SAS Statistics Research and Applications
Paper #2023-02

Comparing Domain Statistics in Survey Data Analysis with PROC SURVEYMEANS

Tony An, SAS Institute Inc., San Francisco, CA

Domain analysis is a frequently performed task in survey data analysis. In SAS/STAT® software, the DOMAIN statement in the survey procedures provides the capability for such analysis. Quite often, researchers are interested in comparing domain statistics such as domain means and domain ratios. The DIFF option in the SURVEYMEANS procedure enables comparisons of domain statistics across domain levels. Through examples, this paper illustrates how to perform comparisons of domain statistics for pairwise levels of a defined domain.

Introduction

Researchers often use the methodology of survey sampling to obtain information about a large population by selecting and measuring a sample from that population. Because of variability among sampled subjects, they apply probability-based scientific designs to select the sample. This reduces the risk of a distorted view of the population and enables them to make statistically valid inferences from the sample.

When a sample is drawn according to a complex sample design, data are collected from both response variables and auxiliary variables. Based on the sample design, sampling weights are created to ensure unbiased estimates for response variables in the population, and the sample design information is also stored in order to be used to estimate variance.

After a sample is collected, you might be interested in estimating some characteristics of a subpopulation, or domain, on the basis of demographic information that might or might not be part of the sample design. This is often called domain analysis. Statistics that are computed for those subpopulations, such as domain means and domain ratios, are often called domain statistics. Inferences about these domain statistics can be of great interest to researchers. Domain mean comparison and domain ratio comparison are tasks that are frequently performed in domain analysis.

Let's take a look at an example of domain mean comparison. A survey of hospital patients is carried out to study information about insurance copayments, and the sample design uses patients' states of residency as strata. For each patient who participates in this survey, along with the copayment information, demographic and auxiliary information is collected, such as race, gender, age, insurance type, and so on. The average patient copay is estimated from these survey data. During the analysis of the survey, however, you might want to estimate whether there is a difference in the average copay amount between two different racial groups. Because race is not part of the sample design, analysis involving race becomes a domain analysis, where the domain levels are racial groups. The average copay amounts by racial groups are examples of domain statistics, and the test of whether there is a significant difference between two racial groups falls within the scope of the comparison of domain means.

Now let's look at an example of the comparison of domain ratios. As the effects of global climate change become more evident, we are interested in determining what types of energy are being used so that we can create policies to increase the use of renewable energy. A nationwide probability sample collects data about both total monthly energy consumption and the amount of renewable energy consumption from customers. The renewable rate is the ratio of renewable energy usage to total energy consumption, and it is estimated at the national level. If we also want to know whether there is a significant difference in the renewable rate between two states, then this becomes a domain ratio comparison problem, where the domain is defined by the states.

Most SAS® procedures include the BY statement, which enables you to perform independent and separate analyses of subsets of your data. Using a BY statement is different from domain analysis in survey sampling. Because the formation of domains can be unrelated to the sample design, the domain sample sizes can be random variables. Domain analysis takes this variability into account by using the entire sample to estimate the variance of domain estimates. Domain analysis is also known as subgroup analysis, subpopulation analysis, or subdomain analysis. For more information about domain analysis, see Fuller (2009); Lohr (2022); Särndal, Swensson, and Wretman (1992); Wolter (2007); and Cochran (1977).

You can perform domain analysis by using the DOMAIN statement in the SURVEYMEANS, SURVEYFREQ, SURVEYREG, SURVEYLOGISTIC, and SURVEYPHREG procedures, which all properly analyze complex survey data by taking into account the sample design and using the entire sample.

Moreover, PROC SURVEYMEANS also enables comparisons among domain statistics. It provides the DIFF option in its DOMAIN statement, which is specifically designed for comparing domain means and domain ratios, as described in the previous examples.

Note that the comparison of domain means applies only to continuous variables in PROC SURVEYMEANS. You can compare the differences in proportions of a categorical variable in PROC SURVEYFREQ.

The following sections provide details and examples of domain means, domain ratios, and their comparisons.

Syntax

To perform domain analysis in PROC SURVEYMEANS, you specify a DOMAIN statement as follows:

```
DOMAIN variables <variable*variable ...> / options;
```

The DOMAIN statement names the variables that identify domains, which are called domain variables; they can be either character or numeric. If a variable appears by itself in a DOMAIN statement, each level of this variable defines a domain in the study population. If two or more variables are joined by asterisks (*), then every possible combination of levels of these variables defines a domain. The procedure performs a descriptive analysis within each domain that is defined by the domain variables.

To perform a comparison of domain means for each continuous analysis variable, you can specify the DIFFMEANS option (or the DIFF option) in the DOMAIN statement as follows, and PROC SURVEYMEANS computes differences between domain means for pairwise levels of a defined domain:

```
DOMAIN variables / DIFFMEANS;
```

To perform a comparison of domain ratios, you can specify the DIFFRATIOS option (or the DIFF option) in the DOMAIN statement as follows, and the procedure computes differences between domain ratios for pairwise levels of each domain that you define:

```
DOMAIN variables / DIFFRATIOS;
```

Alternatively, you can specify the DIFF option as follows. This is equivalent to specifying both the DIFFMEANS and DIFFRATIOS options.

```
  DOMAIN variables / DIFF;
```

Differences in Domain Statistics

When you use a DOMAIN statement to perform a domain analysis, the procedure computes the requested statistics for each domain level.

For a domain D, let I_D be the corresponding indicator variable:

$$I_D(h,i,j) = \begin{cases} 1 & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

where

- $h = 1, 2, \ldots, H$ is the stratum index
- $i = 1, 2, \ldots, n_h$ is the cluster index within stratum h
- $j = 1, 2, \ldots, m_{hi}$ is the unit index within cluster i of stratum h

Let

$$z_{hij} \quad = \quad y_{hij} I_D(h,i,j) \quad = \begin{cases} y_{hij} & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

Denote w_{hij} as the sampling weight for unit j in cluster i of stratum h.

Let

$$v_{hij} = w_{hij} I_D(h,i,j) = \begin{cases} w_{hij} & \text{if observation } (h,i,j) \text{ belongs to } D \\ 0 & \text{otherwise} \end{cases}$$

The requested statistics for variable y in domain D are computed by using the new weights v.

For comparing domain statistics, let $D_1, D_2, \ldots, D_r$ be the r levels of D, and let the corresponding indicator variables be

$$I_{D_k}(h,i,j) = \begin{cases} 1 & \text{if observation } (h,i,j) \text{ belongs to } D_k \\ 0 & \text{otherwise} \end{cases}$$

Domain Means

The estimated mean of Y in the domain D is

$$\hat{\hat{Y}}_D = \left(\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}\ y_{hij} \right) \Big/ \ v_{\ldots}$$

where

$$v_{\ldots} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} v_{hij}$$

PROC SURVEYMEANS estimates the variance of variance of $\hat{\hat{Y}}_D$ by using either the Taylor series method or replication methods, and it computes related statistics such as confidence intervals accordingly. For more information, see the section “Statistical Computations” in the SURVEYMEANS procedure chapter of the *SAS/STAT User’s Guide*.

The difference between the means for domain levels D_{k1} and D_{k2} ($1 \le k1 \ne k2 \le r$) can be expressed as

$$\Delta(Y, D, k1, k2) = \hat{\hat{Y}}_{D_{k1}} - \hat{\hat{Y}}_{D_{k2}}$$

The estimated variance for this difference is

$$\hat{V}(\Delta(Y, D, k1, k2)) = \hat{V}(\hat{\hat{Y}}_{D_{k1}}) + \hat{V}(\hat{\hat{Y}}_{D_{k2}}) - 2\hat{\text{Cov}}(\hat{\hat{Y}}_{D_{k1}}, \hat{\hat{Y}}_{D_{k2}})$$

where the estimated variances $\hat{V}(\hat{\hat{Y}}_{D_{k1}})$ and $\hat{V}(\hat{\hat{Y}}_{D_{k2}})$ for means at corresponding domain levels (in addition to the covariance between these two domain means) are described as in the section “Domain Mean” in the SURVEYMEANS procedure chapter of the *SAS/STAT User’s Guide*.

Domain Ratios

The estimated ratio of Y to X in domain D is

$$\hat{R}_D = \frac{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \ v_{hij}\ y_{hij}}{\sum_{h=1}^{H} \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} \ v_{hij}\ x_{hij}}$$

The difference between domain levels D_{k1} and D_{k2} ($1 \le k1 \ne k2 \le r$) in domain D can be expressed as

$$\Delta(Y/X, D, k1, k2) = \hat{R}_{D_{k1}} - \hat{R}_{D_{k2}}$$

Replication methods are used to estimate the variance for this difference, $\hat{V}(\Delta(Y/X, D, k1, k2))$. It is computed by measuring the variability among the estimates that are derived from each replicate. For more information, see the section "Replication Methods for Variance Estimation" in the SURVEYMEANS procedure chapter of the *SAS/STAT User's Guide*.

Example of Domain Mean Comparison

In this section we use an example to illustrate the comparison of domain means. The data set is generated for the purpose of illustration, and this is not a real survey.

To evaluate the efficacy of a vaccine that has been developed for a communicable disease, a stratified probability sample is collected in which the strata are the 50 states of the United States. A simple random sample is collected in each state, and a participant's vaccination and infection status, along with demographic information about the participant, is collected.

The data are saved in the SAS data set `vaccine`. A total of 400 patients participate in the study. Figure 1 displays the first 10 patients in the `vaccine` data set.

The variable `state` indicates the US state where the patient resides; it is the stratification variable. The variable `gender` identifies the patient's gender. The variable `vaccination` records whether or not the patient has received the vaccine that is being evaluated. The variable `infection` indicates whether the patient has contracted the disease. The variable `weight` contains the sampling weights, which sum to the total population size of the United States.

Figure 1: First 10 Patients in the Vaccine Study

OBS	State	Gender	Vaccination	Infection
1	TN	Female	Yes	0
2	CO	Male	Yes	0
3	SD	Male	No	0
4	ND	Female	No	0
5	MI	Male	No	0
6	HI	Male	Yes	0
7	NC	Female	Yes	0
8	OR	Female	No	0
9	AZ	Male	No	0
10	FL	Male	No	1

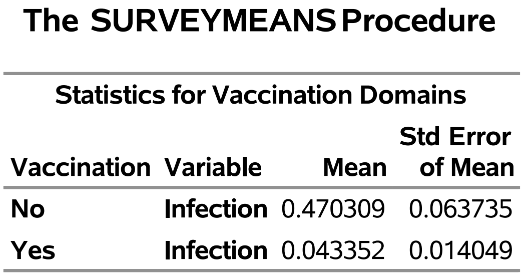
The following SAS code uses PROC SURVEYMEANS to perform domain analysis to study the infection rate, which is the estimated mean of the variable `infection`. The stratum variable `state` is specified in the STRATA statement. Two domains are defined by the DOMAIN statement: one is defined by the variable `vaccination`, and the other is defined by the variable `gender`. To determine whether there is a significant difference in the infection rates between the domain levels, the DIFFMEANS option is specified in the DOMAIN statement.

```
proc surveymeans data=vaccine mean;
   strata state;
   var infection;
```

```
    domain vaccination gender /diffmeans;
    weight weight;
run;
```

Figure 2 displays the domain estimates of the infection rates. The infection rate in the unvaccinated population is 47.03%, but in the vaccinated population it is only 4.33%.

Figure 2: Infection Rates in Each Vaccination Status Group



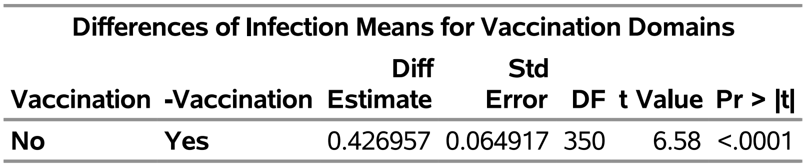
The SURVEYMEANS Procedure

Statistics for Vaccination Domains

Vaccination	Variable	Mean	Std Error of Mean
No	**Infection**	0.470309	0.063735
Yes	**Infection**	0.043352	0.014049

Figure 3 displays the difference in infection rates, 42.70%, between the vaccinated and unvaccinated patients. The difference is statistically significant, indicating that the studied vaccine is highly effective.

Figure 3: Comparison of Infection Rates between Vaccination Status Groups

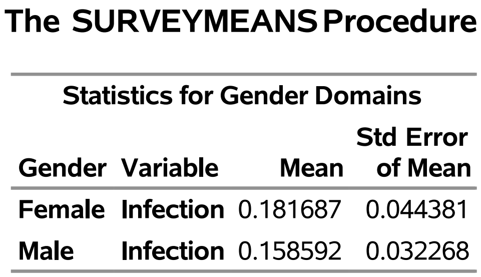


Differences of Infection Means for Vaccination Domains

Vaccination	-Vaccination	Diff Estimate	Std Error	DF	t Value	Pr > |t|
No	**Yes**	0.426957	0.064917	350	6.58	<.0001

For the domain `gender`, Figure 4 shows the infection rates for each gender.

Figure 4: Infection Rates for Each Gender



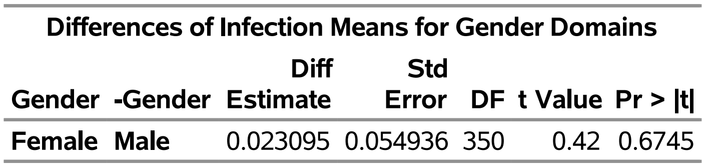
The SURVEYMEANS Procedure

Statistics for Gender Domains

Gender	Variable	Mean	Std Error of Mean
Female	**Infection**	0.181687	0.044381
Male	**Infection**	0.158592	0.032268

Figure 5 shows the difference in infection rates between genders. It appears that the disease infects female patients (18.17%) more than male patients (15.86%). The difference is 2.31%, which is not statistically significant.

Figure 5: Comparison of Infection Rates between Genders



Differences of Infection Means for Gender Domains

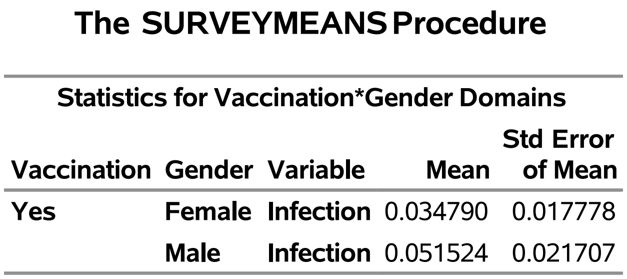
Gender	-Gender	Diff Estimate	Std Error	DF	t Value	Pr > |t|
Female	**Male**	0.023095	0.054936	350	0.42	0.6745

Moreover, we are interested in finding out whether there is any difference in the infection rate between genders in the vaccinated population. To do so, we specify a new domain definition in the following DOMAIN statement as `vaccination*gender`, and we specify `vaccination('Yes')` to restrict the output to display only the results for the vaccinated group:

```
proc surveymeans data=vaccine mean;
   strata state;
   var infection;
   domain vaccination(’Yes’)*gender /diffmeans;
   weight weight;
run;
```

As shown in Figure 6 and Figure 7, the infection rate for males (5.15%) is slightly higher than for females (3.48%) in the vaccinated population. The difference is 1.67%, which is not statistically significant.

Figure 6: Infection Rates for Vaccinated Patients by Gender

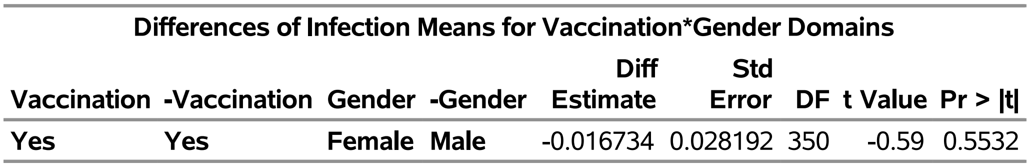


The SURVEYMEANS Procedure

Statistics for Vaccination*Gender Domains

Vaccination	Gender	Variable	Mean	Std Error of Mean
Yes	**Female**	**Infection**	0.034790	0.017778
	Male	**Infection**	0.051524	0.021707

Figure 7: Comparison of Infection Rates between Genders in Vaccinated Group



Differences of Infection Means for Vaccination*Gender Domains

Vaccination	-Vaccination	Gender	-Gender	Diff Estimate	Std Error	DF	t Value	Pr > |t|
Yes	**Yes**	**Female**	**Male**	-0.016734	0.028192	350	-0.59	0.5532

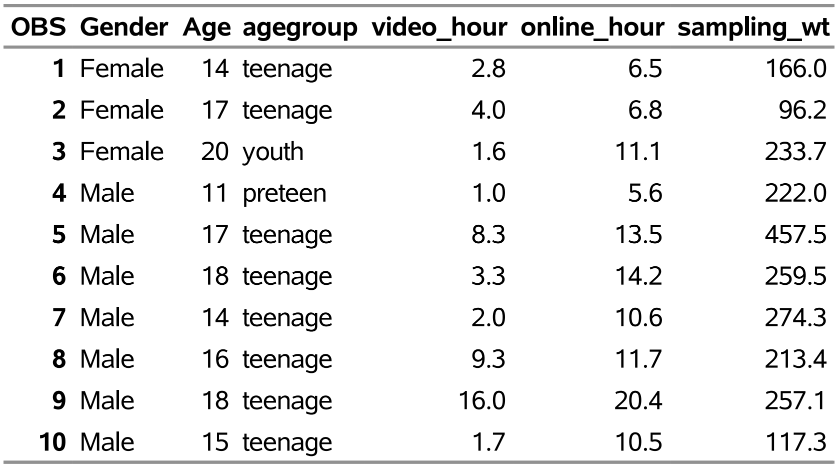
Example of Domain Ratio Comparison

Now let’s look at an example that shows how to perform domain ratio comparison.

Suppose that a company conducts market research to study young people’s online activities, such as viewing videos, reading, gaming, writing, and so on. More specifically, the company wants to find out how much time the young people spend viewing videos when they are online, so that it can make decisions about future products.

A stratified unequal probability sample is drawn. There are a total of 300 participants in the study, ages 11 to 25, and data from their digital devices are collected over a specific period of time. The data are saved in the data set `usage`. The variable `gender` is used for stratification. The variable `online_hour` records the number of hours that a participant spends online during the specified time period, and the variable `video_hour` records how many of those hours are spent viewing videos. The variable `sampling_wt` stores the sampling weights. Figure 8 displays the first 10 observations of the `usage` data set.

Figure 8: First 10 Observations of Online Usage



OBS	Gender	Age	agegroup	video_hour	online_hour	sampling_wt
1	Female	14	teenage	2.8	6.5	166.0
2	Female	17	teenage	4.0	6.8	96.2
3	Female	20	youth	1.6	11.1	233.7
4	Male	11	preteen	1.0	5.6	222.0
5	Male	17	teenage	8.3	13.5	457.5
6	Male	18	teenage	3.3	14.2	259.5
7	Male	14	teenage	2.0	10.6	274.3
8	Male	16	teenage	9.3	11.7	213.4
9	Male	18	teenage	16.0	20.4	257.1
10	Male	15	teenage	1.7	10.5	117.3

We want to compare video viewing habits in the preteen, teenage, and youth age groups and also determine whether there is a difference in viewing habits between genders. Therefore, the variable agegroup and the variable gender are used to define domains in the following program, which uses PROC SURVEYMEANS to compute the ratio of video_hour to online_hour. The DIFFRATIOS option performs the ratio comparisons. We use the jackknife method to estimate the variances.

```
ods graphics on;
proc surveymeans ratio method=jk;
ratio video_hour/online_hour;
strata gender;
domain agegroup gender/diffratios;
weight sampling_wt;
run;
```

Figure 9 shows the domain means of video viewing hours and hours spent online for each age group.

Figure 9: Domain Means of Video Viewing Hours and Online Hours by Age Group

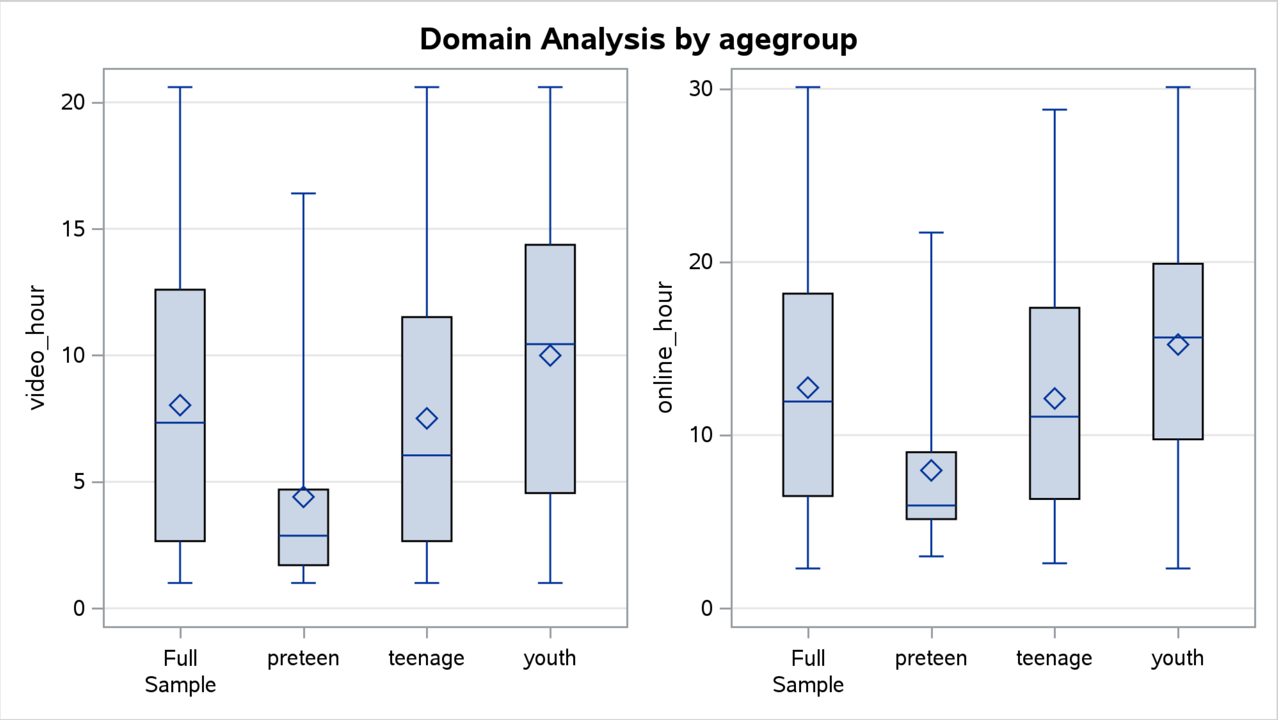
The SURVEYMEANS Procedure

Statistics for agegroup Domains

agegroup	Variable	Mean	Std Error of Mean
preteen	**video_hour**	4.403283	0.861859
	online_hour	7.963177	1.027912
teenage	**video_hour**	7.507407	0.474863
	online_hour	12.117016	0.567894
youth	**video_hour**	9.992479	0.685599
	online_hour	15.233573	0.701921

Figure 10 graphically displays these statistics. Both video viewing hours and online hours increase as the young people grow older.

Figure 10: Domain Means of Video Viewing Hours and Online Hours by Age Group



The ratio of video viewing to online activities in each age group is computed and displayed in Figure 11.

Figure 11: Ratio of Video Viewing to Online Activities by Age Group

The SURVEYMEANS Procedure

Ratios for agegroup Domains

agegroup	Numerator	Denominator	Ratio	Std Error
preteen	video_hour	online_hour	0.552956	0.052945
teenage	video_hour	online_hour	0.619576	0.015377
youth	video_hour	online_hour	0.655951	0.021611

Figure 12 shows the comparison of domain ratios among the age groups, which are all insignificant from group to group. For example, the percentage of video viewing hours increased 6.66% in the teenage group (61.96%) compared to the preteen group (55.30%), but the increase is not significant. Similarly, even though the ratio increases another 3.64% in the youth group (65.60%) compared to the teenage group, this increase is also not significant. The ratio increased 10.30% from the preteen group to the youth group, and this increase is still statistically insignificant. But the overall trend, according to this study, is that as young people get older, they tend to spend more of their time online viewing videos.

Figure 12: Comparison of Ratio of Video Viewing to Online Activities across Age Groups

Ratio Comparison for agegroup Domains

Numerator	Denominator	agegroup	-agegroup	Diff Estimate	Std Error	DF	t Value	Pr > |t|
video_hour	online_hour	preteen	teenage	-0.066620	0.055096	298	-1.21	0.2276
		preteen	youth	-0.102996	0.057166	298	-1.80	0.0726
		teenage	youth	-0.036376	0.026549	298	-1.37	0.1717

For the domain that is defined by gender, Figure 13 shows the domain means of video viewing hours and hours spent online for each gender.

Figure 13: Domain Analysis of Video Viewing Hours and Online Hours by Gender

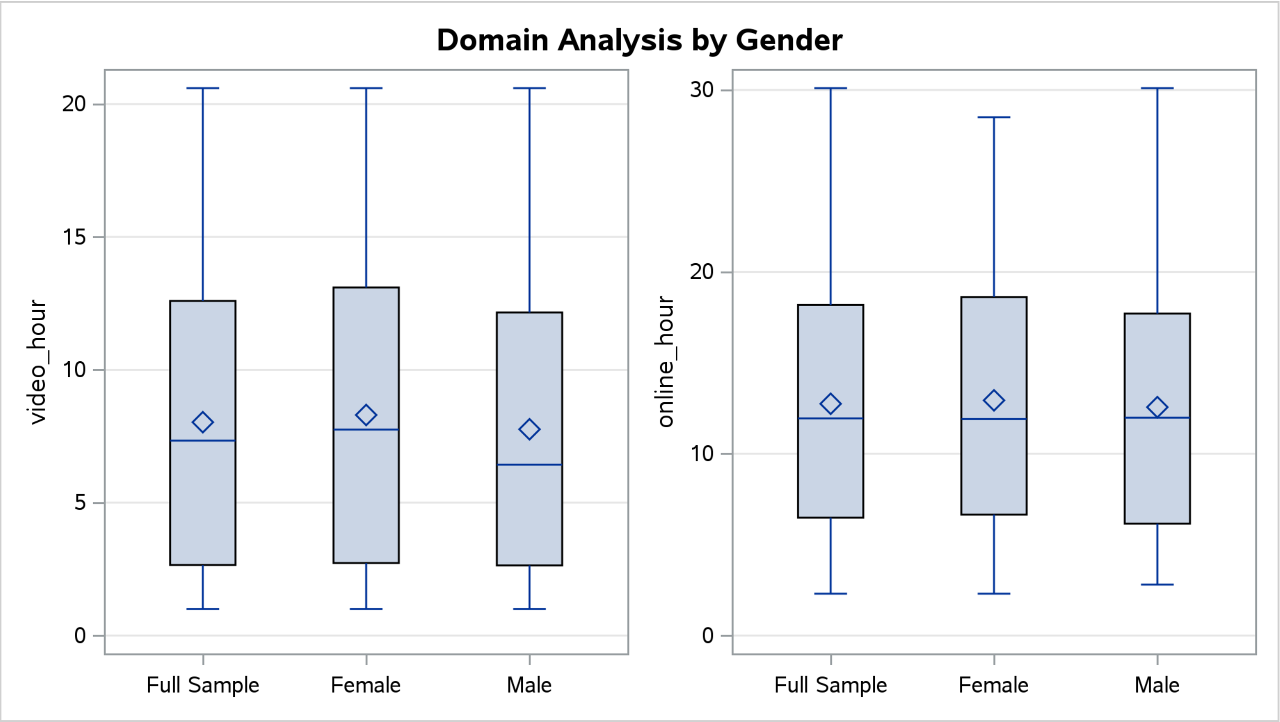
The SURVEYMEANS Procedure

Statistics for Gender Domains

Gender	Variable	Mean	Std Error of Mean
Female	video_hour	8.299641	0.543733
	online_hour	12.932284	0.615519
Male	video_hour	7.762410	0.507902
	online_hour	12.558354	0.584483

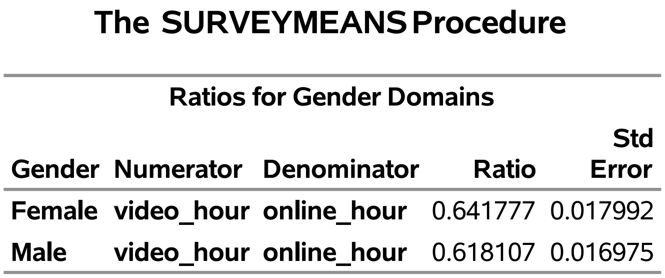
Figure 14 presents these statistics in graphics.

Figure 14: Domain Means of Video Viewing Hours and Online Hours by Gender



For the domain that is defined by gender, Figure 15 shows the ratio of video viewing hours to online hours in each gender group. The female group spends 64.17% of their online hours watching videos, whereas the male group spends 61.81% of their time online doing so. Figure 16 shows the comparison of these ratios in each gender group. The difference is 2.37%, which is not significant.

Figure 15: Domain Ratios of Video Viewing to Online Activities by Gender

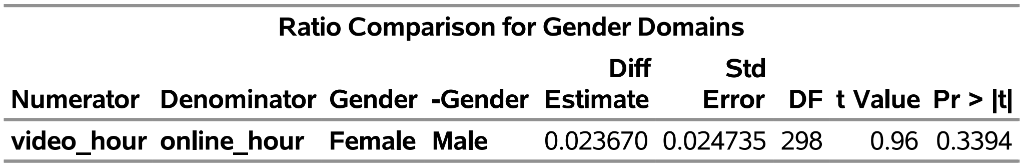


The SURVEYMEANS Procedure

Ratios for Gender Domains

Gender	Numerator	Denominator	Ratio	Std Error
Female	**video_hour**	**online_hour**	0.641777	0.017992
Male	**video_hour**	**online_hour**	0.618107	0.016975

Figure 16: Comparison of Ratio of Video Viewing to Online Activities between Genders



Ratio Comparison for Gender Domains

Numerator	Denominator	Gender	-Gender	Diff Estimate	Std Error	DF	t Value	Pr > |t|
video_hour	**online_hour**	**Female**	**Male**	0.023670	0.024735	298	0.96	0.3394

References

Cochran, W. G. (1977). *Sampling Techniques*. 3rd ed. New York: John Wiley & Sons.

Fuller, W. A. (2009). *Sampling Statistics*. Hoboken, NJ: John Wiley & Sons.

Lohr, S. L. (2022). *Sampling: Design and Analysis*. 3rd ed. Boca Raton, FL: CRC Press.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. 2nd ed. New York: Springer.

Acknowledgment

The author would like to thank Ed Huddleston and Yiu-Fai Yung for their help on the paper.