STA 144 Term Project

06.04.2019

Danli Zheng

Qian Yang

Summary

Through the reading, we learned about the problems of finite population inference and the relationship between design and inference, we also learned about the Neyman revolution, which relative to this quarter's lectures. In this quarter we learned SRS (simple random sample) which is a model-based method and stratified random sample. Based on the idea of inference we learned in the article and knowledge we learned in this lecture. In this project, we use both methods to get our NHANES dataset sample mean, sample variance, standard error, and 95% confidence intervals, and we analyze the results between these two methods.

Introduction

This is a project to utilize all the sampling methods we have learned in class and in the reading to analyze the dataset. There are two main parts of this report. The first part is a summary and discussion of the paper "The Foundations of Survey Sampling: a Review". The second part is the analysis of the dataset using the traditional methods of analysis and using sample survey methods.

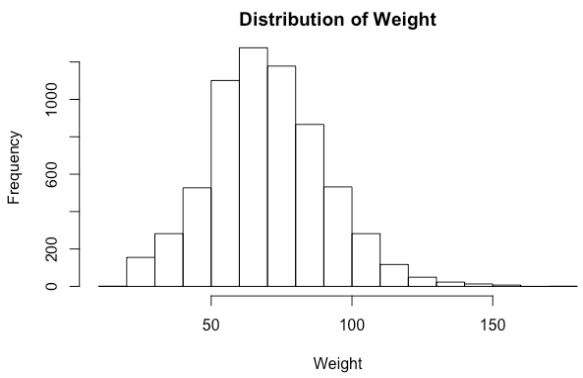
Discussion of part 1 of project

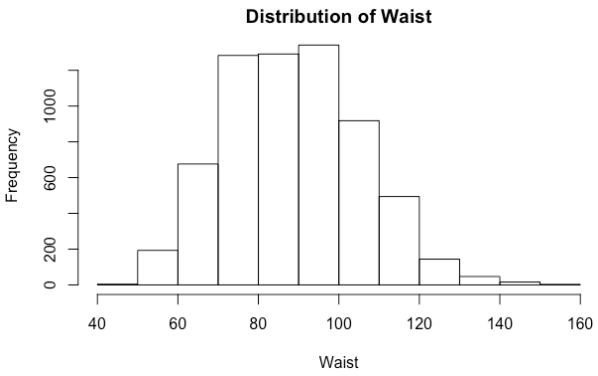
The paper, "The Foundations of Survey Sampling: a Review", written by T.M.F. Smith discusses the historical development of the twin problems of survey design and of finite population inference and the relationship between design and inference. In fact, almost all of the traditional sampling inferences are based on Neyman's fundamental paper. One of the inferences that Neyman introduced is the confidence interval but with a shortcut of consistency for any form of random sampling. In order to improve the efficiency of the sampling, Neyman derives minimum variance unbiased estimators. Within this method, stratified sampling would be more efficient than simple random sampling. Neyman also addressed the idea of optimal allocation of the stratified sampling and provided a framework of the cluster sampling. But all the sampling methods have no best estimators speaking of the efficiency. Alternative approaches are found, one is to ignore some of the data, and another one is the superpopulation argument. The superpopulation model is an assumption that the finite population has been generated as a random sample from an infinite super population. Both of the methods can improve the efficiency of the sampling.

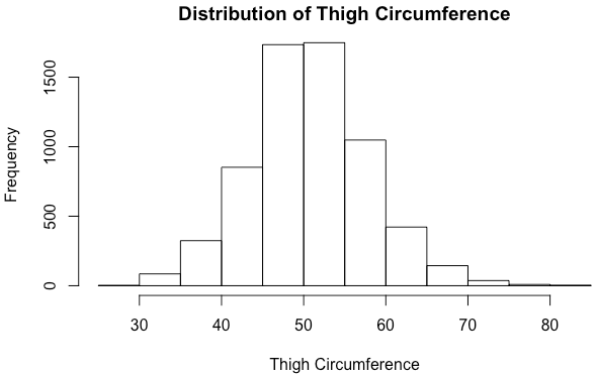
Analysis of NHANES data

Population Estimation

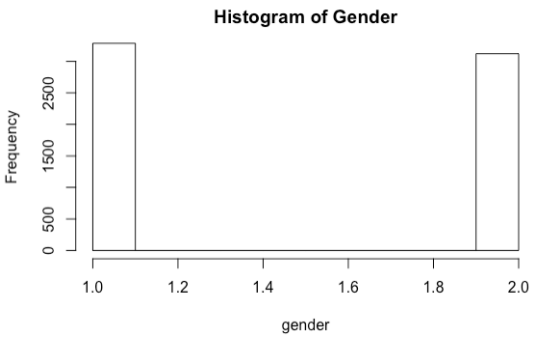
We analyze variables weight, body mass index, waist circumference, and thigh circumference in the dataset that we are given. We plot histogram for variables to check the normal distribution. Apparently, we learned that our variable is normally distributed from our plots.

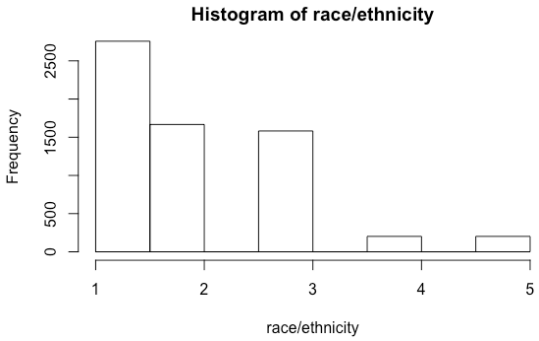


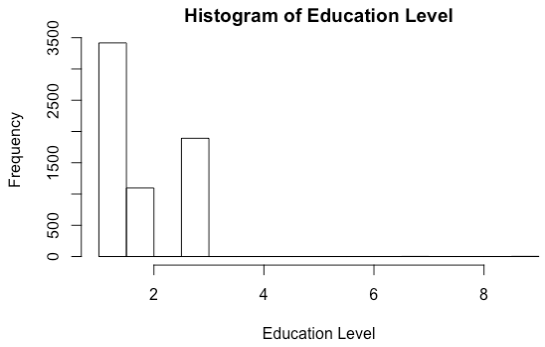


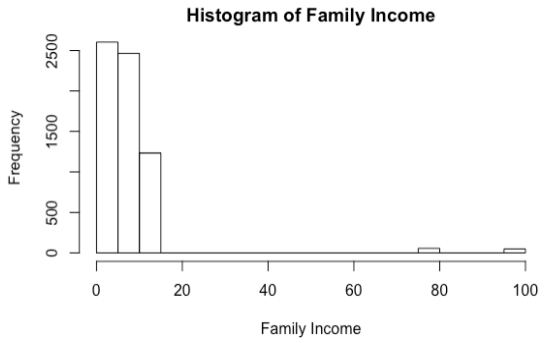


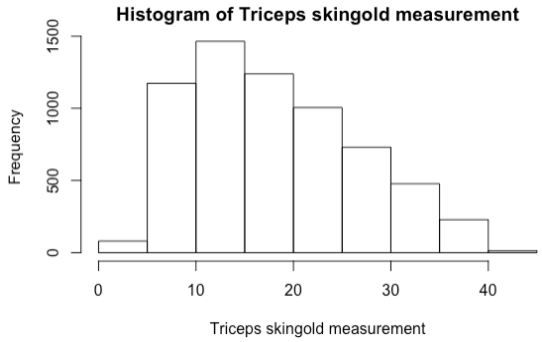
Other Population Estimations

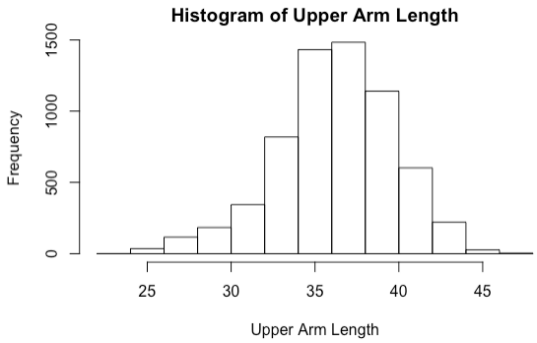





Histogram of Education Level
Frequency
0
500
1500
2500
3500
2
4
6
8
Education Level


Histogram of Family Income
Frequency
0
500
1500
2500
0
20
40
60
80
100
Family Income


Histogram of Triceps skingold measurement
Frequency
0
500
1000
1500
0
10
20
30
40
Triceps skingold measurement


Histogram of Upper Arm Length
Frequency
0
500
1000
1500
25
30
35
40
45
Upper Arm Length

SRS

Then, we take a simple random sample of 100 from the population and calculate the sample mean, sample variance, estimated variance of the sample, estimated total, and estimated variance of sample total.

	Weight	Body Mass Index	Waist circumference	Thigh circumference
Sample mean	70.882	26.2482	90.128	51.495
Sample variance	416.9831	38.84673	278.2443	50.83927
Estimated standard error of the sample mean	2.026	0.6184	1.655	0.7074
Estimated total	454353.6	168251	577720.5	330083
Estimated variance of sample total	168657574	15712377	112541734	20563009
Estimated 95% CI for the mean	(66.91, 74.85)	(25.03615, 27.46025)	(86.884, 93.372)	(50.1084, 52.8816)

Stratified Random Sample

We divided our dataset into three strata, age zero to fourteen, age fifteen to thirty-five and more than thirty-six. First of all, we do equal allocation for each stratum, we take ten percent stratified sample. The sample size of each stratum is the same, which is around 214. We use a mathematical method to find the mean, total, variance, and standard error of each variable in each stratum, and then calculate the confidence interval for the estimated population means. After the calculation for each stratum, we utilize all the data collected from the stratum to estimate the population mean, total, standard error, and confidence interval of the mean of each variable.

Stratum 1: age 0-14

	Weight	Body Mass Index	Waist circumference	Thigh circumference
Sample mean	47.00467	20.5113	72.3393	44.793
Sample total	10059	4389.42	15480.6	9585.7
Sample variance	220.414	20.7458	162.5115	53.7606
Estimated SE	0.02574	0.002423	0.01898	0.00628
Estimated 95% CI for the mean	(46.95422, 47.05513)	(20.5066, 20.5166)	(72.3021, 72.3765)	(44.7807, 44.8053)

Stratum 2: 15-35

	Weight	Body Mass Index	Waist circumference	Thigh circumference
Sample mean	71.4603	24.9638	85.4336	52.7561
Sample total	15292.5	5342.25	18282.8	11289.8
Sample variance	249.9938	22.6386	160.9877	37.135
Estimated SE	0.1407	0.0127	0.09059	0.0209
Estimated 95% CI for the mean	(71.1846, 71.736)	(24.9388, 24.9888)	(85.25608, 85.6112)	(52.7151, 52.797)

Stratum 3: 36+

	Weight	Body Mass Index	Waist circumference	Thigh circumference
Sample mean	77.0509	27.3605	98.0893	50.6991
Sample total	16488.9	5855.14	20991.1	10849.6
Sample variance	254.1547	21.2151	156.5289	34.1145
Estimated SE	0.179	0.0149	0.1103	0.024
Estimated 95% CI for the mean	(76.7001, 77.4018)	(27.3312, 27.3898)	(97.8732, 98.3054)	(50.652, 50.746)

Population estimation

	Weight	Body Mass Index	Waist circumference	Thigh circumference
Estimated sample total	41840.4	15586.81	54754.5	31725.1
Estimate mean	65.172	24.2785	85.2874	49.416
Estimated variance of sample mean	0.3874	0.03376	0.2323	0.05742

Estimated SE	0.6224	0.1837	0.2396	0.2396
Estimated 95% CI for the mean	(63.9521, 66.39182)	(23.91839, 24.63865)	(84.3427, 86.2321)	(48.94539, 49.8856)

Conclusions

In this project, we mainly focus on using inference we learned in classes and the article. we preprocess our dataset, we remove missing values from our datasets, because they affect our results, and then through histograms, we know our dataset is normally distributed.

We use SRS because it is the simplest form of probability sample, and we know that we can use stratified sampling to prevents samples would produce bad estimates and increase the precision of the estimators. Through our results, we notice that the estimate standard errors from stratified sampling method are lower than a simple random sample. Therefore, we conclude for our dataset, the stratified sampling method increases the precision of the estimation.