AGYEMANG ERIC
MAT 450 HOMEWORK 5
Generated SAS Output

QUESTION 6

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
case	12	1 2 3 4 5 6 7 8 9 10 11 12

Number of Observations Read 36
Number of Observations Used 36

Dependent Variable: worms

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	149.6388889	13.6035354	3.00	0.0117
Error	24	108.6666667	4.5277778		
Corrected Total	35	258.3055556			

R-Square	Coeff Var	Root MSE	worms Mean
0.579310	58.47547	2.127858	3.638889

Level of case	N	worms	
		Mean	Std Dev
1	**3**	4.33333333	3.05505046
2	**3**	3.33333333	1.15470054
3	**3**	1.00000000	1.00000000
4	**3**	5.00000000	1.73205081
5	**3**	7.00000000	2.64575131
6	**3**	3.33333333	3.51188458
7	**3**	3.66666667	2.30940108
8	**3**	1.66666667	1.52752523
9	**3**	5.00000000	2.00000000
10	**3**	2.66666667	1.52752523
11	**3**	6.66666667	2.51661148
12	**3**	0.00000000	0.00000000

The SURVEYMEANS Procedure

Data Summary	
Number of Clusters	12
Number of Observations	36
Sum of Weights	13920

Statistics					
Variable	**N**	**Mean**	**Std Error of Mean**	**95% CL for Mean**	
worms	36	3.638889	0.608324	2.29997721	4.97780057

We can also calculate the following quantities.

AGYEMANG ERIC

MAT 450 #HMW-6

CHAPTER 5: (5) Also given $N = 580$, $n = 12$, $m_i = 3$, $M_i = 24$

$\hat{t}_{unb} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$ and $w_{ij} = \frac{N M_i}{n m_i} = \frac{(580)\cdot(24)}{(12)(3)} = 386.667$

$\Rightarrow \hat{t}_{unb} = 386.667 + 1933.34 + 2706.67 + 1546.67 + 773.334 +$
$1546.67 + 386.667 + 773.334 + 1160 + 2320 + 2320 + 1546.68$
$+ 3480 + 3093.34 + 2706.67 + 1160 + \ldots + 0$
$= 50,653.4$

next, $\hat{V}(\hat{t}_{unb}) = N^2\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i \in S}\left(1 - \frac{m_i}{M_i}\right) M_i^2 \frac{s_i^2}{m_i}$ where

$s_t^2 = \frac{1}{n-1}\sum_{i \in S}\left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2$ and $\hat{t}_i = M_i \bar{y}_i$ and $s_i^2 = \frac{1}{m_i - 1}\sum_{j \in S}(y_{ij} - \bar{y}_i)^2$

Then $s_t^2 = \frac{1}{12-1}\left[\left(104 - \frac{50653.4}{580}\right)^2 + \left(80 - \frac{50653.4}{580}\right)^2 + \cdots + \left(0 - \frac{50653.4}{580}\right)^2\right]$

$= \frac{28730.7}{11} = 2611.88$

So $\hat{V}(\hat{t}_{unb}) = 580^2\left(1 - \frac{12}{580}\right)\cdot\frac{2611.88}{12} + \frac{580}{12}(9128) = 72145998.93$

and $SE(\hat{t}_{unb}) = \sqrt{\hat{V}(\hat{t}_{unb})} = \sqrt{72145998.93} = 8493.88$

The CI as required is $\hat{t}_{unb} \pm z_{\alpha/2} \cdot SE(\hat{t}_{unb})$

$= 50653.4 \pm 1.96(8493.88) = [34005.395, 67301.405]$

$\hat{V}_{WR}(\hat{t}_{unb}) = N^2 \frac{s_t^2}{n} = 580^2\left(\frac{2611.88}{12}\right) = 73219702.67$

Comparing the estimated value of variance and the approximate of the estimator with replacement

$\Rightarrow \frac{\hat{V}(\hat{t}_{unb})}{\hat{V}_{WR}(\hat{t}_{unb})} = \frac{72145998.93}{73219702.67} = 0.9853$

The comparison shows there is not much difference between them.

QUESTION 16

```
##QUESTION 16A
> #The percentage of parents who returned a consent form is given by column ybar_i in the table meas_agg below.
> library(dplyr)
> meas_agg <-as.data.frame(measles %>%
+                              group_by(`School No`= school, Mi = Mitotal, ki = mi) %>%
+                              summarize(Return = sum(returnf==1, na.rm = TRUE),
+                                        mi = sum(returnf!=9, na.rm = TRUE),
+                                        ybar_i = Return/mi))
> meas_agg
   School No  Mi ki Return mi    ybar_i
1          1  78 40     19 38 0.5000000
2          2 238 38     19 36 0.5277778
3          3 261 19     13 17 0.7647059
4          4 174 30     18 30 0.6000000
5          5 236 30     12 26 0.4615385
6          6 188 25     13 24 0.5416667
7          7 113 23     15 22 0.6818182
8          8 170 43     21 36 0.5833333
9          9 296 38     23 35 0.6571429
10        10 207 21      7 17 0.4117647
> N<-46
> n<-nrow(meas_agg)
>
> meas_agg$si_sq<-c(.25676,.25635,.19118,.24828,.25846,.25906,.22727,.25,.23193,.25735)
>
> meas_agg$si_sq
 [1] 0.25676 0.25635 0.19118 0.24828 0.25846 0.25906 0.22727 0.25000 0.23193 0.25735
>
> ############################################################################
```

```
> #QUESTION 16B
> #The sampling weight for each observation is given by the "weight" column in the table meas_agg1 below.
> meas_agg11<-as.data.frame(meas_agg %>%
+                               group_by(`School No`, Mi, ki, Return, mi, ybar_i, si_sq) %>%
+                               summarize(est_ti = Mi*Return/mi))
>
>
> ybar_r<-sum(meas_agg11[,"est_ti"])/sum(meas_agg11[, "Mi"])
> ybar_r
[1] 0.5789482
>
> var_ybar_r<-(1/(Mbar^2))*(((1-n/N)*(sr_squared/n))+(sum(final)/(n*N)))
> var_ybar_r
[1] 0.00138099
>
> meas_agg1 <-as.data.frame(meas_agg11 %>%
+                               group_by(`School No`, Mi, ki, Return, mi, ybar_i, est_ti, si_sq)
%>%
+                               summarize(squared_deviation = (est_ti-Mi*ybar_r)^2,
+                                         final = (Mi^2)*(1-mi/Mi)*(si_sq/mi)))
> attach(meas_agg1)


meas_agg1$weight<-(N/n)*(Mi/mi)
```

```
> meas_agg1
School    No  Mi ki Return mi    ybar_i    est_ti   si_sq squared_deviation    final    weight
1          1  78 40     19 38 0.5000000  39.00000 0.25676        37.9204848  21.08135  9.442105
2          2 238 38     19 36 0.5277778 125.61111 0.25635       148.3174465 342.34118 30.411111
3          3 261 19     13 17 0.7647059 199.58824 0.19118      2350.5770575 716.18277 70.623529
4          4 174 30     18 30 0.6000000 104.40000 0.24828        13.4176413 207.36346 26.680000
5          5 236 30     12 26 0.4615385 108.92308 0.25846       767.7722103 492.66452 41.753846
6          6 188 25     13 24 0.5416667 101.83333 0.25906        49.1251259 332.80575 36.033333
7          7 113 23     15 22 0.6818182  77.04545 0.22727       135.1244850 106.22806 23.627273
8          8 170 43     21 36 0.5833333  99.16667 0.25000         0.5557246 158.19444 21.722222
9          9 296 38     23 35 0.6571429 194.51429 0.23193       535.7193943 511.94240 38.902857
10        10 207 21      7 17 0.4117647  85.23529 0.25735      1197.6435659 595.38679 56.011765
```

#QUESTION 16C

> #The overall percentage of parents who received a consent form along with a 95% CI is given below

```
> sr_squared<-sum(meas_agg1[,"squared_deviation"])/(nrow(meas_agg1)-1)
> ybar_r<-sum(meas_agg1[,"est_ti"])/sum(meas_agg1[, "Mi"])
> attach(meas_agg1)
N<-46
> Mbar<-sum(Mi)/n
> var_ybar_r<-(1/(Mbar^2))*(((1-n/N)*(sr_squared/n))+(sum(final)/(n*N)))
> var_ybar_r
[1] 0.00138099
> ###point estimate of percentage of parents
> ybar_r
[1] 0.5789482
>
> ##confidence interval for ybar_r
>
> ybar_r-1.96*sqrt(var_ybar_r); ybar_r+1.96*sqrt(var_ybar_r)
[1] 0.5061113
[1] 0.6517851
```

(0.5061113,0.6517851), As the CI required.

#QUESTION 16D

(16)

i) Analysing the data as SRS, $\hat{t}_{unb} = \sum_{i \in S} \sum_{j \in S_i} w_{ij} y_{ij}$

where $w_{ij} = \frac{N M_i}{n m_i}$.

But $\hat{V}(\hat{t}_{unb}) = N^2\left(1 - \frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\sum_{i \in S}\left(1 - \frac{m_i}{M_i}\right)M_i^2 \frac{s_i^2}{m_i}$

where $s_t^2 = \frac{1}{n-1}\sum_{i \in S}\left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2$; $\hat{t}_i = M_i \bar{y}_i$ is the estimated total of psu and $s_i^2 = \frac{1}{m_i - 1}\sum_{j \in S_i}(y_{ij} - \bar{y}_i)^2$

school	Number of parents	M_i	m_i	w_{ij}	$w_{ij} y_{ij}$
1	38	78	40	8.190	311.22
2	35	238	38	26.305	920.684
3	17	261	19	57.695	980.811
4	30	174	30	24.360	730.80
5	26	236	30	33.040	859.04
6	24	188	25	31.584	758.016
7	22	113	23	20.635	453.965
8	36	170	43	16.605	597.767
9	35	296	38	32.716	1145.05
10	17	207	21	41.400	703.8
				Total	7461.16

$\therefore \hat{t}_{unb} = 7461.16$

~~$V(\hat{t}_{unb})$~~ $s_t^2 = \frac{1}{10-1}\left\{\left(2184 - \frac{7461.16}{46}\right)^2 + \left(6664 - \frac{7461.16}{46}\right)^2 + \cdots + \left(5796 - \frac{7461}{46}\right)^2\right.$

$= 34828617$

$\hat{V}(\hat{t}_{unb}) = 46^2\left(1 - \frac{10}{46}\right)\frac{34828617}{10} + \frac{46}{10} \times 40593.94 = 5767308907.3$

$SE(\hat{t}_{unb}) = \sqrt{\hat{V}(\hat{t}_{unb})} = \sqrt{5767308907.34} = 75942.80$

$CI = \hat{t}_{unb} \pm z_{\alpha/2} SE(\hat{t}_{unb}) = (7461.16 \pm 1.96(75942.80))$

$= [-141386.73, 156309.048]$

The ratio, $\frac{\text{Estimated variance from (C)}}{\text{Estimated variance if the data were analyzed as SRS}} = \frac{85665.34}{5767308907.34}$

$= 0.000015$

Therefore, the effecting of clustering vary low that is deff = 0.000015

(25) (a) If $M_i = M$ for $\forall\, i$ and $m_i = m$ for $\forall\, i$, then

$\cancel{\hat{\bar{y}}_r = \frac{\sum_{i\in S} M\bar{y}_i}{nM}}$

$$\hat{t}_{unb} = \frac{N}{n}\sum_{i\in S} t_i = \frac{N}{N}\sum_{i\in S} t_i = \sum_{i\in S} t_i = \hat{t}_r$$

Hence, $\hat{\bar{y}}_r = \dfrac{\sum_{i\in S} M\bar{y}_i}{nM} = \dfrac{1}{NM}\hat{t}_{unb} = \hat{\bar{y}}_{unb}$ shown.

(b) It follows that $\hat{\bar{y}} = \hat{\bar{y}}_r = \hat{\bar{y}}_{unb}$

source	df	SS	MS
Between clusters	$n-1$	$\sum_{i\in S}\sum_{j\in S_i}(\bar{y}_i - \hat{\bar{y}})^2$	msb
within clusters	$n(m-1)$	$\sum_{i\in S}\sum_{j\in S_i}(y_{ij} - \bar{y}_i)^2$	msw
Total	$nm-1$	$\sum_{i\in S}\sum_{j\in S_i}(y_{ij} - \hat{\bar{y}})^2$	msto

(c) Let $z_i = \begin{cases} 1; & \text{if psu } i \text{ in sample} \\ 0; & \text{otherwise} \end{cases}$

But $\widehat{SSW} = \sum_{i\in S}\sum_{j\in S_i}(y_{ij} - \bar{y}_i)^2$

so $E(\widehat{SSW}) = E\left(\sum_{i=1}^{N} z_i \sum_{j\in S_i}(y_{ij} - \bar{y}_i)^2\right) = E\left\{\sum_{i=1}^{N} z_i\, E\left[\sum_{j\in S_i}(y_{ij} - \bar{y}_i)^2 / z\right]\right\}$

$$= E\left\{\sum_{i=1}^{N} z_i (m-1) E\left[s_i^2 / z\right]\right\} = E\left\{\sum_{i=1}^{N} z_i (m-1) S_i^2\right\}$$

$$= (m-1)\frac{n}{N}\sum_{i=1}^{N} S_i^2$$

But $MSW = \dfrac{SSW}{n(m-1)} \Rightarrow E(MSW) = \dfrac{\cancel{(m-1)}\frac{\cancel{n}}{N}\sum_{i=1}^{N} S_i^2}{\cancel{n}\cancel{(m-1)}}$

$\Rightarrow E(msw) = \dfrac{1}{N}\sum_{i=1}^{N} S_i^2 = MSW$ shown

Since $M_i = M$ and $m_i = m$ for $\forall\, i$ $\hat{\bar{y}}_{unb} = \frac{1}{n}\sum \bar{y}_i = \frac{1}{n}\sum_{i=1}^{N} z_i \bar{y}_i$

so $\hat{SSB} = \sum_{i\in S}\sum_{j\in S_i}(\bar{y}_i - \hat{\bar{y}}_{unb})^2 = m\sum_{i\in S}(\bar{y}_i - \hat{\bar{y}}_{unb})^2$

$$E(\hat{SSB}) = mE\left(\sum(\bar{y}_i^2 - 2\bar{y}_i\hat{\bar{y}}_{unb} + \hat{\bar{y}}_{unb}^2)\right] = E\left(\sum(\bar{y}_i^2 - n\hat{\bar{y}}_{unb}^2)\right)$$

$$= mE\left(\left[\sum_{i=1}^{N} z_i\bar{y}_i^2 \,/\, z_1, \ldots, z_n\right]\right) - mnE\left[\hat{\bar{y}}_{unb}^2\right]$$

$$= mE\left(\sum z_i\bar{y}_i^2 \,/\, z_1, \ldots, z_n\right) - mnE(\hat{\bar{y}}_{unb}^2$$

$$= mE\left(\sum z_i\{V(\bar{y}_i \mid z) + \bar{y}_{iu}^2\}\right] - mn\left[V(\hat{\bar{y}}_{unb}) + \bar{y}_u^2\right]$$

$$= mE\left[\sum_{i=1}^{N} z_i\left\{\left(1-\frac{m}{M}\right)\frac{s_i^2}{m} + \bar{y}_{iu}^2\right\}\right] - mn\left[\frac{1}{M^2}\left(1-\frac{n}{N}\right)\frac{S_t^2}{n} + \frac{1}{nN}\sum_{i=1}^{N}\left(1-\frac{m}{M}\right)\frac{s_i^2}{m} + \bar{y}_u^2\right]$$

$$= \frac{mn}{N}\sum_{i=1}^{N}\left[\left(1-\frac{m}{N}\right)\frac{s_i^2}{m} + \bar{y}_{iu}^2\right] - \frac{m}{M^2}\left(1-\frac{n}{N}\right)S_t^2 - \frac{m}{n}\sum_{i=1}^{N}\left(1-\frac{m}{M}\right)\frac{s_i^2}{m} - mn\bar{y}_u^2$$

$$= \frac{m(n-1)}{N}\left(1-\frac{m}{M}\right)\sum_{i=1}^{N}\frac{s_i^2}{m} + mn\left[\frac{1}{N}\sum\bar{y}_{iu}^2 - \bar{y}_u^2\right] - \frac{m}{M}\left(1-\frac{n}{N}\right)(MSB)$$

$$= \frac{m(n-1)}{N}\left(1-\frac{m}{M}\right)\sum_{i=1}^{N}\frac{s_i^2}{m} + \frac{(N-1)mn - n(N-n)}{NM(N-1)}SSB$$

$$= (n-1)\left(1-\frac{m}{M}\right)MSW + (n-1)\frac{m}{M}MSB$$

so $E(msb) = \left(1-\frac{m}{M}\right)MSW + \frac{m}{M}MSB$

$$= \frac{m}{M}MSB + \left(1-\frac{m}{M}\right)MSW \quad \text{shown}$$

(d) $E(\hat{MSB}) = \frac{M}{m}\left(1-\frac{m}{M}\right) MSW + \frac{M}{m}\frac{m}{M} MSB - \left(\frac{M}{m}-1\right) MSW$

$= MSB$ hence unbiased estimator of MSB.

(e) $s_t^2 = \frac{1}{n-1}\sum_{i\in S}\left(\hat{t}_i - \frac{\hat{t}_{unb}}{N}\right)^2 = \frac{1}{n-1}\sum_{i\in S}\left(M\bar{y}_i - M\hat{\bar{y}}\right)^2$

$= \frac{1}{n-1}\frac{M^2}{m}\sum_{i\in S}\sum_{j\in S_i}\left(\bar{y}_i - \hat{\bar{y}}\right)^2 = \frac{M^2}{m} msb$

and $\sum_{i\in S} s_i^2 = \frac{1}{m-1}\sum_{i\in S}\sum_{j\in S_i}\left(y_{ij} - \bar{y}_i\right)^2 = n\,msw$

Then $\hat{V}(\hat{\bar{y}}_{unb}) = \frac{1}{(NM)^2}\left[N^2\left(1-\frac{n}{N}\right)\frac{s_t^2}{n} + \frac{N}{n}\left(1-\frac{m}{M}\right)\frac{M^2}{m}\sum_{i\in S} s_i^2\right]$

$= \left(1-\frac{n}{N}\right)\frac{msb}{nm} + \frac{1}{N}\left(1-\frac{m}{M}\right)\frac{msw}{m}$. Shown

```
/* AGYEMANG ERIC*/
/*MAT 450 HOMEWORK */
/*Generated SAS Code*/
/*QUESTION 6*/
data worms;
do case = 1 to 12;
do can = 1 to 3;
input worms @@;
wt = (580/12)*(24/3);
output;
end;
end;
cards;
1 5 7
4 2 4
0 1 2
3 6 6
4 9 8
0 7 3
5 5 1
3 0 2
7 3 5
3 1 4
4 7 9
0 0 0
;
proc print data=worms;
run;
proc glm data=worms;
class case;
model worms = case;
mean case;
run;
/* Due to the 2-stage sampling, SAS do not calculate the extra term for variance */
proc surveymeans data=worms total = 580;
weight wt;
cluster case;
var worms;
run;
```

```
#R-CODES
##QUESTION 16A
#The percentage of parents who returned a consent form is given by column ybar_i in the table meas_agg below.
```

```
library(dplyr)
meas_agg <-as.data.frame(measles %>%
                group_by(`School No`= school, Mi = Mitotal, ki = mi) %>%
                summarize(Return = sum(returnf==1, na.rm = TRUE),
                      mi = sum(returnf!=9, na.rm = TRUE),
                      ybar_i = Return/mi))
meas_agg

N<-46
n<-nrow(meas_agg)

meas_agg$si_sq<-c(.25676,.25635,.19118,.24828,.25846,.25906,.22727,.25,.23193,.25735)

meas_agg$si_sq

##############################################################################################
####
```

#QUESTION 16B

```
#The sampling weight for each observation is given by the "weight" column in the table meas_agg1 below.
meas_agg11<-as.data.frame(meas_agg %>%
                group_by(`School No`, Mi, ki, Return, mi, ybar_i, si_sq) %>%
                summarize(est_ti = Mi*Return/mi))

ybar_r<-sum(meas_agg11[,"est_ti"])/sum(meas_agg11[, "Mi"])
ybar_r

var_ybar_r<-(1/(Mbar^2))*(((1-n/N)*(sr_squared/n))+(sum(final)/(n*N)))
var_ybar_r

meas_agg1 <-as.data.frame(meas_agg11 %>%
                group_by(`School No`, Mi, ki, Return, mi, ybar_i, est_ti, si_sq) %>%
                summarize(squared_deviation = (est_ti-Mi*ybar_r)^2,
                      final = (Mi^2)*(1-mi/Mi)*(si_sq/mi)))
attach(meas_agg1)

meas_agg1$weight<-(N/n)*(Mi/mi)
meas_agg1

##############################################################################################
##
```

#QUESTION 16C

```
#The overall percentage of parents who received a consent form along with a 95% CI is given below
sr_squared<-sum(meas_agg1[,"squared_deviation"])/(nrow(meas_agg1)-1)
ybar_r<-sum(meas_agg1[,"est_ti"])/sum(meas_agg1[, "Mi"])
```

```
attach(meas_agg1)
N<-46

Mbar<-sum(Mi)/n
var_ybar_r<-(1/(Mbar^2))*(((1-n/N)*(sr_squared/n))+(sum(final)/(n*N)))
var_ybar_r
###point estimate of percentage of parents
ybar_r



##confidence interval for ybar_r

ybar_r-1.96*sqrt(var_ybar_r); ybar_r+1.96*sqrt(var_ybar_r)
#(0.5061113,0.6517851), As the CI
```