# How do I utilize the subpopn statement in SUDAAN?

Authored by

**stats writer**

July 1, 2024

The subpopn statement in SUDAAN is a useful tool for analyzing data from specific subgroups within a larger population. It allows researchers to specify a subset of the population based on certain characteristics or criteria, and then perform statistical analyses on that subset. This can be particularly helpful when studying subgroups that may have unique characteristics or experiences that are of interest to the researcher. By using the subpopn statement, researchers can obtain more targeted and accurate results, as well as gain a deeper understanding of the data. Overall, the subpopn statement is a valuable feature in SUDAAN that can enhance the precision and relevance of statistical analyses for researchers.

## How can I use the subpopn statement in SUDAAN? | SUDAAN FAQ

## Below is an example of the subpopn statement. This statement should be used whenever you want to analyze only a subpopulation in your data. You should NOT subset your data in a data step before running the analysis, as this can cause a wide variety of problems, from incorrect results to difficulties running the procedure at all. See the section of the Features and Functions chapter of the SUDAAN manual for more information regarding the subpopn statement, how to use it, and how missing values are handled. It includes a

note with a more complete explanation of why
the subpopn statement should be used instead of subsetting the
data first. Other references on this include Cochran (1977, Section
2.13, pages 35-38) and the Stata Survey Manual. There are a few
basic reasons why you should not subset your data in order to look at just a
subpopulation. One is that the standard errors of the estimates may be
incorrect, and another is that the sampling information for observations not
included in the subpopulation is still used in the calculations. If you
delete these observations before making the calculations, then that
information is not available. Also, depending on how you subset, you may
find that you have strata with too few PSUs to run the procedure.

The example below shows a regression for just the males in the data set (srsex =

**1). We have bolded the note in the output that indicates the subpopulation used. The subgroup and levels statements are used to indicate that racehpra is a categorical variable with four levels. Beginning with SUDAAN 9, you can use the class statement instead of these two statements.**

```
proc regress data=temp1 filetype=sas design = jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
model ae13 = ae14 racehpra;
subpopn srsex = 1;
subgroup racehpra;
levels 4;
run;
```

**Number of observations read : 55428 Weighted count: 23847415**
**Observations in subpopulation : 23002 Weighted count: 11631728**

**Observations used in the analysis : 3744 Weighted count: 2522055**

**Denominator degrees of freedom : 80**

**Maximum number of estimable parameters for the model is 5**

**Weighted mean response is 3.133033**

**Multiple R-Square for the dependent variable AE13: 0.231226**

**Variance Estimation Method: Replicate Weight Jackknife**

**Working Correlations: Independent**

**Link Function: Identity**

**Response variable AE13: Number of drinks on the days drinking alcohol**

**For Subpopulation: SRSEX = 1**

---

**Independent P-value**

**Variables and Beta T-Test**

**Effects Coeff. SE Beta T-Test B=0 B=0**

---

**Intercept 1.71 0.07 24.92 0.0000**

**Number of times**

**having 5 or more**
**drinks in past**
**month 0.38 0.04 9.67 0.0000**
**Race - UCLA CHPR**
**Definition**
**LATINO 1.29 0.11 12.31 0.0000**
**PACIFIC ISLANDER 0.84 0.59 1.44 0.1543**
**AIAN 0.54 0.24 2.20 0.0307**
**ASIAN 0.00 0.00 . .**
**----------**

**----------**

**Contrast Degrees**
**of P-value**
**Freedom Wald F Wald F**
**----------**
**OVERALL MODEL 5 618.86 0.0000**
**MODEL MINUS**
**INTERCEPT 4 63.04 0.0000**
**INTERCEPT . . .**
**AE14 1 93.52 0.0000**
**RACEHPRA 3 50.72 0.0000**
**----------**

**In this example, we have two conditions on the subpopn statement.**
**Hence, the regression results apply only to those cases where both srsex = 1**
**and racehpra = 2 is true.**

```
proc regress data=temp1 filetype=sas design = jackknife;
weight rakedw0;
jackwgts rakedw1--rakedw80 / adjjack=1;
model ae13 = ae14 ;
subpopn srsex = 1 and racehpra = 2;
run;
```

**Number of observations read : 55428 Weighted count: 23847415**
**Observations in subpopulation : 101 Weighted count: 30282**
**Observations used in the analysis : 69 Weighted count: 17998**
**Denominator degrees of freedom : 80**

**Maximum number of estimable parameters for the model is 2**

**Weighted mean response is 3.607368**

**Multiple R-Square for the dependent variable AE13: 0.068544**
**Variance Estimation Method: Replicate Weight Jackknife**
**Working Correlations: Independent**
**Link Function: Identity**
**Response variable AE13: Number of drinks on the days drinking alcohol**
**For Subpopulation: SRSEX = 1 AND RACEHPRA = 2**

---

**Independent P-value**
**Variables and Beta T-Test**
**Effects Coeff. SE Beta T-Test B=0 B=0**

---

**Intercept 3.05 0.63 4.86 0.0000**
**Number of times**
**having 5 or more**
**drinks in past**
**month 0.20 0.13 1.60 0.1145**

---

---

**Contrast Degrees**

**of P-value**

**Freedom Wald F Wald F**

**----------**

**OVERALL MODEL 2 19.02 0.0000**

**MODEL MINUS**

**INTERCEPT 1 2.55 0.1145**

**INTERCEPT 1 23.64 0.0000**

**AE14 1 2.55 0.1145**

**----------**