RANDOM SAMPLING IN SAS: Using PROC SQL and PROC SURVEYSELECT

Monique Ardizzi

TransUnion Canada

OCTOBER 2015

Agenda:

- Why Sample?
- Sampling Terminology
- Example Problem: BWeights Dataset in SAShelp
- Simple Random Sampling using PROC SQL and PROC SURVEYSELECT
- Stratified Random Sampling using PROC SQL and PROC SURVEYSELECT
- Summary and Comparison of Methods
- Q&A

Why Sample?

Not practical or not possible to have data on the entire population of interest

- For example, determining the average height of men in North America

Computational and physical constraints

- You may not have enough space to store such a large dataset

You can save time and money

- Data requests are likely charged based on volume (e.g. Stats Canada)

Testing Purposes

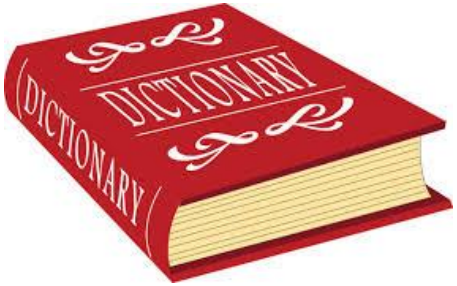
- For example, testing your program

Sampling Terminology 101

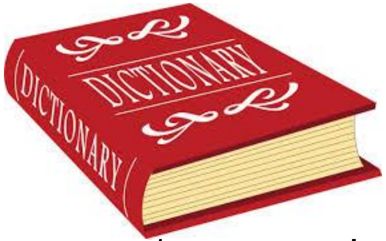
SAMPLE—a subset of the population

SAMPLING—the selection process used the extract the sample

PROBABILITY SAMPLING—a sampling method where each unit in the population is given a known probability of selection and a random mechanism is used to select specific units for the sample



Sampling Terminology 102



SIMPLE RANDOM SAMPLING—a sampling method where n units are randomly selected from a population of N units and every possible sample has an equal chance of being selected

STRATIFIED RANDOM SAMPLING—a sampling method where the population is first divided into mutually exclusive groups called **strata**, and simple random sampling is performed in each strata

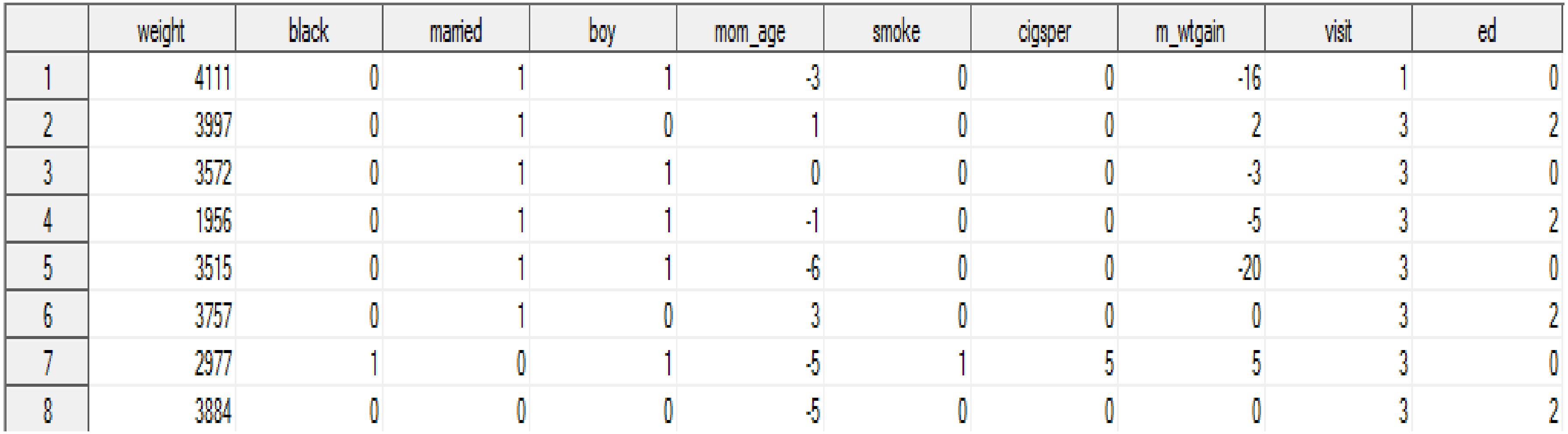
SYSTEMATIC SAMPLING—a sampling method that lists the N members of the population, randomly selects a starting point, and selects every kth member of the list for inclusion in the sample, where k=N/n and n is the sample size

CLUSTER SAMPLING—a sampling method where the population is first divided into mutually exclusive groups called **clusters**, and simple random sampling is performed to select the clusters to be included in the sample

Example Problem: Bweight Dataset in SAShelp

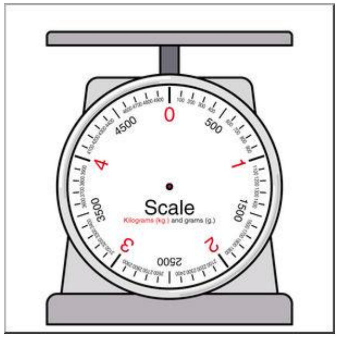
I will be using the data set **Bweight in the SAShelp Library** throughout this presentation.

- There are 50,000 observations
- The data is from the National Center for Health Statistics and record live, single births to mothers aged 18-45 in the United States in 1997 who were classified as black or white



	weight	black	married	boy	mom_age	smoke	cigsper	m_wtgain	visit	ed
1	4111	0	1	1	-3	0	0	-16	1	0
2	3997	0	1	0	1	0	0	2	3	2
3	3572	0	1	1	0	0	0	-3	3	0
4	1956	0	1	1	-1	0	0	-5	3	2
5	3515	0	1	1	-6	0	0	-20	3	0
6	3757	0	1	0	3	0	0	0	3	2
7	2977	1	0	1	-5	1	5	5	3	0
8	3884	0	0	0	-5	0	0	0	3	2

Example Problem: The Goal



Our Goal

Suppose that only 50,000 babies were born in the U.S. in 1997, thus we have data available on the entire population of interest. We want to measure:

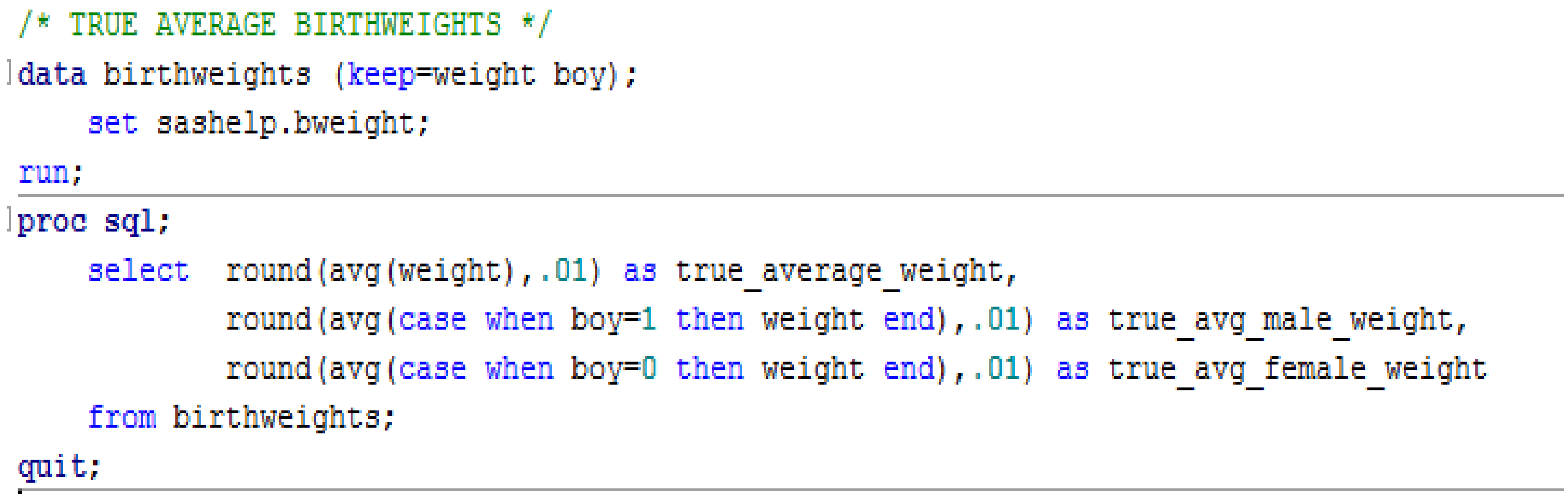
1. The average birthweight of an American child in 1997
2. The average birthweight of an American female child and an American male child in 1997

Sampling Methods to be Used

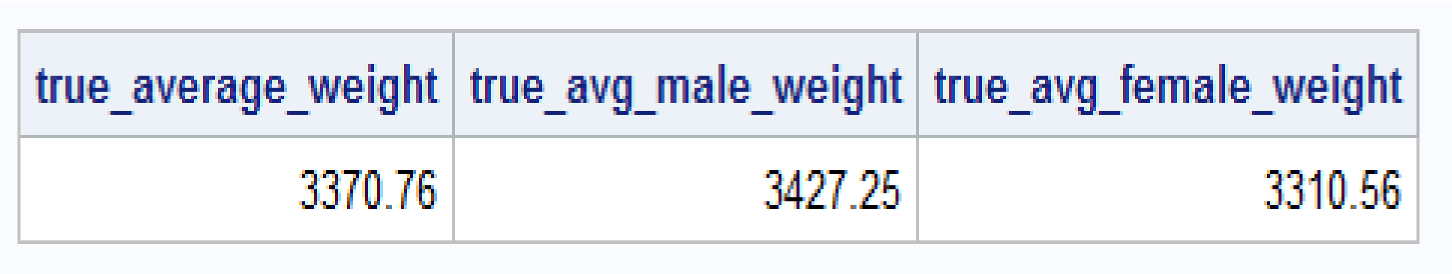
1. Simple random sampling
2. Stratified random sampling

Example Problem: What if we Didn’t Sample?

Let’s calculate the metrics of interest by using the entire population.



```
/* TRUE AVERAGE BIRTHWEIGHTS */
data birthweights (keep=weight boy);
     set sashelp.bweight;
run;
proc sql;
     select  round(avg(weight),.01) as true_average_weight,
             round(avg(case when boy=1 then weight end),.01) as true_avg_male_weight,
             round(avg(case when boy=0 then weight end),.01) as true_avg_female_weight
     from birthweights;
quit;
```



true_average_weight	true_avg_male_weight	true_avg_female_weight
3370.76	3427.25	3310.56

RANUNI Function

What is it?

A function that returns a pseudo-random number generated from the uniform (0,1) distribution.

Syntax

RANUNI*(seed*)

Notes

Seed can be any integer less than 2^(31) – 1 and is the initial starting point for the series of numbers generated by the function. The time on the computer clock is used as the seed if a non-positive integer is supplied or the value is left blank.

As an example, we expect RANUNI to give us a number between 0.25 and 0.5 approximately 25% of the time.

Simple Random Sampling

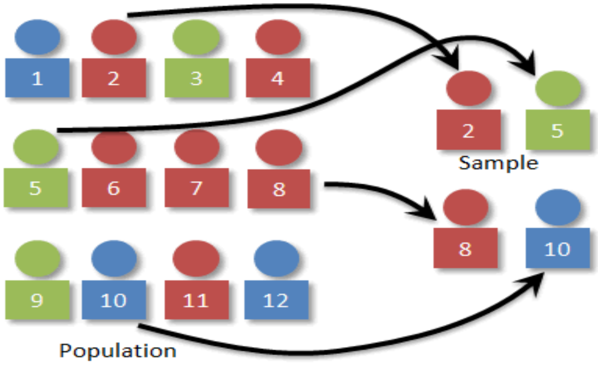
We'll do this in two ways:

1. Sample randomly a percentage of observations from the large dataset (10%)

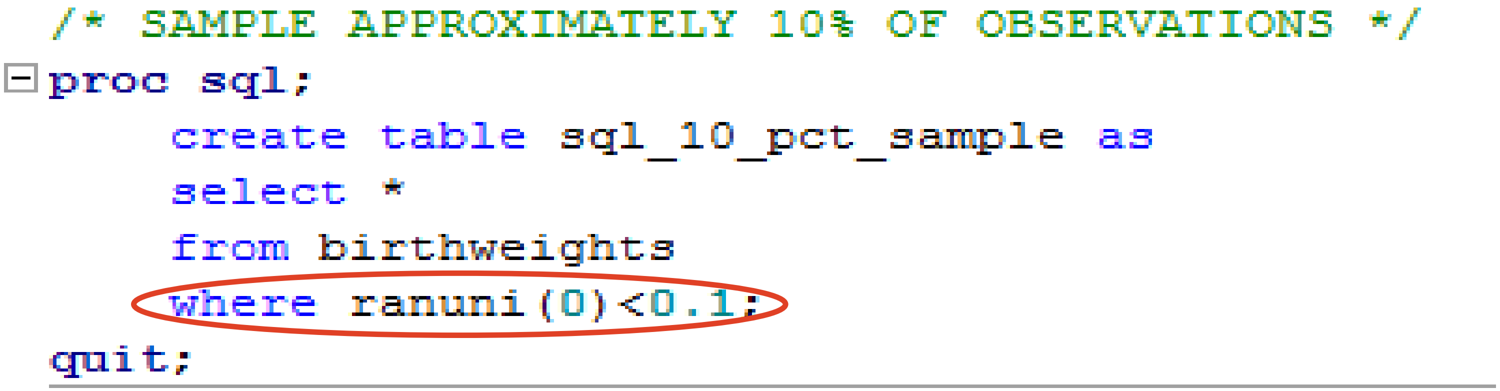
2. Sample randomly a fixed number of observations from the large dataset (5,000)

In our case we know that both should give us about the sample size we want because we know the actual number of observations in the population.

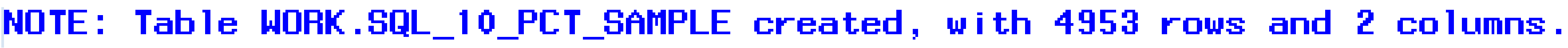
Method (1) is very useful when we don't know on hand the observation count of the large dataset, but we know what proportion of observations we'd like to sample.



Simple Random Sampling a % of the Population: PROC SQL



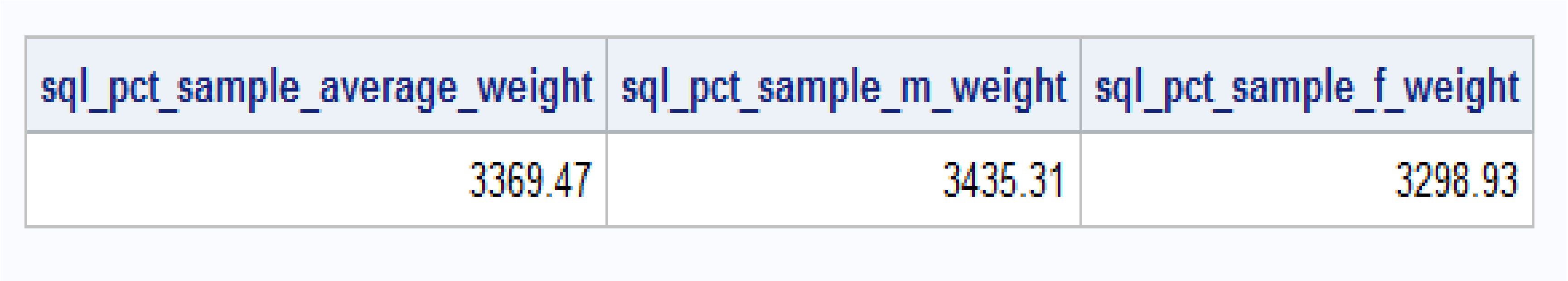
```
/* SAMPLE APPROXIMATELY 10% OF OBSERVATIONS */
proc sql;
    create table sql_10_pct_sample as
    select *
    from birthweights
    where ranuni(0)<0.1;
quit;
```



```
NOTE: Table WORK.SQL_10_PCT_SAMPLE created, with 4953 rows and 2 columns.
```

Each time a record is considered for selection a random number between 0 and 1 is generated and if it falls in the range (0,0.1) the record is selected.

Simple Random Sampling a % of the Population: PROC SQL

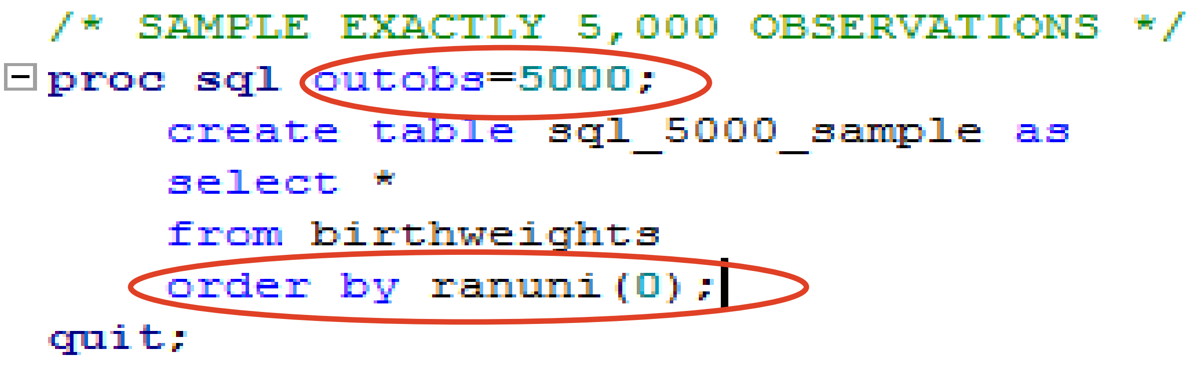


sql_pct_sample_average_weight	sql_pct_sample_m_weight	sql_pct_sample_f_weight
3369.47	3435.31	3298.93

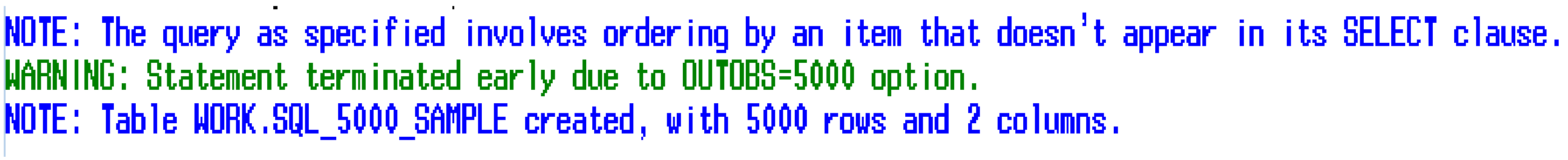
Actual Average Weight	Actual Average Male Weight	Actual Average Female Weight
3370.76	3427.25	3310.56

Simple Random Sampling a Fixed Number of Observations: PROC SQL

We use the OUTOBS and ORDERBY statements to sample an exact amount of observations from our large dataset.

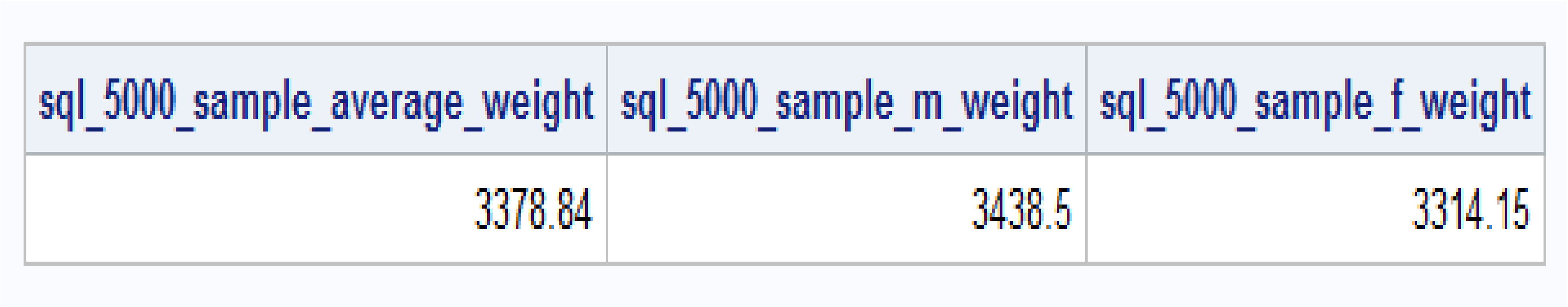


```
 /* SAMPLE EXACTLY 5,000 OBSERVATIONS */
 proc sql outobs=5000;
     create table sql_5000_sample as
     select *
     from birthweights
     order by ranuni(0);
 quit;
```



```
NOTE: The query as specified involves ordering by an item that doesn't appear in its SELECT clause.
WARNING: Statement terminated early due to OUTOBS=5000 option.
NOTE: Table WORK.SQL_5000_SAMPLE created, with 5000 rows and 2 columns.
```

Simple Random Sampling a Fixed Number of Observations: PROC SQL



sql_5000_sample_average_weight	sql_5000_sample_m_weight	sql_5000_sample_f_weight
3378.84	3438.5	3314.15

Actual Average Weight	Actual Average Male Weight	Actual Average Female Weight
3370.76	3427.25	3310.56

The SURVEYSELECT Procedure

What is it?

A procedure that provides a variety of methods for choosing probability-based random samples, including simple random sampling, stratified random sampling, and systematic random sampling.

Syntax

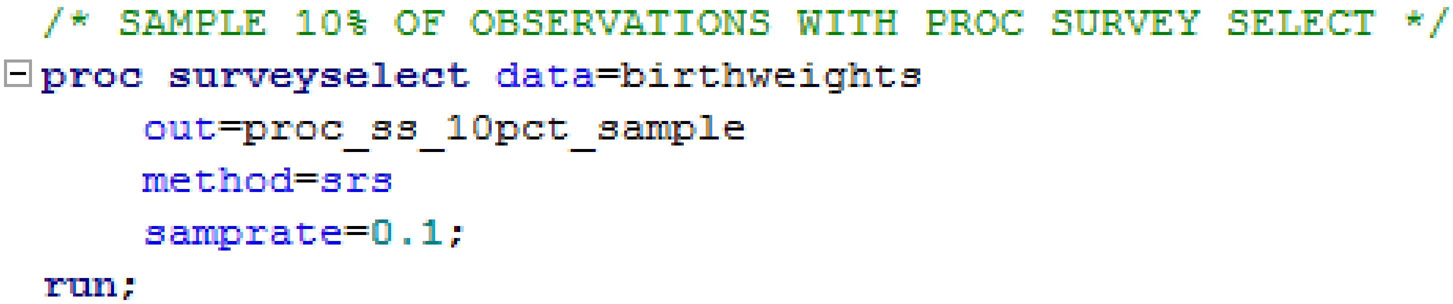
```
PROC SURVEYSELECT options ;
                optional statements;
RUN;
```

Notes

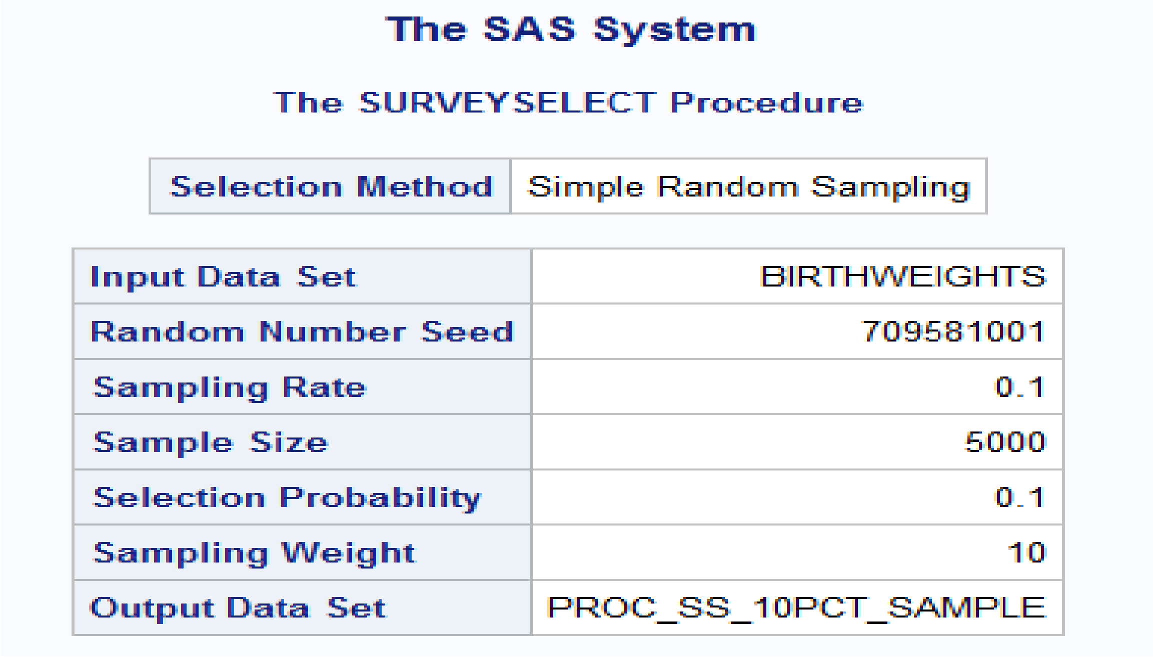
Some of the options we will utilize in the PROC SURVEYSELECT statement are:

1. DATA=, the input dataset
2. OUT=, the output dataset
3. METHOD=, the selection method (SRS is default if not specified)
4. SAMPSIZE=, the number of observations to select for the sample
5. SAMPRATE=, the proportion of observations to select for the sample

Simple Random Sampling a % of the Population: PROC SURVEYSELECT



```
/* SAMPLE 10% OF OBSERVATIONS WITH PROC SURVEY SELECT */
proc surveyselect data=birthweights
     out=proc_ss_10pct_sample
     method=srs
     samprate=0.1;
run;
```



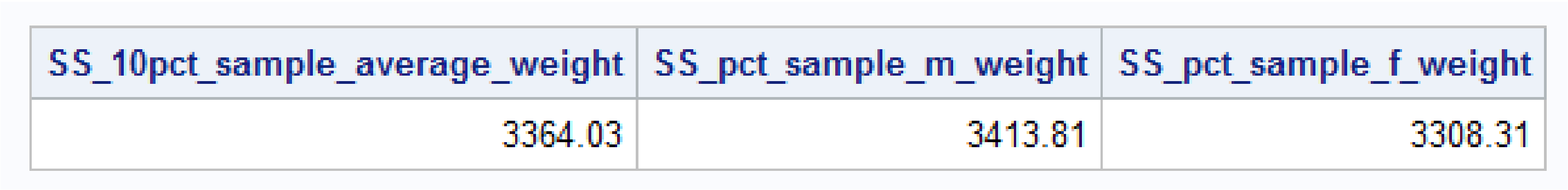
The SAS System

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	BIRTHWEIGHTS
Random Number Seed	709581001
Sampling Rate	0.1
Sample Size	5000
Selection Probability	0.1
Sampling Weight	10
Output Data Set	PROC_SS_10PCT_SAMPLE

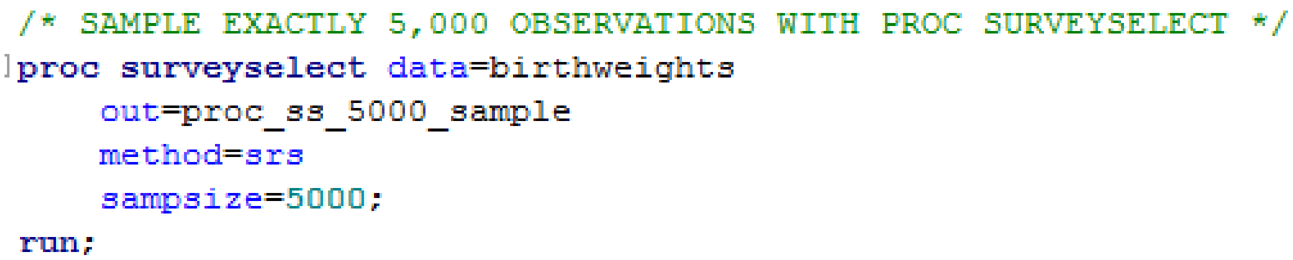
Simple Random Sampling a % of the Population: PROC SURVEYSELECT



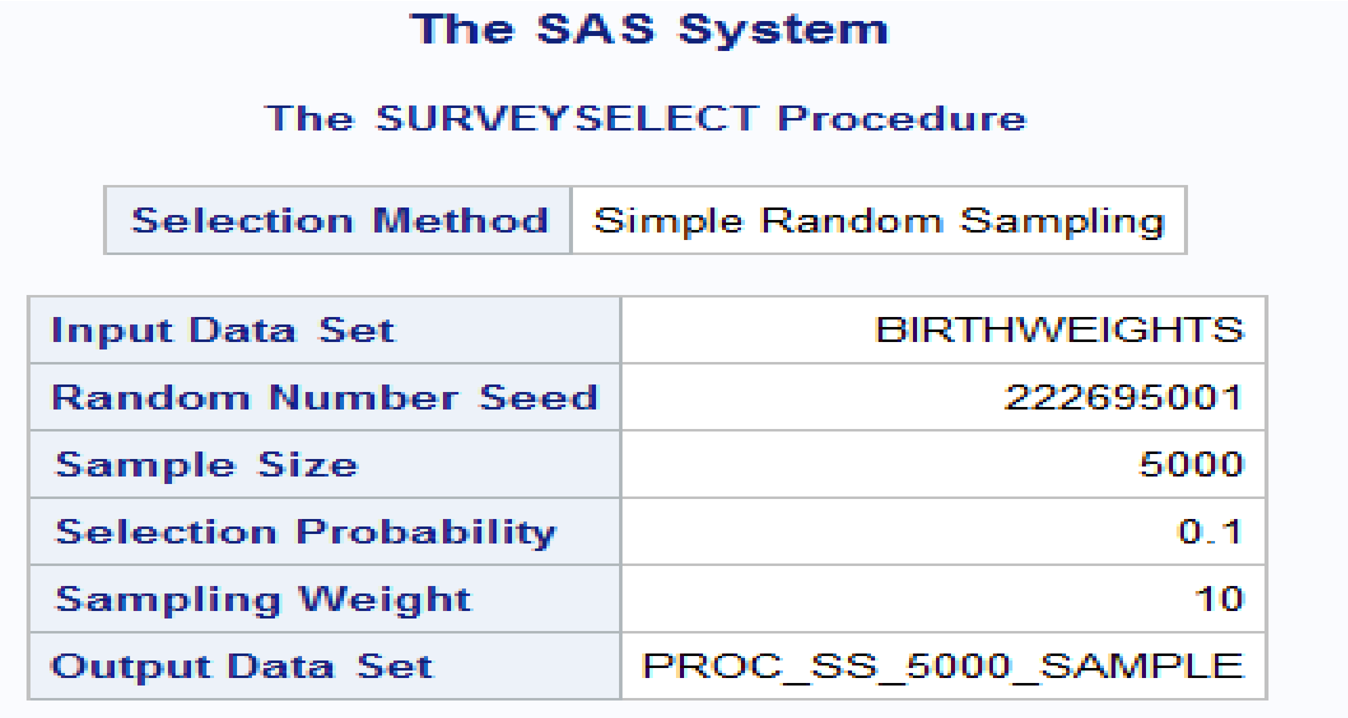
SS_10pct_sample_average_weight	SS_pct_sample_m_weight	SS_pct_sample_f_weight
3364.03	3413.81	3308.31

Actual Average Weight	Actual Average Male Weight	Actual Average Female Weight
3370.76	3427.25	3310.56

Simple Random Sampling a Fixed Number of Observations: PROC SURVEYSELECT



```
/* SAMPLE EXACTLY 5,000 OBSERVATIONS WITH PROC SURVEYSELECT */
proc surveyselect data=birthweights
    out=proc_ss_5000_sample
    method=srs
    sampsize=5000;
run;
```



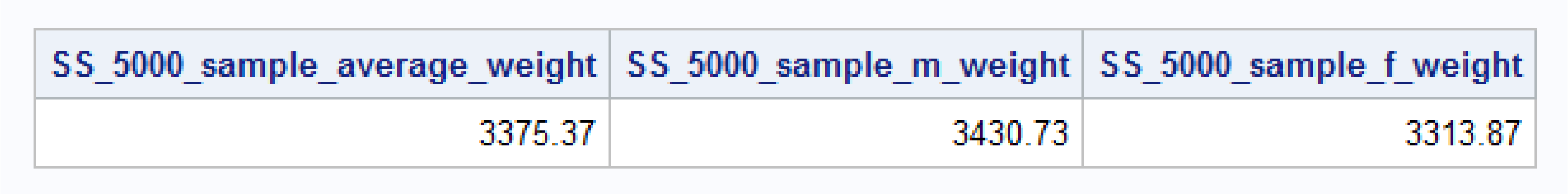
```
                    The SAS System

             The SURVEYSELECT Procedure

     | Selection Method | Simple Random Sampling |

| Input Data Set           |         BIRTHWEIGHTS |
| Random Number Seed       |            222695001 |
| Sample Size              |                 5000 |
| Selection Probability    |                  0.1 |
| Sampling Weight          |                   10 |
| Output Data Set          | PROC_SS_5000_SAMPLE  |
```

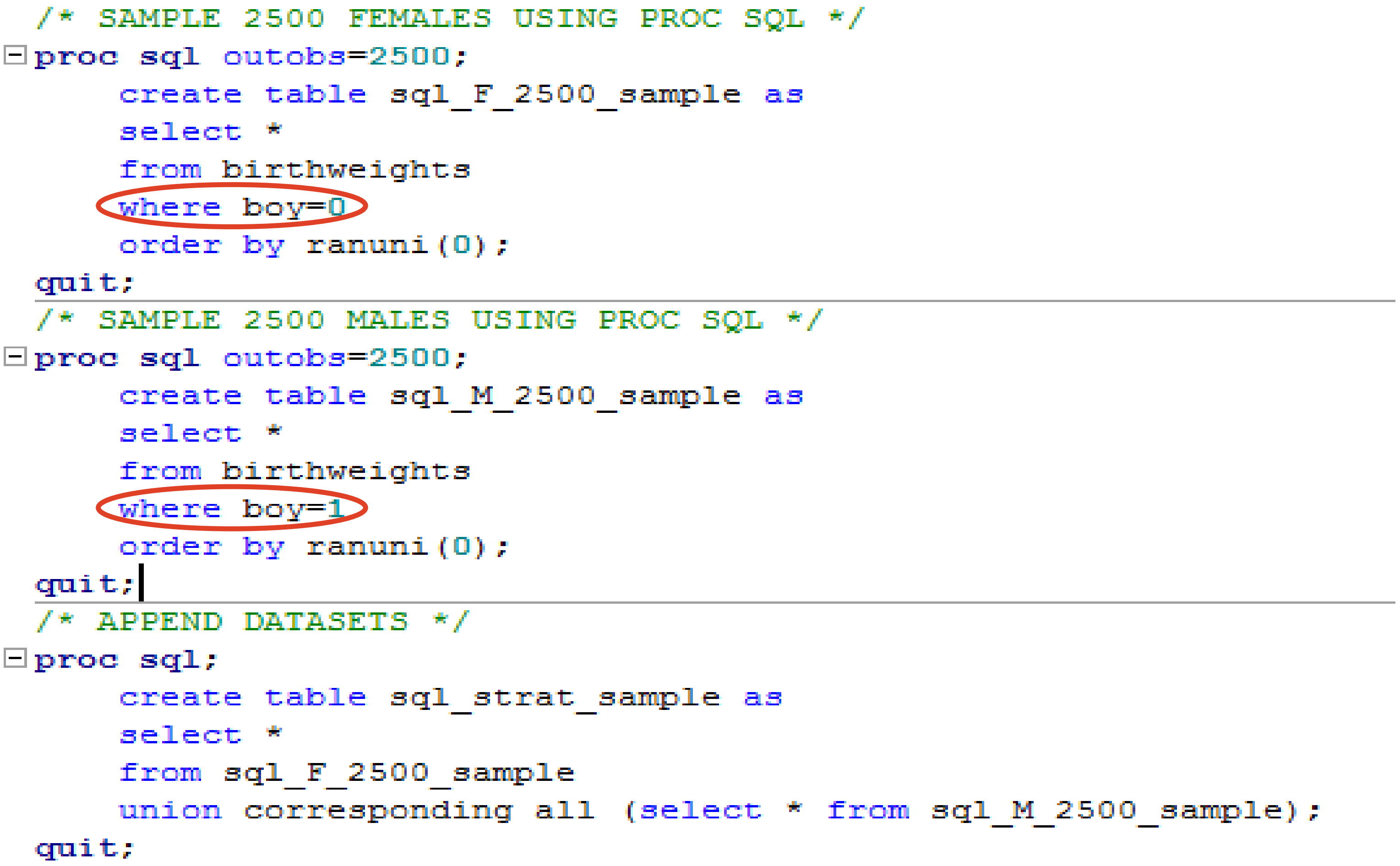
Simple Random Sampling a Fixed Number of Observations: PROC SURVEYSELECT



SS_5000_sample_average_weight	SS_5000_sample_m_weight	SS_5000_sample_f_weight
3375.37	3430.73	3313.87

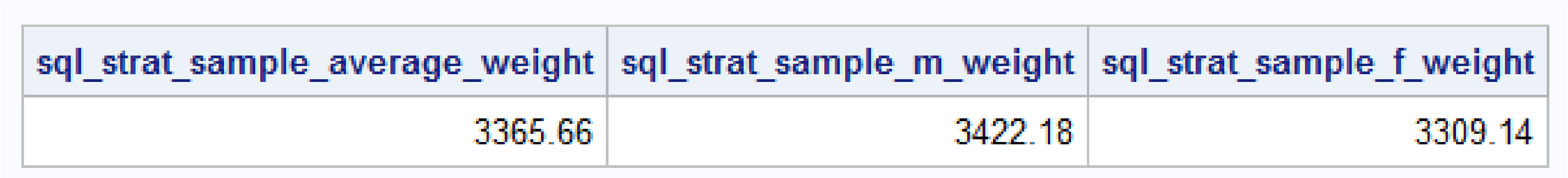
Actual Average Weight	Actual Average Male Weight	Actual Average Female Weight
3370.76	3427.25	3310.56

Stratified Random Sampling : PROC SQL



```
/* SAMPLE 2500 FEMALES USING PROC SQL */
proc sql outobs=2500;
    create table sql_F_2500_sample as
    select *
    from birthweights
    where boy=0
    order by ranuni(0);
quit;
/* SAMPLE 2500 MALES USING PROC SQL */
proc sql outobs=2500;
    create table sql_M_2500_sample as
    select *
    from birthweights
    where boy=1
    order by ranuni(0);
quit;
/* APPEND DATASETS */
proc sql;
    create table sql_strat_sample as
    select *
    from sql_F_2500_sample
    union corresponding all (select * from sql_M_2500_sample);
quit;
```

Stratified Random Sampling : PROC SQL



sql_strat_sample_average_weight	sql_strat_sample_m_weight	sql_strat_sample_f_weight
3365.66	3422.18	3309.14

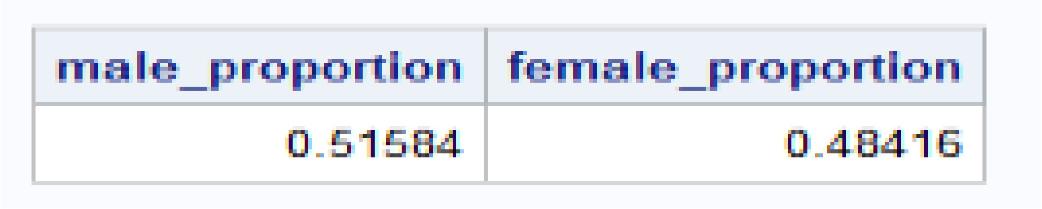
Actual Average Weight	Actual Average Male Weight	Actual Average Female Weight
3370.76	3427.25	3310.56

Stratified Random Sampling : PROC SQL

!

Q: What is the potential problem with what we’ve done here?

We sampled an equal amount from each strata and/or assumed that the population is 50/50.



male_proportion	female_proportion
0.51584	0.48416

In this case it is a pretty reasonable assumption, but in general we cannot just sample equal amounts from each strata and assume it is representative of the population.

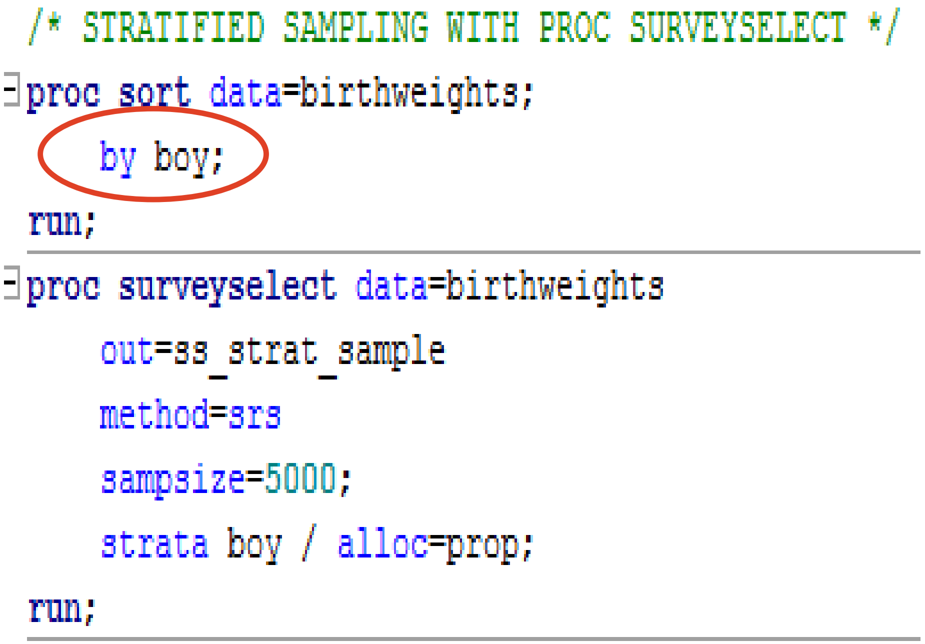
Examples:

1. Estimating average credit card balance in Canada, stratifying by province
2. Estimating the average number of hours worked per week in a company, stratifying by department

Stratified Random Sampling with Proportional Allocation: PROC SURVEYSELECT

PROPORTIONAL ALLOCATION allocates the total sample size amongst the strata using their proportion in the actual population, improving representativeness

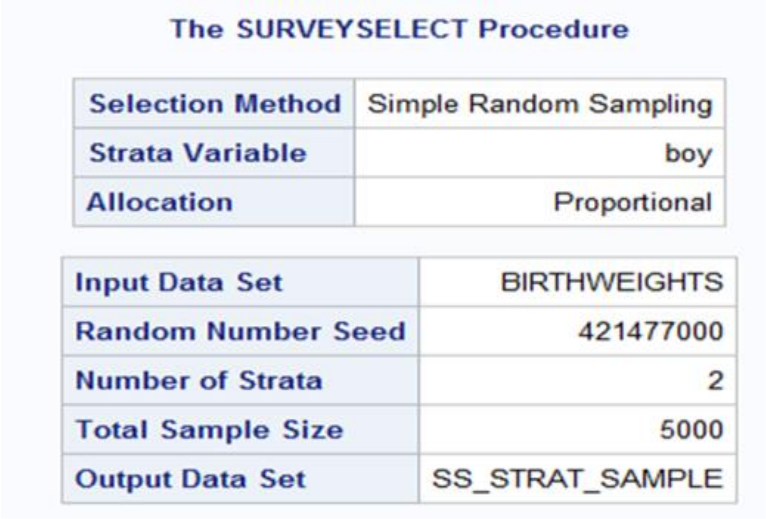
In our case, based on the true proportion of males and females in the population, for a sample of 5000 we should select 2579 males and 2421 females.



```
/* STRATIFIED SAMPLING WITH PROC SURVEYSELECT */
proc sort data=birthweights;
    by boy;
run;
proc surveyselect data=birthweights
    out=ss_strat_sample
    method=srs
    sampsize=5000;
    strata boy / alloc=prop;
run;
```

Quirk Alert!
PROC SURVEYSELECT expects the dataset to be sorted by the strata variable(s).

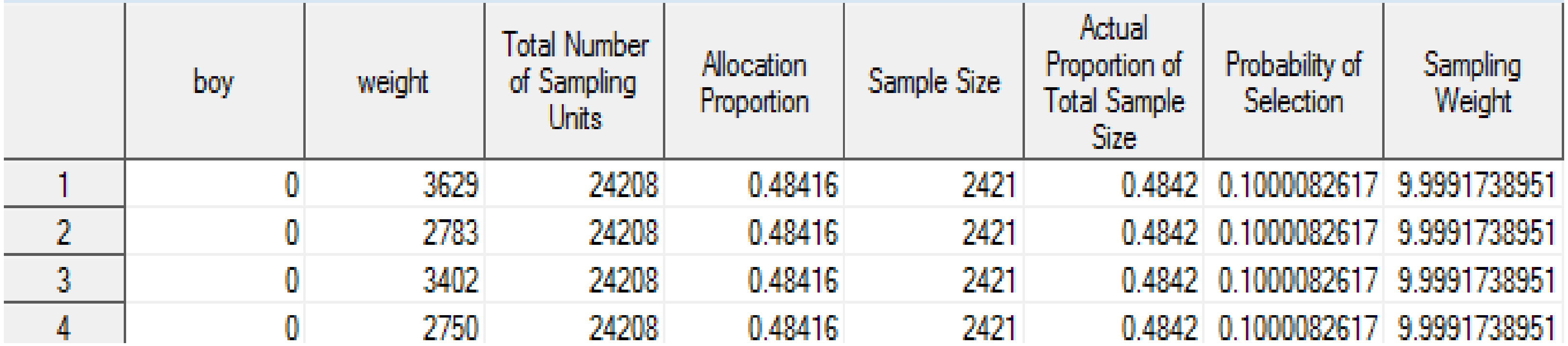
Stratified Random Sampling with Proportional Allocation: PROC SURVEYSELECT



The SURVEYSELECT Procedure

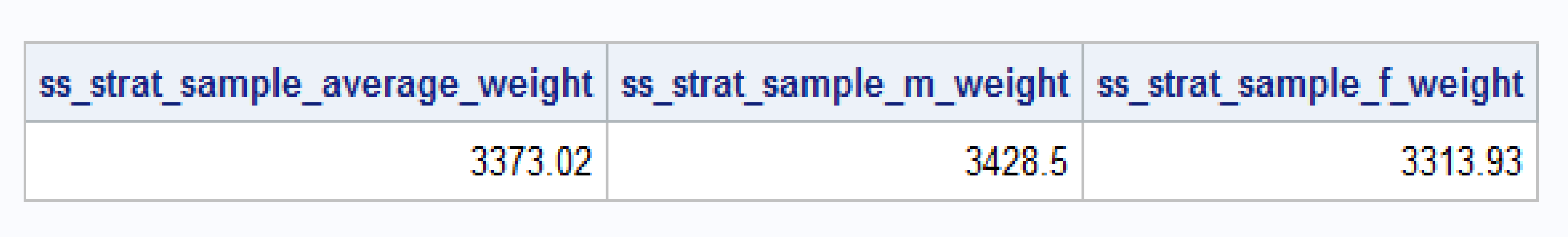
Selection Method	Simple Random Sampling
Strata Variable	boy
Allocation	Proportional

Input Data Set	BIRTHWEIGHTS
Random Number Seed	421477000
Number of Strata	2
Total Sample Size	5000
Output Data Set	SS_STRAT_SAMPLE



	boy	weight	Total Number of Sampling Units	Allocation Proportion	Sample Size	Actual Proportion of Total Sample Size	Probability of Selection	Sampling Weight
1	0	3629	24208	0.48416	2421	0.4842	0.1000082617	9.9991738951
2	0	2783	24208	0.48416	2421	0.4842	0.1000082617	9.9991738951
3	0	3402	24208	0.48416	2421	0.4842	0.1000082617	9.9991738951
4	0	2750	24208	0.48416	2421	0.4842	0.1000082617	9.9991738951

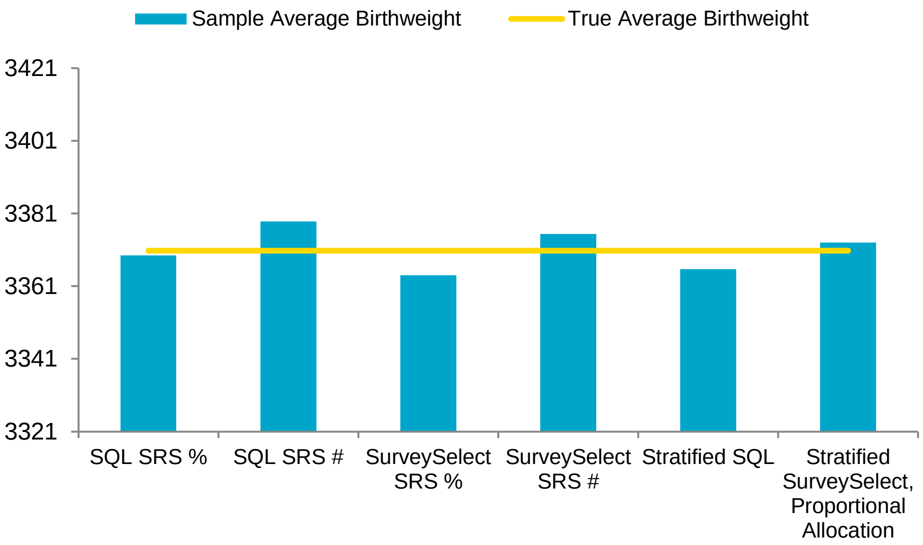
Stratified Random Sampling with Proportional Allocation: PROC SURVEYSELECT

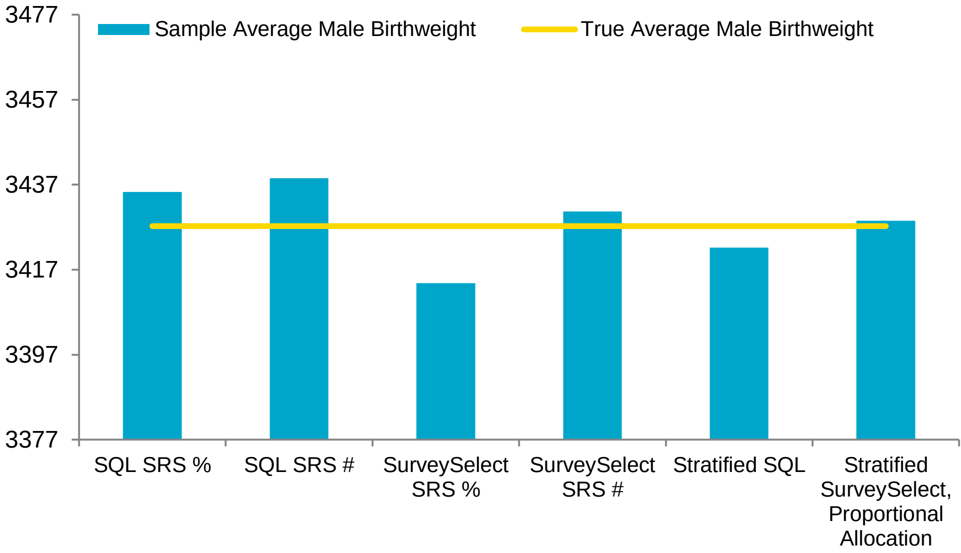


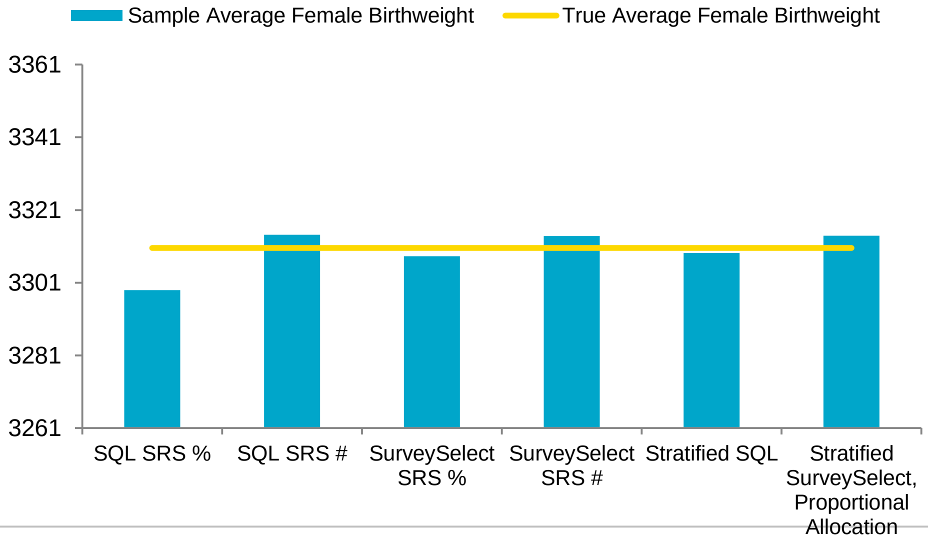
ss_strat_sample_average_weight	ss_strat_sample_m_weight	ss_strat_sample_f_weight
3373.02	3428.5	3313.93

Actual Average Weight	Actual Average Male Weight	Actual Average Female Weight
3370.76	3427.25	3310.56

Sampling Results vs. Actual results—How Close Were We?







Comparison of SAS Procedures for Sampling

PROC SQL	PROC SURVEYSELECT
Pros - Procedure is very familiar to most users - Possible to sample directly from your database	**Pros** - Can sample an exact % of the population even if you don't know the population size - Has built in sampling methods
Cons - Not always possible to sample exact proportion of the population - Doesn't have built in sampling methods - Proportional allocation cannot be easily done	**Cons** - Cannot sample directly from your database - Need to sort large dataset before stratifying - May be a new procedure for many users

Thank You for Listening!

Advanced Analytics Intern
TransUnion
May 2015 – December 2015
Email: mardizz@transunion.com
Phone: 905-340-1000 ext. 2049

Honours Actuarial and Financial Mathematics Co-op, Level V
McMaster University
Graduation Date: April 2016
Email: ardizzm@mcmaster.ca

Thank you to the TransUnion Advanced Analytics Team for their contributions to this presentation!

Q&A

??

References

Richard, Severino. "Getting Your Random Sample in Proc SQL." Accessed October 16, 2015. http://www2.sas.com/proceedings/sugi31/168-31.pdf.

"The SURVEYSELECT Procedure." SAS/STAT(R) 9.2 User's Guide, Second Edition. Accessed October 20, 2015. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveyselect_sect001.htm.

Tortora, Cristina. "Probability Samples." Lecture, McMaster University, Hamilton, Ontario, Fall 2014.

Tortora, Cristina. “Stratified Sampling." Lecture, McMaster University, Hamilton, Ontario, Fall 2014.

“Why Sample?" QMSS E-Lessons. Accessed October 5, 2015. http://ccnmtl.columbia.edu/projects/qmss/samples_and_sampling/why_sample.html.

Appendix

Sample	Difference from True Average Weight	Difference from True Average Male Weight	Difference from True Average Female Weight
SQL SRS %	-1.29	+8.06	-11.63
SQL SRS %	+8.08	+11.25	+3.59
SurveySelect SRS %	-6.73	-13.44	-2.25
SurveySelect SRS #	+4.61	+3.48	+3.31
SQL Stratified	-5.10	-5.07	-1.42
SurveySelect Stratified, Optimal Allocation	+2.26	+1.25	+3.37