EPID 642: Sampling and Power **Stephanie Mecham**
Homework 1
Total points: 25

Due: Feb. 6, 2018 at 10:00 am

Statement of task: As a group of two or three, you are asked to sample using a probability sampling method and answer the questions below. **Note that for questions 5, 6, 7, and 9, please provide your SAS codes and outputs.**

Synopsis:
Suppose that you are a chief epidemiologist in the Arbor County Public Health Department.

A physician with an interest in diabetes at your local hospital sends you a copy of the following paper and wishes to know if the finding can be used to target a sub-population susceptible to diabetes in relation to sedentary lifestyle in the area.

George et al., Chronic disease and sitting time in middle-aged Australian males: findings from the 45 and Up Study. International Journal of Behavioral Nutrition and Physical Activity 2013;10:20.

1. What was the study design and what were the objectives and hypotheses of the study by *George et al.*? (2pts)

This is a cross-sectional study whose aim was to build on the existing knowledge about the association between sitting-time/sedentary lifestyles and the development of chronic diseases in an understudied population (middle-aged Australian males). The researchers' hypotheses were that people with higher reported sitting-times would have higher rates of chronic diseases compared with those reporting less-sitting time, independently of BMI or age.

2. What were the inclusion and exclusion criteria? To answer this question, please specify the source population (i.e. sampling frame), eligible sample, and describe how study participants (i.e. actual sample) were selected. (3pts)

The source population/sampling frame was the 267,153 participants from the 45 and Up study. The eligible sample was all adults 45 years of age or older who were currently residing in New South Wales state of Australia from the source population (the eligible sample consisted of 70,416 individuals). The inclusion criteria for this study was that the participants had to have filled out a mailed baseline questionnaire and provided signed consent. Participants were excluded if they were missing/had invalid data for 2 or more of the physical activity categories. The actual sample consisted of 63,048 individuals.

A total of 10,018 adults aged 20 and older live in your area that has 10 census blocks. To identify a susceptible sub-population, you decided to conduct a preliminary study with a sample of **500** adults (including all age groups and both males and females).

3. Which sampling method would you choose? Please describe your sampling strategies and explain why you would like to use this method? (3pts)

I would choose a stratified random sampling method, where I would apply random sampling techniques among each census block separately. Advantages of stratified random sampling are that it is a good way to minimize selection bias and ensure equal representation of different groups. It's good for smaller sample sizes (like ours) and allows us to see the differences between strata and potentially identify the most at-risk groups.

4. What are the potential limitations of this method? (2pts)

Stratified random sampling may not be very helpful if the among-strata groups are heterogenous. Additionally, if there is a population pattern of some other variable that gives rise to these specific strata, this could lead to systematic bias in our sample. It also may inadvertently exclude people who are homeless or otherwise escape classification into a specific census block.

Assume that after you had sampled, you conducted a survey and collected information for your study including diabetes, the total amount of time, in hours, spending for watching TV or using a computer per day (hr_tvcomp) and other important covariates.

5. What is the prevalence of diabetes in your sample? (1.5pts)

The prevalence of diabetes in our sample is 12.83%.

The SAS System

The FREQ Procedure

self-reported physician diagnosis of DM				
dm	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	435	87.17	435	87.17
1	64	12.83	499	100.00
Frequency Missing = 1				

6. What is the mean and standard error of the variable 'hr_tvcomp' in your sample? (1.5pts)

The mean is 2.968 hours of TV watched or computer used per day, and the standard error is 0.0913.

The SAS System

The MEANS Procedure

Analysis Variable : hr_tvcomp Hours watch TV or use computer per day over past 30 days		
N	Mean	Std Error
500	2.9680000	0.0913380

7. What is the association between 'hr_tvcomp' and diabetes in your sample? Please provide measures of association and 95% confidence intervals. To answer this question, you must categorize 'hr_tvcomp' into quartiles* and compare the lowest quartile with the upper three quartiles. Also, in order to account for confounding effects, please control for age, BMI, gender, race-ethnicity, and poverty-to-income ratio in the model. (3.5pt)

   The association between hr_tvcomp and diabetes in my sample appears to get stronger as you go up in quartiles of hr_tvcomp . With the first quartile as the reference group, the second quartile has a β_{x1} of -0.0413 (-0.1241, 0.0416) in the model, the third has a β_{x2} of 0.000346 (-0.0780, 0.0787), and the highest quartile has a of β_{x3} of 0.0578 (-.0272, 0.143). The second quartile has a slight negative linear association with diabetes, the third quartile has a very slight positive linear association with diabetes (but for all practical purposes is essentially no association), and the fourth quartile has a very slight positive linear association with diabetes. The hr_tvcomp quartiles do not appear to have a statistically significant association with diabetes based on the results of our sample. The P-value for the second quartile was 0.3286, third was 0.9931, and fourth was 0.1820. These are all higher than our alpha threshold for statistical significance of 0.05. As a whole, the hr_tvcomp quartiles resulted in a P value of 0.0514 for type I sum of squares and 0.1536 for type III sum of squares. Neither of these are below our significance threshold either.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-.2264410905	B	0.10736402	-2.11	0.0355	-.4374234127	-.0154587683
tech_quart 1	-.0412507280	B	0.04217784	-0.98	0.3286	-.1241349119	0.0416334560
tech_quart 2	0.0003463643	B	0.03987685	0.01	0.9931	-.0780161092	0.0787088378
tech_quart 3	0.0578436874	B	0.04327283	1.34	0.1820	-.0271922742	0.1428796490
tech_quart 0	0.0000000000	B	.	.	.	.	.
age	0.0046422968		0.00074469	6.23	<.0001	0.0031789034	0.0061056901
bmi	0.0111512567		0.00229140	4.87	<.0001	0.0066483914	0.0156541221
gender 1	-.0118644718	B	0.02915890	-0.41	0.6843	-.0691649772	0.0454360336
gender 2	0.0000000000	B	.	.	.	.	.
raceeth 1	-.1716647234	B	0.08500355	-2.02	0.0440	-.3387062292	-.0046232177
raceeth 2	-.1905250971	B	0.10627099	-1.79	0.0737	-.3993595062	0.0183093121
raceeth 3	-.1747164324	B	0.08213324	-2.13	0.0339	-.3361174424	-.0133154224
raceeth 4	-.1407018145	B	0.08463378	-1.66	0.0971	-.3070166718	0.0256130428
raceeth 5	0.0000000000	B	.	.	.	.	.
lowpir 0	-.0258597094	B	0.03926892	-0.66	0.5105	-.1030275390	0.0513081203
lowpir 1	0.0000000000	B	.	.	.	.	.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tech_quart	3	0.76387505	0.25462502	2.60	0.0514
age	1	4.43924175	4.43924175	45.38	<.0001
bmi	1	2.15609116	2.15609116	22.04	<.0001
gender	1	0.04429746	0.04429746	0.45	0.5013
raceeth	4	0.51190280	0.12797570	1.31	0.2659
lowpir	1	0.04241979	0.04241979	0.43	0.5105

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tech_quart	3	0.51705874	0.17235291	1.76	0.1536
age	1	3.80134654	3.80134654	38.86	<.0001
bmi	1	2.31666669	2.31666669	23.68	<.0001
gender	1	0.01619473	0.01619473	0.17	0.6843
raceeth	4	0.51640335	0.12910084	1.32	0.2616
lowpir	1	0.04241979	0.04241979	0.43	0.5105

8. Does your sample support the finding of George et al.? Why or why not? (2pts)

No, it does not, because my sample shows a non-significant association of the amount of time spent watching TV/using the computer and diabetes, at all quartile levels. George et.al, on the other hand, found a statistically significant positive linear association between sedentary behavior and diabetes diagnosis among people who reported sitting for 6 or more hours a day.

Assume that you conducted the survey in the entire population and collected the same information.

9. What is the association between 'hr_tvcomp' and diabetes in the population? To answer this question, you must categorize 'hr_tvcomp' into quartiles* and compare the lowest quartile with the upper three quartiles. Again, in order to account for confounding effects, please control for age, BMI, gender, race-ethnicity, and poverty-to-income ratio in the model. Please provide measures of association and 95% confidence intervals? (3.5pt)

   The association between hr_tvcomp and diabetes in the population appears to increase slightly as you go up the hr_tvcomp quantiles. However, these associations are slight. Compared to the first quartile of hr_tvcomp, the second quartile has a β_{x1} estimate of -.0038 (-0.1942, 0.0118), the third has a β_{x2} estimate of 0.0114 (-0.0033, 0.0261), and the fourth has a β_{x3} estimate of 0.0465 (0.0308, 0.0622). We can see that all of these are close to zero. The fourth quartile of hr_tvcomp is the only quartile that has a p-value that suggests statistical significance ($p < .0001$, which is below our alpha of 0.05). The second and third quartiles had p-values of 0.6341 and 0.1298, respectively, and thus are not statistically significant in the association with diabetes.

Parameter	Estimate		Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	-.2247555487	B	0.01924140	-11.68	<.0001	-.2624724123	-.1870386852
tech_quart 1	-.0037933370	B	0.00797021	-0.48	0.6341	-.0194164945	0.0118298205
tech_quart 2	0.0113592599	B	0.00749726	1.52	0.1298	-.0033368235	0.0260553433
tech_quart 3	0.0464860532	B	0.00802205	5.79	<.0001	0.0307612867	0.0622108196
tech_quart 0	0.0000000000	B	.	.	.	.	.
age	0.0039766238		0.00013823	28.77	<.0001	0.0037056685	0.0042475792
bmi	0.0059183342		0.00042175	14.03	<.0001	0.0050916251	0.0067450432
gender 1	0.0003537700	B	0.00548650	0.06	0.9486	-.0104008235	0.0111083636
gender 2	0.0000000000	B	.	.	.	.	.
raceeth 1	0.0121367286	B	0.01471644	0.82	0.4096	-.0167103461	0.0409838033
raceeth 2	-.0060264846	B	0.02055413	-0.29	0.7694	-.0463165671	0.0342635979
raceeth 3	-.0494489273	B	0.01398797	-3.54	0.0004	-.0768680494	-.0220298052
raceeth 4	-.0014508168	B	0.01463644	-0.10	0.9210	-.0301410681	0.0272394345
raceeth 5	0.0000000000	B	.	.	.	.	.
lowpir 0	-.0286743234	B	0.00695126	-4.13	<.0001	-.0423001475	-.0150484994
lowpir 1	0.0000000000	B	.	.	.	.	.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
tech_quart	3	6.12922853	2.04307618	26.54	<.0001
age	1	58.97290973	58.97290973	766.10	<.0001
bmi	1	17.19505603	17.19505603	223.38	<.0001
gender	1	0.00088668	0.00088668	0.01	0.9145
raceeth	4	8.61460584	2.15365146	27.98	<.0001
lowpir	1	1.30986600	1.30986600	17.02	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
tech_quart	3	3.66012593	1.22004198	15.85	<.0001
age	1	63.70883012	63.70883012	827.62	<.0001
bmi	1	15.15861798	15.15861798	196.92	<.0001
gender	1	0.00032005	0.00032005	0.00	0.9486
raceeth	4	6.93736451	1.73434113	22.53	<.0001
lowpir	1	1.30986600	1.30986600	17.02	<.0001

10. Is the finding in your sample consistent with that in the population? Briefly discuss consistency in the findings between your sample and the entire population in terms of random errors and systematic bias. (3pts)

    Yes, the findings in my sample were very close to the results in the population. The only notable difference was that in the population regression, the fourth quantile of hr_tvcomp was statistically significant whereas that was not the case in my sample. Overall the trend was similar though, with the association being slightly negative in the second quartile and increasing up to slightly positive in the fourth quartile. The slight difference between my sample data and the population data may be due simply to random error. Though stratifying by census block was meant to reduce selection bias, that does not eliminate random error, and the fourth quartile may have shown no significant association with diabetes simply by the individuals included in the sample purely by chance. Systematic bias could have occurred if the population was structured in a way that the populations of each census block followed some sort of pattern; that pattern then would have been reflected in the sampling and thus wouldn't be truly random.

    CODE:

```
*EPID 642 Homework 1 | Stephanie Mecham;

libname epi "C:\Users\smecham\Desktop\epi";
run;

*Sampling by Census Block;

proc sort data=epi.dm_sit_sec1;
by block;
run;

proc surveyselect data=epi.dm_sit_sec1 method=SRS
sampsize=50 seed=79589 out=epi.dm_samp;
strata block;
run;

*Question 5: Finding Diabetes Prevalence in Sample;

proc freq data=epi.dm_samp;
tables dm;
run;

*Question 6: Finding Sample Mean & Standard Error of HR_TVCOMP;

proc means data=epi.dm_samp n mean stderr;
var hr_tvcomp;
run;

*Question 7: Regression Analysis of HR_TVCOMP by Diabetes in Sample;
```

```
proc rank data= epi.dm_samp out=epi.recoded groups=4;
var hr_tvcomp;
ranks tech_quart;
run;

proc glm data=epi.recoded;
class tech_quart (ref='0') gender (ref='2') raceeth lowpir;
model dm = tech_quart age bmi gender raceeth lowpir / solution clparm;
run;

*Question 9: Regression Analysis of HR_TVCOMP by Diabetes in Population;

proc rank data= epi.dm_sit_sec1 out=epi.population groups=4;
var hr_tvcomp;
ranks tech_quart;
run;

proc glm data=epi.population;
class tech_quart (ref='0') gender (ref='2') raceeth lowpir;
model dm = tech_quart age bmi gender raceeth lowpir / solution clparm;
run;

*End of code;
```

Description of data

#	Variable	Type	Len	Format	Informat	Label
1	id	Num	8	BEST12.	F12.	Subject ID
2	gender	Num	8			Gender (1=male, 2=female)
3	age	Num	8			Age (years)
4	raceeth	Num	8			Race/Ethnicity 1=Mexican American 2=Other Hispanic 3=Non-Hispanic White 4=Non-Hispanic Black 5=Other race (including multi-racial)
5	bmi	Num	8			Body Mass Index (kg/m**2)
6	SMK	Num	8			smoking status 0=Never 1=Former 2=Current
7	dm	Num	8			self-reported physician diagnosis of DM 0=No 1=Yes
8	hr_tvcomp	Num	8			Hours watch TV or use computer per day over past 30 days
9	block	Num	8			Census block (1-11)
10	lowpir	Num	8			Low poverty-to-income ratio (1: yes, 0: no)

```
* To categorize into quartiles;
proc rank data=your_data out=new_data groups=4;
 var variable;
 ranks new_quartile_variable;
run;
```