5 = 0.1196	contributes to ½ of 11th grade
	Add all 3: 0.1263 + 0.2442 + 0.1196 = **0.4901**	
12th	0. 2391* 0.5 = 0.1196	contributes to ½ of 11th grade for the 12th grade
	0. 2640* 1 = 0.2629	
	Add both: 0.1196 + 0.2629 = **0.3836**	

Total = (gr8*0.1263) + (gr10 *0.4901) + (gr12*0.3836)

The coding for generating 2021 synthetic high school estimates is:

```
gen hswt=.
replace hswt=0.1263*100 if grade==8
replace hswt=0.4901*100 if grade==10
replace hswt=0.3836*100 if grade==12
```

Or if you want STATA to do the math for you, you can use the following formula:

```
gen hswt=.
replace hswt=(347539/87795*.5) if grade==8
replace hswt=((347539/87795*.5)+( 347539/84871*1)+( 347539/83110*.5)) if grade==10
replace hswt=((347539/83110*.5)+( 347539/91763*1)) if grade==12
svyset [pweight=hswt], psu(schgrd)
svy:tab d21_16use grade, col se obs per format(%3.2f)
```

Used marijuana - Any use in the past 30 days	Grade 8	Grade 10	Grade 12	Total
no	97.16	92.79	84.08	90.90
	(0.31)	(0.60)	(1.37)	(0.69)
	6804.00	7807.00	4279.00	18890.00
yes	2.84	7.21	15.92	9.10
	(0.31)	(0.60)	(1.37)	(0.69)
	199.00	607.00	810.00	1616.00
Total	100.00	100.00	100.00	100.00
	7003.00	8414.00	5089.00	20506.00

```
  Key:  column percentage
        (linearized standard error of column percentage)
        number of observations

  Pearson:
    Uncorrected   chi2(2)         =  498.7252
    Design-based  F(1.31, 204.77) =   58.9912     P = 0.0000
```

Interpretation: The weighted synthetic high school estimate for current marijuana use is 9.1% ±1.4.

Data Book Combined Grade Estimates

The Washington State Health Care Authority – Division of Behavioral Health and Recovery (DBHR) creates Community Needs Assessment Data Books for Community Prevention and Wellness Initiative (CPWI) coalitions to help with prevention strategic planning. To produce estimates for CPWI coalitions with small populations, grade levels are combined to produce Grades 8 and 10 results and Grades 8-12 results. These results are weighted by the number of students enrolled and the number of valid HYS respondents in the grades. According to the 2021-2022 OSPI enrollment data for the state, there are:

Grade	Enrolled	Valid N
8th	85,819	57,978
9th	87,795	5,304
10th	84,871	50,288
11th	83,110	3,727
12th	91,763	33,634

Combined Grade 8 and 10 Estimates

To create a weight for each grade, we include the enrollment for grades 8 and 10 divided by the number of valid survey respondents for grades 8 and 10. The enrollment and validn for each community can be found in the Report List at: www.AskHYS.net/Past. This example uses the 2021 census dataset.

```
use "C:2021 census.dta"
gen weight810=.
replace weight810= 85819/57978 if grade==8
replace weight810=84871/50288 if grade==10
svyset [pweight=weight810]
gen grade810=grade
recode grade810 6=. 7=. 8=1 9=. 10=1 11=. 12=.
svy:tab d21_16use grade810, col se obs per format(%3.2f)
```

Used marijuana - Any use in the past 30 days	column	se	obs
no	94.77	0.07	92131.00
yes	5.23	0.07	4935.00
Total	100.00		97066.00

Key: column = column percentage
se = linearized standard error of column percentage
obs = number of observations

Interpretation: The weighted combined grade 8 and 10 estimate for current marijuana use is 5.2% ±0.1.

Combined Grade 8 through 12 Estimates

For communities that surveyed their extra grades. create a weight for each grade, we include the enrollment for grades 8, 9, 10, 11, and 12 divided by the number of valid survey respondents for grades 8, 9, 10, 11, and 12. The enrollment and valid n for each community can be found in the Report List at: www.AskHYS.net/Past.

For this example, we'll use Wahkiakum County.

Grade	Enrolled	Valid N
8th	33	28
9th	32	11
10th	46	36
11th	50	43
12th	46	26

```
keep if conum==35
gen weight812=.
replace weight812=28/33 if grade==8
replace weight812=11/32 if grade==9
replace weight812=36/46 if grade==10
replace weight812=43/50 if grade==11
replace weight812=26/46 if grade==12
svyset [pweight=weight812]
```

```
gen grade812=grade
recode grade812 6=. 7=. 8=1 9=1 10=1 11=1 12=1
svy:tab d21_16use grade812, col se obs per format(%3.2f)
```

```
Used marijuana - Any use in the past 30 days |      column          se         obs
---------------------------------------------+---------------------------------------
                                          no |       89.43        2.85      108.00
                                         yes |       10.57        2.85       13.00
                                             |
                                       Total |      100.00                   121.00
---------------------------------------------+---------------------------------------
  Key:  column    =  column percentage
        se        =  linearized standard error of column percentage
        obs       =  number of observations
```

The weighted combined grade 8, 9, 10, 11, and 12 estimate for current marijuana use is 10.6% ±5.6.

For communities that didn't survey their extra grades. create a weight for each grade, we include the enrollment for grades 8, 10, and 12 divided by the number of valid survey respondents for grades 8, 10, and 12:

For this example, we'll use Adams County.

Grade	Enrolled	Valid N
8th	438	354
10th	396	316
12th	367	211

```
keep if conum==1
gen weight81012=.
replace weight81012=438/354 if grade==8
replace weight81012=396/316 if grade==10
replace weight81012=367/211 if grade==12
gen grade81012=grade
recode grade81012 8=1 10=1 12=1 6=. 7=. 9=. 11=.
svyset, clear
svyset [pweight=weight810]
svy:tab d21_16use grade81012, col se obs per format(%3.2f)
```

```
Used marijuana - Any use in the past 30 days |      column          se         obs
---------------------------------------------+---------------------------------------
                                          no |       92.30        0.98      748.00
                                         yes |        7.70        0.98       58.00
                                             |
                                       Total |      100.00                   806.00
---------------------------------------------+---------------------------------------
  Key:  column    =  column percentage
        se        =  linearized standard error of column percentage
        obs       =  number of observations
```

Interpretation: The weighted combined grade 8, 9, 10, 11, and 12 estimate for current marijuana use is 7.7% ±1.9.

Adding Additional Data

This section describes how to add more data onto your HYS data. It includes information about how to use the merge command. Merge allows you to add additional data to your original data by joining a common variable.

NOTE: STATA defines your original data as the "master" data and the new data you are appending on as your "using" data.

Merging

Merging is used when the data you want to add has at least one variable in common with your original dataset, like school building number or county number.

For example, if you wanted to conduct analysis of the state sample data according to the four classifications for urban/rural and you had a dataset with that classification by school building number, you could add the classification to your HYS data with merge.

Data Preparation:

Keep your merge simple; don't include unnecessary variables. Sometimes both datasets have the same (duplicate) variables. Duplicates can confuse STATA and cause problems with your merge. Only keep the duplicate variables that you need to make a proper merge. If you want to keep other duplicate variables, rename them so they will be distinct variables in your new dataset.

You also need to make sure that the variable in your new data is in the same format as your HYS data. For example, the variable schgrd in the HYS dataset is numeric. If you are going to merge your new data with schgrd, you need to make sure that the schgrd variable in your new data is also numeric. If the schgrd variable in your new data is a string, change it to numeric using the encode command:

```
encode schgrd, gen(school)
drop schgrd
rename school schgrd
```

Sort Using Data:

Your "using data" should be the dataset that you are adding on to the HYS dataset. Prior to merging, you need to sort your new dataset by your merge variable(s).

```
sort schgrd
```

After sorting, save your dataset with a new name. This is now referred to as your "using dataset."

```
save "C:new using data.dta"
```

Sort Master Data:

Open your HYS dataset (your "master dataset") and sort it by your merge variable(s).

```
sort schgrd
```

Merge the Data:

Once your master data is sorted you can merge on your new dataset:

```
merge (schgrd) using "C:new using data.dta"
```

Merge Investigation:

While merging, STATA will try to tell you if something looks wrong; look for any error messages. Error messages are usually in red font. Messages in green font are usually just letting you know that there were commonalities between the two datasets, like the same variable labels.

After the merge you will have a new variable "_merge." You can use this variable to check the results of your merge. The response options provided for _merge are 1, 2 and 3:

```
tab _merge
```

1 = the using data did not have a match

2 = the master data did not have a match

3 = matched

Depending on the dataset you are adding, you may or may not be expecting all of the data to match. E.g., when we check our merge of the urban rural classifications to HYS we get mostly 3's (matches), but we get some 2's (non-matches in the master). This is OK because we know that the urban-rural classification data includes all schools in the state and our HYS data only includes schools that participated in the survey. We expect that the schools that did not participate should not match. So in this case we would simply get rid of the non-matched data by dropping them:

```
drop if _merge==2
```

If we get some 1s with this same merge we would need to do more investigation. We expect that every school in our HYS dataset should have an urban-rural classification. We can find out the names of the schools that didn't match by:

```
tab schname if _merge==1
```

Then we would want to check our urban rural classification to see if that school existed. If not, we would need to figure out why. For example, maybe the school is new or it changed its location in the past year, etc. If this is the case, you can update your urban rural dataset and then re-merge.

You can also look at your actual data to see what happened in your merge by looking in the "data browser" by click on the toolbar icon  or in the "data editor" by clicking on the toolbar icon  . In either the browser or the editor you can sort by the _merge variable to see exactly which data are not matching up. The data browser opens faster than the data editor, but the data editor allows you to add and delete data. When you exit the data editor it will ask you if you want to preserve your changes – only keep changes you are sure you want.

Once you're satisfied with your merge you can get rid of the "_merge" variable:
drop _merge

NOTE: You cannot merge on additional data until you drop the "_merge" variable or rename it.

For some reason, it usually takes most of us multiple attempts to get our merges correct. So don't worry if you it takes you a few tries, and always investigate your merge to make sure it did what you wanted it to.

Checking Findings and Significance Online

This section describes the information available on the AskHYS.net website to verify your analysis results. When running data analysis in STATA it's always a good idea to verify your results by looking at previously produced results.

This section also includes information about an online tool for testing statistical significance when you are comparing two estimates that have 95% confidence intervals.

AskHYS.net Website

AskHYS.net is the primary location of most HYS related information and results.

Address is: http://www.askhys.net

AskHYS Fact Sheets

Currently, topical fact sheets are available with results from 2008 through 2021. State, ESD, and County fact sheets are available to the public. District and Building fact sheets are available to those with permission from district superintendents (through an approval process with OSPI – see the Log On page for more information).

Fact sheets can also be produced by gender, but the general rules for crosstabs apply (at least 5 respondents in every cell for state fact sheets and 10 per cell for local).

- Current grade-level fact sheets include the following topics:
- Unintentional Injury
- Violent Behaviors & School Safety
- Harassment and Bullying
- Community Risk Factors
- Community Protective Factors
- School Risk Factors
- School Protective Factors
- Peer-Individual Risk Factors
- Family Protective Factors
- Weight and Obesity
- Dietary Behaviors
- Oral Health
- Physical Activity
- Mental Health and Well-being
- Hope
- WA HYS Adverse Childhood Experiences
- Sexual Behavior
- Current Substance Use
- Alcohol Use
- Commercial Tobacco Product Use
- Marijuana Use
- Migratory Students
- Polysubstance Use

Multiple-grade fact sheets are available for the following topics:

- Alcohol Use
- Marijuana Use
- School Safety
- Depressive Feelings & Suicide
- Prescription Medication Use

Most fact sheets also include a chart with topical results, trend data, comparisons to the state, and relationships between a topic and academic achievement, e.g., cigarette smoking and academic achievement. There is also an information factsheet on what Risk and Protective Factors are.

Instructional videos are available on the Fact Sheet webpage.

Q x Q Analysis

The Q x Q is an interactive data query system to analyze state and local frequencies and crosstabs. HYS data from 2002 through 2021 are available to analyze. State, ESD, and County data can be accessed by all. District and Building data are available to those with permission from district superintendents (through an approval process with OSPI – see the Log On page for more information).

When running a crosstab, you need to think about how you want your results to turn out before you select your variables. The variable that you drop into the first box will be the group you are interested in finding out more information about. The second variable you select is your outcome variable. For example:

1. Do you want to know the prevalence of marijuana use among different race groups? Then select Demographics – Race/Ethnicity –[G06] Race/Ethnicity as your first variable and Marijuana – Current Use – [D21_16] Current Marijuana Use as your second variable.
2. Do you want to know the prevalence of drinking alcohol among youth who use marijuana? Then select Marijuana – Current Use – [D21_16] Current Marijuana Use as your first variable and Alcohol – Current Use – [D20] Current Alcohol Drinking as your second.
3. Or do you what to know the flip side, what is the prevalence of marijuana use among youth who drink alcohol? Then select Alcohol first and Marijuana second.

Crosstabs on the Q x Q also have to follow the requirements for a minimum number of respondents per cell in order to produce results:

- For state level analysis you must have 5 or more respondents in each cell.
- For sub-state level analysis, you must have 10 or more respondents in each cell.

Example 1: Statewide HYS 2021 - Grade 10 Race/Ethnicity and Current Marijuana Use

Selected:

Output:

Race/Ethnicity	Current Marijuana Use: no days	any days	Total
White non-Hispanic	92.0% ± 1.6% 3,770	8.0% ± 1.6% 330	100.0% 4,100
Hispanic	92.9% ± 1.4% 1,755	7.1% ± 1.4% 135	100.0% 1,890
American Indian or Alaskan Native non-Hispanic	88.4% ± 5.7% 99	11.6% ± 5.7% 13	100.0% 112
Asian or Asian American non-Hispanic	98.8% ± 0.7% 818	1.2% ± 0.7% 10	100.0% 828
Black or African-American non-Hispanic	93.2% ± 2.8% 262	6.8% ± 2.8% 19	100.0% 281
Native Hawaiian or other Pacific Islander non-Hispanic	93.9% ± 5.3% 77	6.1% ± 5.3% 5	100.0% 82
Other non-Hispanic	95.4% ± 2.6% 208	4.6% ± 2.6% 10	100.0% 218
Multiracial non-Hispanic	90.6% ± 2.7% 723	9.4% ± 2.7% 75	100.0% 798

Interpretation: When reviewing your results, you should read them by each row. Notice that the "Total" for each row is 100%. Statewide in 2021, current marijuana use by 10th grade race/ethnicity groups was:

White non-Hispanic 8.0%
Hispanic 7.1%
American Indian or Alaskan Native non-Hispanic 11.6%
Asian or Asian American non-Hispanic 1.2%
Black or African-American non-Hispanic 6.8%
Native Hawaiian or other Pacific Islander non-Hispanic 6.1%
Other non-Hispanic 4.6%
Multiracial non-Hispanic 9.4%

Example 2: Statewide HYS 2021 - Grade 10 Current Marijuana Use and Current Alcohol Drinking

Selected:

Output:

		Current Alcohol Drinking		
		no days	any days	Total
Current Marijuana Use	no days	95.1% ± 1.1% 7,397	4.9% ± 1.1% 383	100.0% 7,780
	any days	47.1% ± 7.4% 281	52.9% ± 7.4% 316	100.0% 597

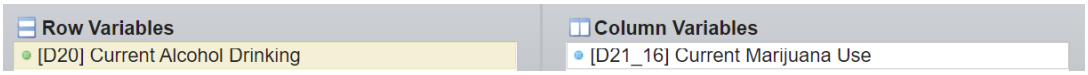
Interpretation: Statewide in 2021, 10th grade current alcohol drinking was:

4.9% among those who didn't use marijuana on any day in the past 30 days

52.9% among current marijuana users

Example 3: Statewide HYS 2021 - Grade 10 Current Alcohol Drinking and Current Marijuana Use

Selected:



Output:

		Current Marijuana Use		
		no days	any days	Total
Current Alcohol Drinking	no days	96.3% ± 0.8% 7,397	3.7% ± 0.8% 281	100.0% 7,678
	any days	54.8% ± 5.0% 383	45.2% ± 5.0% 316	100.0% 699

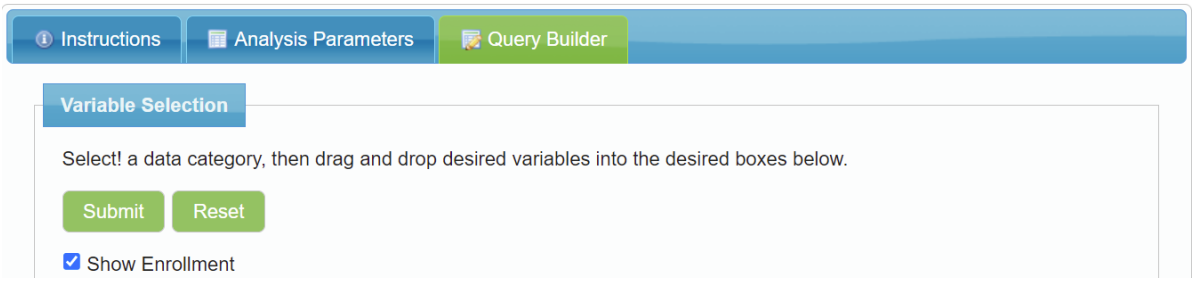
Interpretation: Statewide in 2020, 10th grade current marijuana use was:

3.7% among those who didn't drink alcohol on any day in the past 30 days

45.2% among current alcohol drinkers

Number of Students Represented by Results

The QxQ includes an option to include the number of students who are represented in your results. Click the Show Enrollment box near the top of the Query Builder tab. The number of students represented by a survey result is calculated by multiplying the survey results percentage by the number students enrolled at HYS-eligible schools.



Example 4: Statewide HYS 2021 - Grade 10 Current Alcohol Drinking and Current Marijuana Use with number of students represented.

Output:

		Current Marijuana Use		
		no days	any days	Total
Current Alcohol Drinking	no days	96.3% ± 0.8% 7,397 *78,791*	3.7% ± 0.8% 281 *2,993*	100.0% 7,678 *81,784*
	any days	54.8% ± 5.0% 383 *44,812*	45.2% ± 5.0% 316 *36,972*	100.0% 699 *81,784*

Interpretation: Statewide, there were 81,784 10th-grade students attending HYS-eligible schools. So statewide, among 10th graders, there are about 40,000 current alcohol drinkers who also used marijuana in the past month (45.2% * 81,784 = 36,966).

When extrapolating a survey percentage to the number of students represented, always include the word “about” or “approximately” and round your number to account the uncertainty in the survey estimate.

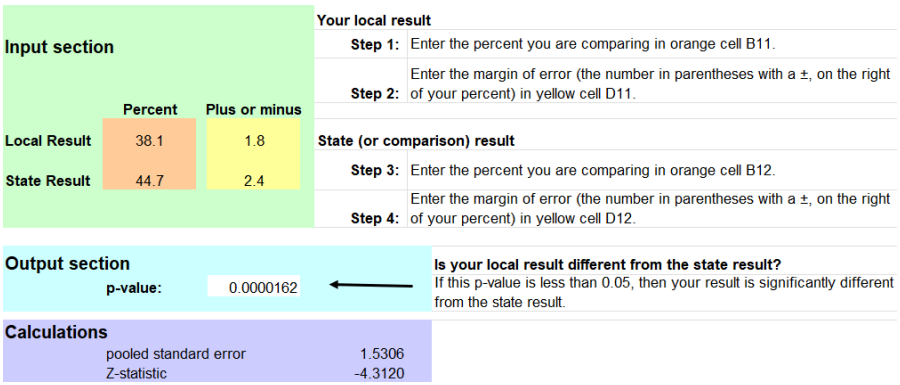
Instructional videos are available on the Q x Q webpage.

Online Tool for Determining Statistical Significance

Address is: http://www.askhys.net/Training

There is an "Excel Tool for Determining Statistical Significance" on the HYS administration website reporting page. Scroll down to Testing for Significant Differences. The tool itself has cells to enter local and state data, but you can use this tool to test the difference between any two estimates with 95% confidence intervals.

1. For example, to test for differences in experiencing depressive feelings between 10th and 12th graders in 2021 statewide using the following results:
    - 10th grade: 38.1% (±1.8)
    - 12th grade: 44.7% (±2.4)



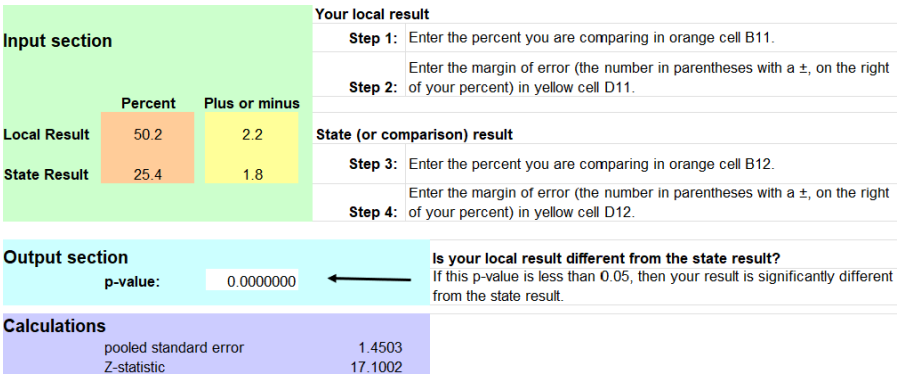
Input section			Your local result	
	Percent	Plus or minus	Step 1:	Enter the percent you are comparing in orange cell B11.
Local Result	38.1	1.8	Step 2:	Enter the margin of error (the number in parentheses with a ±, on the right of your percent) in yellow cell D11.
State Result	44.7	2.4	State (or comparison) result	
			Step 3:	Enter the percent you are comparing in orange cell B12.
			Step 4:	Enter the margin of error (the number in parentheses with a ±, on the right of your percent) in yellow cell D12.

Output section			
p-value:	0.0000162	←	Is your local result different from the state result? If this p-value is less than 0.05, then your result is significantly different from the state result.

Calculations	
pooled standard error	1.5306
Z-statistic	-4.3120

Interpretation: P-value is 0.000, which is less than 0.05, so 12th graders are more likely than 10th graders to experience depressive feelings, in 2021 statewide.

2. Test for differences in experiencing depressive feelings between 10th grade males and females in 2014, statewide using the following results:
    - 10th grade females: 50.2% (±2.2)
    - 10th grade males: 25.4% (±1.8)



Input section			Your local result	
	Percent	Plus or minus	Step 1:	Enter the percent you are comparing in orange cell B11.
Local Result	50.2	2.2	Step 2:	Enter the margin of error (the number in parentheses with a ±, on the right of your percent) in yellow cell D11.
State Result	25.4	1.8	State (or comparison) result	
			Step 3:	Enter the percent you are comparing in orange cell B12.
			Step 4:	Enter the margin of error (the number in parentheses with a ±, on the right of your percent) in yellow cell D12.

Output section			
p-value:	0.0000000	←	Is your local result different from the state result? If this p-value is less than 0.05, then your result is significantly different from the state result.

Calculations	
pooled standard error	1.4503
Z-statistic	17.1002

Interpretation: P-value is 0.0000, which is less than 0.05, so 10th grade females are more likely to experience depressive feelings compared to 10th grade males, in 2021 statewide.

Displaying Results

This section provides some tools to help you display the results of your STATA analysis.

Tables and charts are a common way to present analysis results in a useful and visually-appealing way. STATA provides a number of graphic options that you can use to display your results. Start by selecting Graphics from the drop down menu. The first option on the drop down, Easy Graphs, has some simple graphs such as line graphs, bar charts, and histograms.

It requires some practice to produce meaningful graphs in STATA, or many lines of code. A "do file" is provided in the following Appendix:

- Appendix C: Making Bar Graphs with Error Bars in STATA

This "do file" walks through a number of different commands for creating bar charts, including how to add confidence intervals to your charts. The example used is perception of great risk from regular marijuana use by race and age.

Often it is easiest to copy and paste output results into a more familiar program such as Excel and convert it into tables and charts. Charts can also be produced in Excel using pasted STATA output. Again, the trick is formatting your output so that it can easily be incorporated into a chart using such formatting options as se, ci and vert.

Producing Graphs in STATA

Newer versions of STATA provide a variety of graphing options. Try to experiment with the drop down Graphics menu on the tool bar to create graphics.

NOTE: Type "help graph" in STATA to find more instructions about graphics. There are also a number of helpful STATA graphics books and websites.

The following example takes you through the steps to create a graph of two variables with confidence intervals of current marijuana use by race and grade. It was modified from an example on a UCLA STATA website and found at: http://www.ats.ucla.edu/stat/STATA/faq/barcap.htm

This example does not attempt to thoroughly explain all of the steps involved in the graphic process, but to provide you with some sample commands that you can experiment with. A "do file" is provided in the following Appendix:

- Appendix C: Making Bar Graphs with Error Bars in STATA

Setting up

```
use "C:\HYS State Sample.dta"
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

Creating a recoded race variable

```
gen race=g06
recode race 1=1 2=2 3=3 4=4 5=1 6=5 7=. 8=.
```

lab def newrace 1"API" 2"Indian" 3"Black" 4"Hispanic" 5"White" lab val race newrace

Recode your outcome variable to be 0,1, so you get the correct mean.

Creating a collapsed smoking mean by grade and race

```
collapse (mean) mean_d21_16use= d21_16use (sd) sd_d21_16use= d21_16use (count)
n= d21_16use, by(grade race)
```

Creating the high and low confidence interval values

```
generate hi_d21_16use = mean_d21_16use + invttail(n-1,0.025)*(sd_d21_16use/sqrt(n))
generate lo_d21_16use = mean_d21_16use - invttail(n-1,0.025)*(sd_d21_16use/sqrt(n))
```

Graphing

Creating a simple two-way bar graph

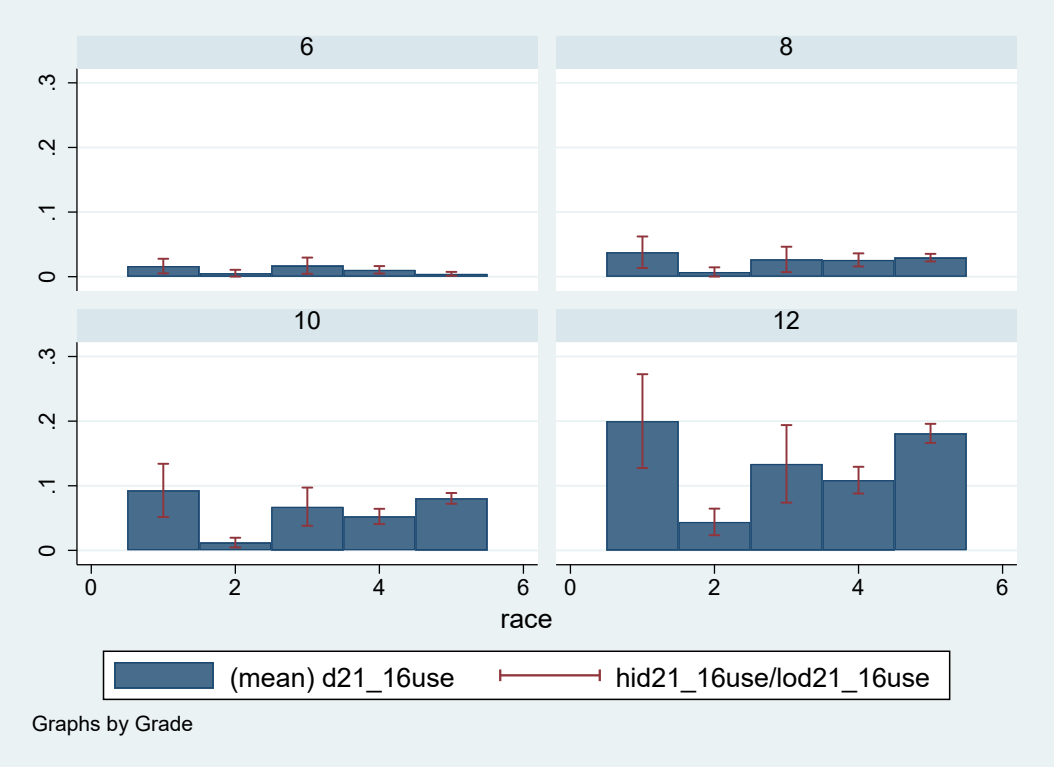
```
graph bar mean_d21_16use, over(race) by(grade)
```

Adding some color

```
graph bar mean_d21_16use, over(race) by(grade) asyvars
```

Adding confidence intervals error bars

```
graph twoway (bar mean_d21_16use race) (rcap hi_d21_16use lo_d21_16use race), by(grade)
```



Changing the graph to be set up by single variables for each race and grade and creating a graph with confidence intervals

```
gen graderace=race if grade==6
replace graderace=race+10 if grade==8
replace graderace=race+20 if grade==10
```

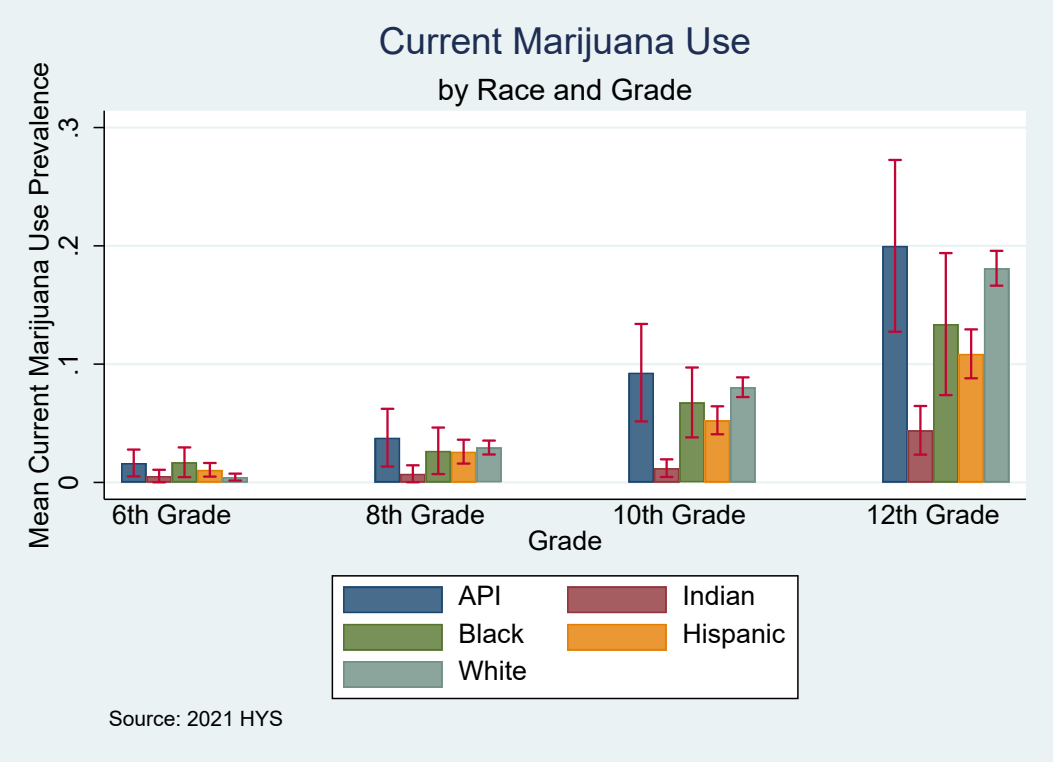
```
replace graderace=race+30 if grade==12
sort graderace
list graderace grade race, sepby(grade)
twoway(bar mean_d21_16use graderace)(rcap hi_d21_16use lo_d21_16use graderace)
```

Adding in more color

```
twoway (bar mean_d21_16use graderace if race==1) ///
(bar mean_d21_16use graderace if race==2) ///
(bar mean_d21_16use graderace if race==3) ///
(bar mean_d21_16use graderace if race==4) ///
(bar mean_d21_16use graderace if race==5) ///
(rcap hi_d21_16use lo_d21_16use graderace)
```

Adding in a legend and labels

```
twoway (bar mean_d21_16use graderace if race==1) ///
(bar mean_d21_16use graderace if race==2) ///
(bar mean_d21_16use graderace if race==3) ///
(bar mean_d21_16use graderace if race==4) ///
(bar mean_d21_16use graderace if race==5) ///
(rcap hi_d21_16use lo_d21_16use graderace), ///
legend(order(1 "API" 2 "Indian" 3 "Black" 4 "Hispanic" 5 "White") ) ///
xlabel(2.5 "6th Grade" 12.5 "8th Grade" 22.5 "10th Grade" 32.5"12th Grade",
noticks) ///
        xtitle(Grade) ytitle(Mean Current Marijuana Use Prevalence) ///
        title(Current Marijuana Use) subtitle(by Race and Grade) note(Source: 2021 HYS)
```



Web Resources

Here are a few helpful resources on the Healthy Youth Survey, STATA, and statistical analysis. The links provided here do not in any way imply that the sources are endorsed by the state agencies involved in HYS. They are just some sites that we have found to be helpful.

Healthy Youth Survey

- AskHYS: http://www.askhys.net
- Office of Superintendent of Public Instruction's HYS webpage: https://www.k12.wa.us/student-success/health-safety/healthy-youth-survey
- Office of Superintendent of Public Instruction's data portal for enrollment data and other school-related data: https://www.k12.wa.us/data-reporting/data-portal

STATA Resources

- STATA: http://www.stata.com
- UCLA: https://stats.idre.ucla.edu/stata/
- Princeton: https://data.princeton.edu/stata and http://www.princeton.edu/~otorres/Stata
- Harvard: https://sociology.fas.harvard.edu/need-help-basic-stata
- Tufts: https://sites.tufts.edu/datalab/learning-statistics/stats-online-tutorials/stata-resources/stata-website/

Statistical Analysis

- JoinPoint regression program: https://surveillance.cancer.gov/joinpoint/
- Allows you to convert data files into STATA datasets. Available free trial at: http://www.stattransfer.com

Appendices

The following appendices are STATA do files that let you replicate many of the analyses presented in this manual.

- Appendix A: State Level Enrollments by Year and Coding for Synthetic High School Weights
- Appendix B: HYS State Data Analysis Examples in STATA
- Appendix C: Quick Examples of HYS Data Analysis in STATA
- Appendix D: Making Bar Graphs with Error Bars in STATA

Appendix A: County-level Analysis Coding by Year

Use the following code to drop any counties with less than 40% response rate or if they don't have enough respondents or districts participating.

For 2021, the following counties and grades should be dropped:

```
drop if conum==1 & grade==6
drop if conum==2 & (grade==10 | grade==12)
drop if conum==4 & grade==12
drop if conum==8 & grade==12
drop if conum==10 & (grade==10 | grade==12)
drop if conum==11
drop if conum==12 & grade==12
drop if conum==19 & grade==12
drop if conum==21 & grade==12
drop if conum==23 & (grade==6 | grade==12)
drop if conum==24 & grade==12
drop if conum==26 & grade==10
drop if conum==27 & grade==12
drop if conum==32 & grade==12
drop if conum==33 & (grade==10 | grade==12)
drop if conum==34 & grade==12
drop if conum==35
drop if conum==36 & grade==6
drop if conum==38 & grade==6
```

For 2018, the following counties and grades should be dropped:

```
drop if conum==1 & (grade==6 | grade==12)
drop if conum==10 & grade==8
drop if conum==11
drop if conum==19 & grade==12
drop if conum==33 & grade==6
drop if conum==36 & (grade==10 | grade==12)
```

For 2018, the following counties and grades should be dropped:

```
drop if conum==1 & (grade==6 | grade==12)
drop if conum==10 & grade==8
drop if conum==11
drop if conum==19 & grade==12
drop if conum==33 & grade==6
drop if conum==36 & (grade==10 | grade==12)
```

For 2016, the following counties and grades should be dropped:

```
drop if conum==2 & grade==6
drop if conum==10 & (grade==6 | grade==8)
drop if conum==11
drop if conum==30 & grade==12
drop if conum==32 & grade==12
```

```
drop if conum==37 & grade==12
drop if conum==38 & grade==12
```

For 2014, the following counties and grades should be dropped:

```
drop if conum==1
drop if conum==5 & (grade==6 | grade==12)
drop if conum==10 & (grade==6 | grade==12)
drop if conum==11
drop if conum==16 & (grade==6 | grade==8)
drop if conum==19 & (grade==6 | grade==8)
drop if conum==20 & (grade==6 | grade==10 | grade==12)
drop if conum==21 & grade==12
drop if conum==26 & (grade==6 | grade==12)
drop if conum==28 & grade==8
drop if conum==30 & grade==12
```

For 2012, the following counties and grades should be dropped:

```
drop if conum==1
drop if conum==5 & (grade==8 | grade==12)
drop if conum==10 & grade==8
drop if conum==11
drop if conum==19 & (grade==6 | grade==8)
drop if conum==26 & grade==12
drop if conum==33 & (grade==6 | grade==10 | grade==12)
drop if conum==34 & grade==6
drop if conum==36 & (grade==10 | grade==12)
```

For 2010, the following counties and grades should be dropped:

```
drop if conum==5 & (grade==6 | grade==10 | grade==12)
drop if conum==11
drop if conum==16 & grade==12
drop if conum==30 & (grade==6 | grade==8)
drop if conum==33 & (grade==6 | grade==10 | grade==12)
```

For 2008, the following counties and grades should be dropped:

```
drop if conum==5 & (grade==10 | grade==12)
drop if conum==10 & grade==12
drop if conum==30 & (grade==6 | grade==8)
drop if conum==33 & grade==12
```

For 2006, the following counties and grades should be dropped:

```
drop if conum==3 & grade==12
drop if conum==5 & (grade==8 | grade==10 | grade==12)
drop if conum==10 & (grade==6 | grade==8 | grade==10)
drop if conum==12 & grade==6
drop if conum==13 & (grade==10 | grade==12)
drop if conum==14 & (grade==6 | grade==12)
drop if conum==16 & grade==12
drop if conum==23 & grade==6
drop if conum==28 & grade==8
```

```
drop if conum==30 & (grade==6 | grade==8)
drop if conum==33 & (grade==10 | grade==12)
```

For 2004, the following counties and grades should be dropped:

```
drop if conum==2 & grade==12
drop if conum==3 & (grade==8 | grade==12)
drop if conum==5 & (grade==6 | grade==10 | grade==12)
drop if conum==7
drop if conum==10 & (grade==6 | grade==12)
drop if conum==11 & (grade==10 | grade==12)
drop if conum==12
drop if conum==14 & grade==6
drop if conum==18 & grade==12
drop if conum==22
drop if conum==30 & (grade==10 | grade==12)
drop if conum==32 & grade==12
drop if conum==33 & (grade==10 | grade==12)
drop if conum==35
```

For 2002, the following counties and grades should be dropped:

```
drop if conum==2 & grade==12
drop if conum==3 & (grade==10 | grade==12)
drop if conum==5
drop if conum==6 & (grade==6 | grade==10 | grade==12)
drop if conum==7
drop if conum==11 & (grade==6 | grade==10 | grade==12)
drop if conum==12
drop if conum==14 & (grade==6 | grade==10 | grade==12)
drop if conum==16 & grade==12
drop if conum==17 & (grade==8 | grade==10 | grade==12)
drop if conum==18
drop if conum==22
drop if conum==23 & (grade==10 | grade==12)
drop if conum==24 & grade==10
drop if conum==26
drop if conum==30
drop if conum==32 & (grade==10 | grade==12)
drop if conum==33 & (grade==6 | grade==10 | grade==12)
drop if conum==35
drop if conum==36 & grade==8
drop if conum==37 & grade==12
```

Appendix B: State Level Enrollments by Year and Coding for Synthetic High School Weights

For more information on calculating synthetic high school estimates, see Combining Grade Levels.

2021-2022 State Enrollment

Grade	Enrolled	% High School
9th	87795	0.2526
10th	84871	0.2442
11th	83110	0.2391
12th	91763	0.2640
Total	347539	1.0000

2021 weight coding:

```
gen hswt=.
replace hswt=(347539/87795*.5) if grade==8
replace hswt=((347539/87795*.5)+(342713/87795*1)+( 342713/83110*.5)) if grade==10
replace hswt=((347539/83110*.5)+(342713/91763*1)) if grade==12
```

2018-2019 State Enrollment

Grade	Enrolled	% High School
9th	84,224	0.2461
10th	83,450	0.2438
11th	84,612	0.2472
12th	89,963	0.2629
Total	342,713	1.0000

2018 weight coding:

```
gen hswt=.
replace hswt=(342713/84224*.5) if grade==8
replace hswt=((342713/84224*.5)+(342713/83450*1)+( 342713/84612*.5)) if grade==10
replace hswt=((342713/8461*.5)+(342713/89963*1)) if grade==12
```

2016-2017 State Enrollment

Grade	Enrolled	% High School
9th	82,113	0.2418
10th	83,687	0.2464
11th	83,320	0.2453
12th	90,522	0.2665
Total	339,642	1.0000

2016 weight coding:

```
gen hswt=.
replace hswt=(339642/82113*.5) if grade==8
replace hswt=((339642/82113*.5)+( 339642/83687*1)+( 339642/83320*.5)) if grade==10
replace hswt=((339642/83320*.5)+( 339642/90522*1)) if grade==12
```

2014-2015 State Enrollment

Grade	Enrolled	% High School
9th	83,277	0.2499
10th	82,136	0.2465
11th	81,040	0.2432
12th	86,821	0.2605
Total	333,274	1.0000

2014 weight coding:

```
gen hswt=.
replace hswt=(83277/333274*100*.5) if grade==8
replace hswt=((83277/333274*100*.5)+(82136/333274*100*1)+(81040/333274*100*.5)) if grade==10
replace hswt=((81040/333274*100*.5)+(86821/333274*100*1)) if grade==12
```

2012-2013 State Enrollment

Grade	Enrolled	% High School
9th	82,921	0.2535
10th	81,141	0.2480
11th	80,702	0.2467
12th	82,397	0.2519
Total	327,161	1.0000

2012 weight coding:

```
gen hswt=.
replace hswt=(82921/327161*100*.5) if grade==8
replace hswt=((82921/327161*100*.5)+(81141/327161*100*1)+(80702/327161*100*.5)) if grade==10
replace hswt=((80702/327161*100*.5)+(82397/327161*100*1)) if grade==12
```

2010-2011 State Enrollment

Grade	Enrolled	% High School
9th	84,113	0.2551
10th	81,966	0.2486
11th	79,874	0.2422
12th	83,818	0.2542
Total	329,771	1.0000

2010 weight coding:

```
gen hswt=.
replace hswt=(84113/329771*100*.5) if grade==8
replace hswt=((84113/329771*100*.5)+(81966/329771*100*1)+(79874/329771*100*.5)) if grade==10
replace hswt=((79874/329771*100*.5)+(83818/329771*100*1)) if grade==12
```

2008-2009 State Enrollment

Grade	Enrolled	% High School
9th	87,638	0.2635
10th	83,359	0.2506
11th	81,601	0.2453
12th	80,013	0.2406
Total	332,611	1.0000

2008 weight coding:

```
gen hswt=.
replace hswt=(87638/332611*100*.5) if grade==8
replace hswt=((87638/332611*100*.5)+(83359/332611*100*1)+(81601/332611*100*.5)) if grade==10
replace hswt=((81601/332611*100*.5)+(80013/332611*100*1)) if grade==12
```

2006-2007 State Enrollment

Grade	Enrolled	% High School
9th	90,444	0.2721
10th	84,476	0.2542
11th	80,193	0.2413
12th	77,242	0.2324
Total	332,355	1.0000

2006 weight coding:

```
gen hswt=.
replace hswt=(90444/332355*100*.5) if grade==8
replace hswt=((90444/332355*100*.5)+(84476/332355*100*1)+(80193/332355*100*.5)) if grade==10
replace hswt=((80193/332355*100*.5)+(77242/332355*100*1)) if grade==12
```

2004-2005 State Enrollment

Grade	Enrolled	% High School
9th	89,970	0.2769
10th	83,315	0.2564
11th	77,443	0.2383
12th	74,248	0.2285
Total	324,976	1.0000

2004 weight coding:

```
gen hswt=.
replace hswt=(89970/324976*100*.5) if grade==8
replace hswt=((89970/324976*100*.5)+(80877/324976*100*1)+(77443/324976*100*.5)) if grade==10
replace hswt=((77443/324976*100*.5)+(74248/324976*100*1)) if grade==12
```

2002-2003 State Enrollment

Grade	Enrolled	% High School
9th	87,842	0.2763
10th	80,877	0.2544
11th	76,759	0.2415
12th	72,404	0.2278
Total	317,882	1.0000

2002 weight coding:

```
gen hswt=.
replace hswt=(87842/317882*100*.5) if grade==8
replace hswt=((87842/317882*100*.5)+(83315/317882*100*1)+(76759/317882*100*.5)) if grade==10
replace hswt=((76759/317882*100*.5)+(72404/317882*100*1)) if grade==12
```

Appendix C: Do File ~ HYS State Data Analysis Examples in STATA

*For use with State Sample data

*The following "do file" runs through examples see the HYS Data Analysis in STATA section

*To run a line of command highlight the command text and hit the icon above that looks like a page with text on it

*Instructions for this file are preceded by an asterisk, they are just informational. Actual STATA commands are indented and don't have an asterisk

*The commands and instructions presented here are suggestions and only one method in which STATA can be used to analyze survey data

*This section covers the following topics: Opening your dataset Analysis by Grade, Frequencies and summaries of statistics, Creating new variables, Labeling new variables, General set up for survey analysis, Two-way tables and crosstabs, More options for using "svy", Additional tips for formatting, Analysis by grade, Stratified analysis and subpopulations

```
*============================================================
```

*Open your 2021 State Sample dataset and Setup for Survey Analysis

```
*============================================================
```

*start your do file with the clear command to get rid on any previous data or add clear to the end of your use command

```
clear
*use "hys21 state dataset.dta"

*use "hys21 state dataset.dta", clear
```

*Put in the pathway to **your** dataset or you can also open your data file by using the File drop down menu

```
*============================================================
*General Setup for Survey Analysis - state sample
*============================================================
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
keep if staterec==1

*============================================================
*Frequencies and Summaries of Statistics
*============================================================
```

*you can look at basic frequencies using the tab command

```
tab d20
```

*for variable exploration use the summarize commands

```
summarize d20
summarize d20, detail
```

*use a histogram to look at the distribution of your variable

```
histogram d20
```

*you can also explore combinations of variables by demographics

```
tab d20use grade
```

*look at your variables graphically with a histogram by grade

```
histogram d20, by(grade)
```

*==

*Creating and Recoding New Variables

*==

*GENERATING new Variables

*you can create a new variable that has the same value as an original variable this can be useful if you plan to modify the variable in any way, so you still have the original in tact

```
gen alc30=d20
tab d20
tab alc30
```

*notice that they have exactly the same output

*you can generate combined variables of one or more original variables

```
gen cigchew30 = d14use + d15use
tab d14use
tab d15use
tab cigchew30
```

*notice that there are more response options (2=yes to both cigarettes and chew, 1=yes to one but not both, 0=no to both)

*you can create variables with no respondents, only missing values

```
gen new=.
tab new
```

*you can also create new dummy variables for each response option from an original variable

```
tab grade, gen(gradecat)
tab grade gradecat1
tab grade gradecat2
```

*notice that gradecat1 are the respondents from grade 6, gradecat2 are the respondents from grade 8 and notice that you have new variables at the bottom of your variable list

*All of these generated variables come in handy when trying to recode your data

*RECODING

*Recode the original current smoking variable to see if you get the same results as the pre-collapsed variable (d14use)

*Codebook your new cig30 variable to see the response options before recoding

```
codebook d20
gen alcohol30=d20
recode alcohol30 1=0 2=1 3=1 4=1 5=1
tab alcohol30 grade
```

*here's another way to recode

```
gen alcthirty=d20
recode alcthirty 1=0 2/6=1
tab alcthirty alc30
tab alcthirty grade
```

*in this case you can also check your recode with a pre-collapsed variable

```
tab d20use alc30
```

*REPLACING

*For more complex coding you will need to use the replace command

*In this example we will combine the variable for visiting a doctor (h24) with visiting a dentist (h25) to create an any visit variable

*Always a good idea to codebook your variables first

```
codebook h24 h25
```

*Create the new combined variable by designating with location of the response options from the original variables

```
gen visitboth=.
replace visitboth=1 if (h24==1 & h25==1)
replace visitboth=0 if (h24==2 | h24==3 | h24==4 | h25==2 | h25==3 | h25==4)
tab visitboth grade
```

*If you only wanted to include respondents who answered both questions you need one more line of command to set those who only answered one to missing

```
replace visitboth=. if (h24==. & h25==.)
tab visitboth grade
```

*==

*Labeling

*==

*Labeling newly created variables helps to keep response options clear

```
*to label a variable with a description:
        lab var visitboth "visited both a doctor and a dentist in the past year"

*to label response options you have two steps, first you have to create a label and then you
have to attach it
        lab def visit 1"both" 0"one or none"
        lab val visitboth visit

*run a tab to see if the labels were applied
        tab visitboth

*============================================================

*Two-Way Tables or Crosstabs

*============================================================

*SETUP

*Before you can run actual survey analysis, you need to provide STATA with setup commands to
account for weighting, primary sampling units and strata

*For these examples we're using state sample data, so we will set up STATA for that type of
analysis.

*If you are running a different type of analysis, for example county, then see the setup
commands under the section General Setup for Survey Analysis or see the examples in Appendix
B: Quick Examples of HYS Data Analysis in STATA
        gen fakewt=1
        svyset [pweight=fakewt], psu(schgrd)
        keep if staterec==1

*SURVEY ANALYSIS

*svy:tab allows you to cross two variables this simple tab splits up the data into four cells with
the totals of the cells = 100%

*the tab will also give you the results of a chi-squared test to let you know if one of the cells is
different from the others
        svy:tab h53 g05_18

*============================================================

*Additional Options with "Svy"

*============================================================

*COLUMN AND ROW PERCENTAGES

*Use "col" and "row" to get a cross tab with column or row percents
        svy:tab h53 g05_18, col
        svy:tab h53 g05_18, row
```

*notice how row and col produce different point estimates

*col gives you the prevalence of depressive feelings for females and males

*row tells you among those with depressive feelings, what proportion are female and what proportion are male

*OBSERVATIONS

*Obs - you can also add the obs command to get the number of observations used to calculate each point estimate

```
svy:tab h53 g05_18, col obs
```

*STANDARD ERROR and CONFIDENCE INTERVALS

*Use "se" and "ci" to add confidence intervals and standard errors to your output

*for standard error (to get symmetrical confidence intervals multiply by 1.96)

```
svy:tab h53 g05_18, col se
```

*for asymmetrical confidence intervals at the 95% confidence level, (95% is the default, you can change it with formatting)

```
svy:tab h53 g05_18, col ci
```

*PERCENTAGES

*Use "per" to display your estimate as percentage points

```
svy:tab h53 g05_18, col per
```

*you can add as many of these commands as you need

```
svy:tab h53 g05_18, col se ci obs per
```

*WIDENING TABLE COLUMNS

*You can create output with columns wide enough to display your response option labels and estimates

*stubwidth changes the width of response labels, cellwidth changes the width for the estimates

```
svy:tab s01 g05_18, row ci stubwidth (20) cellwidth (15)
```

*compare your results without designating the column widths

```
svy:tab s01 g05_18, row ci per
```

*STATA displays your estimates by 2 decimal points, so usually you only need to include the stubwidth command, not the cellwidth

```
svy:tab d20 grade, col ci stubwidth (15)
```

*ROUNDING

*to modify the number of decimal places in the output use the format command

```
svy:tab grade g05_18, per row ci format(%3.2f)
svy:tab grade g05_18, per row ci format(%9.3f)
```

*notice the difference changing the number after the decimal point makes .3 gives 3 decimal points and .0 rounds to the whole number

*REMOVING SCIENTIFIC NOTATION

*sometimes making the formatting number bigger can help if your observations are coming out in scientific notation

```
svy:tab grade g05_18, row per obs
svy:tab grade g05_18, row per obs format(%9.3f)
```

*VERTICAL ALIGNMENT

*to display upper and lower bound confidence intervals in a vertical fashion without the bracket and comma use the vert option

*this can be handy if you are pasting results into an excel table

```
svy:tab grade g05_18, row ci per vert
```

*==

*Stratified Analysis and Subpopulations

*==

*STATA provides a number of ways to create and run stratified analysis. Below are a few ways to generate subpop variables to use in analysis. The important thing is they need to be coded as 1, 0

*Also remember that if you drop something from your dataset that you cannot get it back unless you reopen your dataset

*Proceed with caution!

```
*use "hys21 state dataset.dta"
```

*removes students from all other grades, keeps only 8th grade

```
keep if grade==8
```

*keeps only current alcohol drinkers

```
keep if d20use==1
```

*creates a subpop of only Black-African American students

```
gen black=1 if g06==3
recode black 1=0 2=0 3=1 4=1 5=1 6=1 7=1 8=1
```

*creates a subpop of only 8th graders

```
tab grade, missing
gen eight=1 if grade==8
```

*creates a subpop of only 8th grade Black-African American students

```
gen black8=1 if g06==3
recode black8 1=0 2=0 3=1 4=1 5=1 6=1 7=1 8=1
```

*DUMMY VARIABLES

*!!!If you ran the drop commands, you will need to reopen and re-set up STATA for your analysis!!!

```
*use "hys21 state dataset.dta"
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

*The best way to create subpops is to make dummy variables. This creates four new dummy variables gradecat1 (for 6th grade), gradecat2 (for 8th grade), gradecat3 (for 10th grade) and gradecat4 (for 12th grade)

```
tab grade, gen(gradecat)
```

*Then use to cross current marijuana use by household marijuana use 8th graders, first looking at among current marijuana users/non-marijuana users, what proportion live with a marijuana user?

```
svy:tab d21_16use d99, subpop(gradecat2) row per
```

*Then among 8th graders who live/or don't live with a marijuana user, what proportion use marijuana

```
svy:tab d99 d21_16use, subpop(gradecat2) row per
```

*USING OVER

*You can also use the over command to run stratified analysis

```
recode d21_16use 1=1 2=0
svy:mean d21_16use, over(grade g05_18)
```

*_subpop_1 represents 6th grade females, so current marijuana use for 6th grade females is 1.0%. Current smoking for 8th grade males is 0.9%.

Appendix D: Do File ~ Quick Examples of HYS Data Analysis in STATA

*The following "do file" runs through examples see the Data Analysis - Quick Examples section

*Select the proper set up according to your data, to replicate the results in the manual – use the 2021 state sample dataset

*To run a line of command highlight the command text and hit the icon above that looks like a page with text on it

*Instructions for this file are preceded by an asterisk, they are just informational. Actual STATA commands are indented and don't have an asterisk

*The commands and instructions presented here are suggestions and one method in which STATA can be used to analyze survey data

```
*============================================================
```

*Set ups

```
*============================================================
```

```
*STATE SAMPLE
   *use "hys21 state dataset.dta", clear
        gen fakewt=1
        svyset [pweight=fakewt], psu(schgrd)
```

*STATE CENSUS

```
use "E:\1 Data\HYS\HYS 2021\hys2021clean04152022.dta", clear
        *use "hys21 census dataset.dta", clear
        gen fakewt=1
        svyset [pweight=fakewt]
```

*ANALYSIS of a COUNTY with a CENSUS in 2021

*Adams, Asotin, Benton, Chelan, Clallam, Columbia, Cowlitz, Douglas, Ferry, Franklin, Garfield, Grant, Grays Harbor, Island, Jefferson, Kitsap, Kittitas, Klickitat, Lewis, Lincoln, Mason, Okanogan, Pacific, Pend Oreille, San Juan, Skagit, Skamania, Stevens, Wahkiakum, Walla Walla, Whatcom, Whitman, Yakima

*For COUNTY level setup, you need to replace the "x" with the proper county number

```
        *use "hys21 census dataset.dta", clear
        keep if conum==x
        keep if corec==1
        gen fakewt=1
        svyset [pweight=fakewt]
```

*Analysis of a COUNTY with a COUNTY SAMPLE in 2021

*King, Pierce, Snohomish, and Spokane 6th grade and 8th grade

```
*use "hys21 census dataset.dta", clear
keep if conum==X
keep if corec==1
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

*ANALYSIS of a COUNTY with both SAMPLE and CENSUS in 2021, Spokane

```
*use "hys21 census dataset.dta", clear
keep if conum==32
keep if corec==1
gen fakewt=1
gen id=_n
gen psu=id +10000
replace psu=schgrd if (grade==6 | grade==8)
svyset [pweight=fakewt], psu(psu)
```

*DROP any COUNTY/GRADE levels that cannot be reported for 2021

```
*use "hys21 census dataset.dta", clear
drop if conum==1 & grade==6
drop if conum==10 & (grade==10 | grade==12)
drop if conum==11
drop if conum==12 & grade==12
drop if conum==26 & grade==10
drop if conum==35
drop if conum==36 & grade==6
drop if conum==38 & grade==6
```

*drops counties and grades without 40% participation

```
drop if conum==2 & (grade==10 | grade==12)
drop if conum==4 & grade==12
drop if conum==8 & grade==12
drop if conum==19 & grade==12
drop if conum==21 & grade==12
drop if conum==23 & (grade==6 | grade==12)
drop if conum==24 & grade==12
drop if conum==27 & grade==12
drop if conum==32 & grade==12
drop if conum==33 & (grade==10 | grade==12)
drop if conum==34 & grade==12
```

*ESD ANALYSIS

```
*use "hys21 census dataset.dta", clear
keep if esdrec==1
svyset [pweight=esdwt], psu(esdpsu) strata(conum)
```

*if you only want to analyze one ESD then also use, designate the ESD number for "X"

```
keep if esdnum==x
```

*drops ESD and grades without 40% participation

```
drop if esdnum==101 & grade==12
```

```
drop if esdnum==113 & grade==12
drop if esdnum==123 & grade==12

*ANALYSIS of MULTIPLE COUNTIES using a Complete State Data Set (Census 2021)
*use "hys21 census dataset.dta", clear

*drops counties and grades that cannot be reported
drop if conum==1 & grade==6
drop if conum==10 & (grade==10 | grade==12)
drop if conum==11
drop if conum==12 & grade==12
drop if conum==26 & grade==10
drop if conum==35
drop if conum==36 & grade==6
drop if conum==38 & grade==6

*drops counties and grades without 40% participation
drop if conum==2 & (grade==10 | grade==12)
drop if conum==4 & grade==12
drop if conum==8 & grade==12
drop if conum==19 & grade==12
drop if conum==21 & grade==12
drop if conum==23 & (grade==6 | grade==12)
drop if conum==24 & grade==12
drop if conum==27 & grade==12
drop if conum==32 & grade==12
drop if conum==33 & (grade==10 | grade==12)
drop if conum==34 & grade==12

*generates psu that takes sampling into account
keep if corec==1
gen id=_n
gen psu=id +10000
replace psu=schgrd if (conum==17)
replace psu=schgrd if (conum==27)
replace psu=schgrd if (conum==31)
replace psu=schgrd if (conum==32 & grade==6)
replace psu=schgrd if (conum==32 & grade==8)
gen fakewt=1
svyset [pweight=fakewt], psu(psu)

*SCHOOL DISTRICT ANALYSIS - Never conduct without permission from school district
*use "hys21 census dataset.dta", clear
keep if distnum==x
keep if distrec==1
gen fakewt=1
svyset [pweight=fakewt]

*SCHOOL BUILDING ANALYSIS - Never conduct without permission from school district
*use "hys21 census dataset.dta", clear
```

```
keep if schnum==x
gen fakewt=1
svyset [pweight=fakewt]

*============================================================

*Analysis Examples

*============================================================
*use "hys21 state dataset.dta", clear
gen fakewt=1
svyset [pweight=fakewt], psu(schgrd)
```

*For crosstabs, the coding after the comma helps format your STATA output. Only include the options you need:

```
*col for column percentages
*row for row percentages
*per if you want the point estimates in percentage format
*ci for confidence intervals
*obs for "n"
*format(%3.2f) designates the numbers before and after the decimal (3 before the decimal, 2 after the decimal)
```

Current marijuana use by grade

To run one variable, d21_16use (current marijuana – already coded as no use or any use 0,1) by grade and include the following formatting options after the comma.

- col for column percentages
- per for results displayed in %
- se for standard error (to convert se to 95% ci, you need to *1.96)
- ci for upper/lower confidence intervals
- obs for "n"
- format(%3.2f) to designate the numbers before and after the decimal

```
svy:tab d21_16use grade, col per se ci obs format(%3.2f)
```

Current marijuana use by grade and sex assigned at birth

Generate a binary (1,0) sex dummy variable for subpopulations for marijuana use by grade among a specific sex.

```
tab g05_18, gen(sex)
rename sex1 female
rename sex2 male
svy:tab d21_16use grade, subpop(girl) col per se ci obs format(%3.2f)
svy:tab d21_16use grade, subpop(boy) col per se ci obs format(%3.2f)
```

Generate a binary (1,0) grade dummy variable for subpopulations for marijuana use by sex at birth for a specific grade.

```
tab grade, gen (gradecat)
svy:tab d21_16use g05_18, subpop(gradecat1) col per se ci obs format(%3.2f)
svy:tab d21_16use g05_18, subpop(gradecat2) col per se ci obs format(%3.2f)
svy:tab d21_16use g05_18, subpop(gradecat3) col per se ci obs format(%3.2f)
```

```
svy:tab d21_16use g05_18, subpop(gradecat4) col per se ci obs format(%3.2f)
```

**Current marijuana by grade and chronic absenteeism*

Create a chronic absenteeism variable (absent 3 or more days) with recode, then define and attach the new response option labels and add the new variable with a description. Crosstab marijuana by race and grade.

```
codebook g27
gen chronicabsent= g27
recode chronicabsent 1/2=0 3=1
lab def chronicabsent 1 "Yes-3 or more days" 0"No"
lab val chronicabsent chronicabsent
lab var chronicabsent "Absent from school on 3 or more days in past month for any reason"
svy:tab d21_16use chronicabsent, subpop(gradecat2) col per se ci obs format(%3.2f)
svy:tab d21_16use chronicabsent, subpop(gradecat3) col per se ci obs format(%3.2f)
svy:tab d21_16use chronicabsent, subpop(gradecat4) col per se ci obs format(%3.2f)
```

**Current marijuana use by depressive feelings*

Crosstab marijuana by depressive feelings and grade.

```
svy:tab d21_16use h53, subpop(gradecat2) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(gradecat3) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(gradecat4) col per se ci obs format(%3.2f)
```

**Current marijuana use by depressive among chronic absentee students*

Generate binary (0,1) for chronic absenteeism and grade subpopulations and crosstab marijuana use by depressive feelings among chronically absent students.

```
gen absent8=1 if chronicabsent==1 & grade==8
replace absent8=0 if chronicabsent==0 & (grade==10 | grade==12)
gen absent10=1 if chronicabsent==1 & grade==10
replace absent10=0 if chronicabsent==0 & (grade==8 | grade==12)
gen absent12=1 if chronicabsent==1 & grade==12
replace absent12=0 if chronicabsent==0 & (grade==8 | grade==10)
svy:tab d21_16use h53, subpop(absent8) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(absent10) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(absent12) col per se ci obs format(%3.2f)
```

**Current marijuana use by depressive feelings among boys*

Generate binary (0,1) for boys and grade subpopulations and crosstab marijuana use by depressive feelings among boys.

```
gen boy8=1 if g05_18==2 & grade==8
replace boy8=0 if g05_18==1 &  (grade==10 | grade==12)
gen boy10=1 if g05_18==2 & grade==10
replace boy10=0 if g05_18==1& (grade==8 | grade==12)
gen boy12=1 if g05_18==2 & grade==12
replace boy12=0 if g05_18==1 & (grade==8 | grade==10)
svy:tab d21_16use h53, subpop(boy8) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(boy10) col per se ci obs format(%3.2f)
svy:tab d21_16use h53, subpop(boy12) col per se ci obs format(%3.2f)
```

NOTE: Use caution with crosstabs of variables with low prevalence, or when you are using small subpopulations. Do NOT report results if there are less than 5 observations per cell when

running state level data or less than 10 observations per cell when running sub-state-level analysis.

Appendix E: Do File – Making Bar Graphs with Error Bars in STATA

*The following "do file" runs through see the Displaying Results section

*Select the proper set up according to your data, to replicate the results in the manual – use the 2021 state sample dataset

*To run a line of command highlight the command text and hit the icon above that looks like a page with text on it

*Instructions for this file are preceded by an asterisk, they are just informational. Actual STATA commands are indented and don't have an asterisk

*The commands and instructions presented here are suggestions and one method in which STATA can be used to analyze survey data

*Modified from UCLA/s STATA website at:
http://www.ats.ucla.edu/stat/STATA/faq/barcap.htm

*==========

*Chart Example

*==========

```
use "E:\1 Data\HYS\HYS 2021\hys2021state04152022.dta", clear
        *use "hys21 state dataset.dta", clear
        gen fakewt=1
        svyset [pweight=fakewt], psu(schgrd)

*generate a race variable with the groups you want in the graph
gen race=g06
recode race 1=1 2=2 3=3 4=4 5=1 6=5 7=. 8=.
lab def newrace 1"API" 2"Indian" 3"Black" 4"Hispanic" 5"White"
lab val race newrace

*create a mean current marijuana use prevalence
        collapse (mean) d21_16use= d21_16use (sd) sdd21_16use=d21_16use (count) n=d21_16use,
        by(grade race)

*create the high and low values of the confidence interval
        generate hid21_16use = d21_16use + invttail(n-1,0.025)*(sdd21_16use/sqrt(n))
        generate lod21_16use = d21_16use - invttail(n-1,0.025)*(sdd21_16use/sqrt(n))

*generate a simple two-way bar graph
        graph bar d21_16use, over(race) over(grade)

*add some color to the graph and make it a bit easier to read by adding asyvars
        graph bar d21_16use, over(race) over(grade) asyvars

*add error bars to the graph
```

```
graph twoway (bar d21_16use race) (rcap hid21_16use lod21_16use race), by(grade)
```

*to make a color two-way bar graph with error bars set up single variables

*for each race and grade

```
gen graderace=race if grade==6
replace graderace=race+10 if grade==8
replace graderace=race+20 if grade==10
replace graderace=race+30 if grade==12
sort graderace
list graderace grade race, sepby(grade)
```

*create a single graph with all of the data

```
twoway (bar d21_16use graderace)
```

*add confidence intervals

```
twoway (bar d21_16use graderace) (rcap hid21_16use lod21_16use graderace)
```

*to add color overlay four separate graphs

```
twoway (bar d21_16use graderace if race==1) (bar d21_16use graderace if race==2) (bar
d21_16use graderace if race==3) (bar d21_16use graderace if race==4) (bar d21_16use graderace
if race==5) (rcap hid21_16use lod21_16use graderace)
```

*add a legend and labels

```
twoway (bar d21_16use graderace if race==1)  (bar d21_16use graderace if race==2)  (bar
d21_16use graderace if race==3)  (bar d21_16use graderace if race==4)  (bar d21_16use
graderace if race==5)  (rcap hid21_16use lod21_16use graderace), legend(order(1 "API" 2 "Indian"
3 "Black" 4 "Hispanic" 5 "White")) xlabel(2.5 "6th Grade" 12.5 "8th Grade" 22.5 "10th Grade" 32.5
"12th Grade", noticks)xtitle(Grade)  ytitle(Mean Current Marijuana Use Prevalence) title(Current
Marijuana Use) subtitle(by Race and Grade) note(Source: 2021 HYS)
```