Leo Do
Professor Teevrat Garg
Quantitative Methods 1
6 December 2019

QM Project Writeup: Questions 1-2

Row Labels	Count of Villages (cluster_id)
2005	366
2008	318
2010	385
(blank)	
Grand Total	**1069**

svy_yr and unique housholds	Households Counts
svy_yr	*COUNTUNIQUE of momid*
2005	6648
2008	11326
2010	9455
Grand Total	**27429**

1A)
I made a pivot table utilizing uniquely counted cell blocks on a windows computer to find the number of households, which is seen through the "momids" being the representation of a household, that are surveyed in each survey year. Reutilizing the pivot chart, I was able to find the count of unique "DHS clusters" to find the number of villages in each survey year by creating a column of the dhs clusters and removing the duplicates. With these dublicates, a pivot table was used to find the counts of unique villages by survey year. These two tables are found above illustrate the number of villages and households that were unique in each survey year. In conclusion, we can see that there are around 300 unique villages being surveyed over the three survey sessions. Additionally, there are 27429 unique households being surveyed in each survey year. It can be seen that in 2008 there are the most unique households despite 2010 having the most unique villages.

1B)
Table of births by year

Birth Years	Count of children born
2001	6204
2002	7271
2003	7401
2004	8035
2005	6040
2006	5288
2007	5317
2008	5565
2009	2438
2010	2574
Grand Total	**56133**

Leo Do
Professor Teevrat Garg
Quantitative Methods 1
6 December 2019

Survey Years	Count of Deaths(Age_of_Deaths)
2005	691
2008	2106
2010	1851
Grand Total	**4648**

QM Project Writeup: Questions 1-2

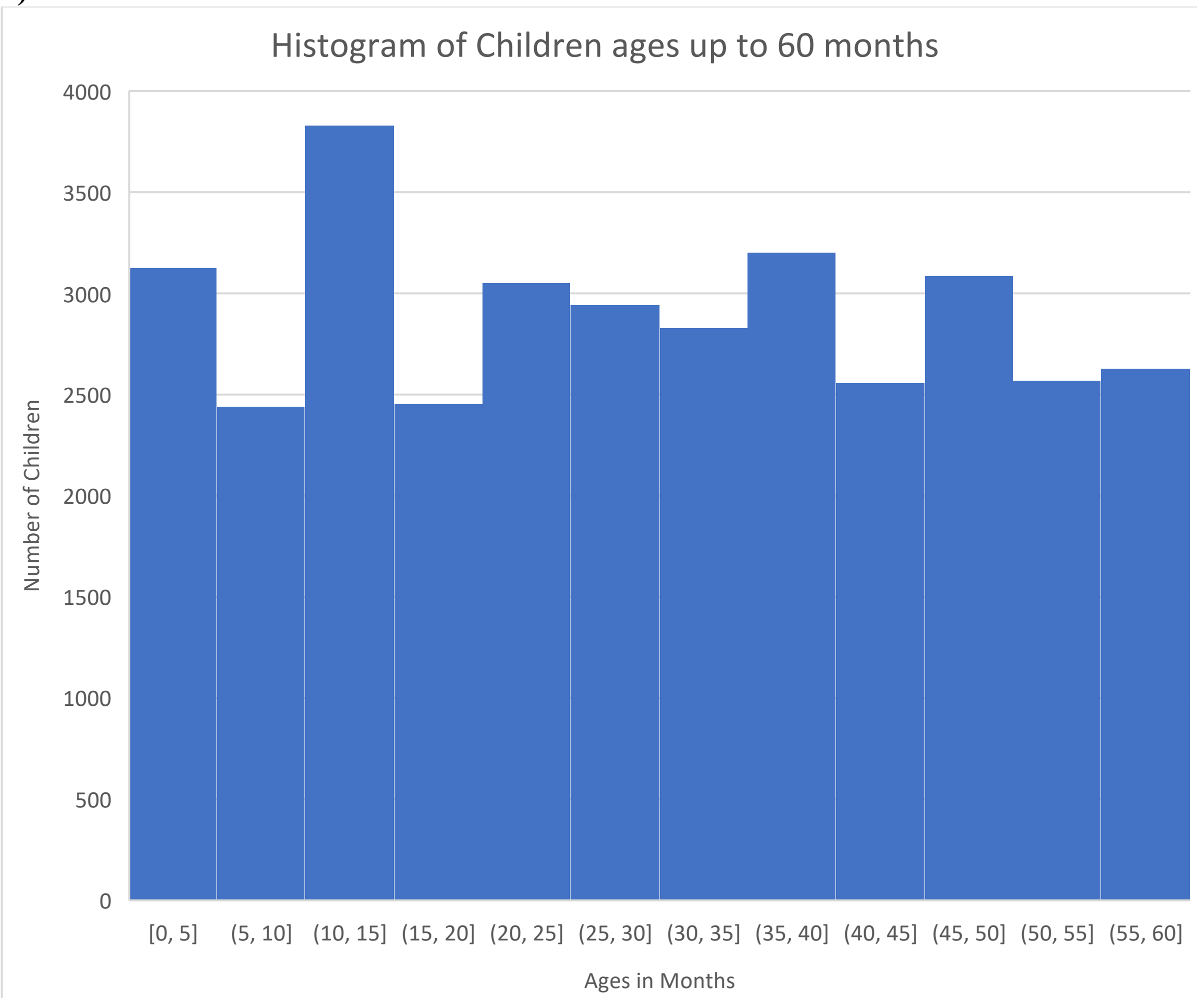
Age of Deaths(months)	Survey Years			
	2005	**2008**	**2010**	**Grand Total**
0	312	922	758	1992
1	39	107	93	239
2	27	63	70	160
3	23	79	59	161
4	17	49	42	108
5	12	32	39	83
6	32	52	45	129
7	17	38	29	84
8	10	37	29	76
9	15	41	35	91
10	7	26	12	45
11	8	21	24	53
12	56	124	112	292
Grand Total	**575**	**1591**	**1347**	**3513**

I created tables of total births over the years using count of birth years and sorting it into their respective years. For the count of deaths per survey, I again also used a pivot table to the total number of deaths reported in each survey year using the age of death column as indications of a death. The same can be said for the counts children that were reported dead aged 12 months and younger at the time of their deaths for each survey year. We can conclude that there is a no major difference of birth numbers through the years. We can also see that there are large proportions of deaths within the first month, month 0, for babies that are being reported. In 2008 it is reported to have the most deaths ages 1 and younger as seen in the grand totals, in addition to having the most deaths reported.

Leo Do
Professor Teevrat Garg
Quantitative Methods 1
6 December 2019

QM Project Writeup: Questions 1-2

1)



I created a histogram to illustrate the number of children that were alive and that were between the ages of 0 and 60 months with the widths of each bar being 5 months in width. The children that are alive are inferred to be the cell blocks that are blank in the "age of death column" To calculate the ages, I found the difference between the survey year and a child's birth year in months and subtracted out the birth month to account for the time they were still in the womb and added 12 to make the survey date each being December 31st. After calculating the ages, I filtered out all ages that were above 60 months for the histogram illustrated above. In the histogram a multimodal or plateau distribution is seen rather than a normal bell curve distribution, due to the even levelness of all the category widths.

2a)

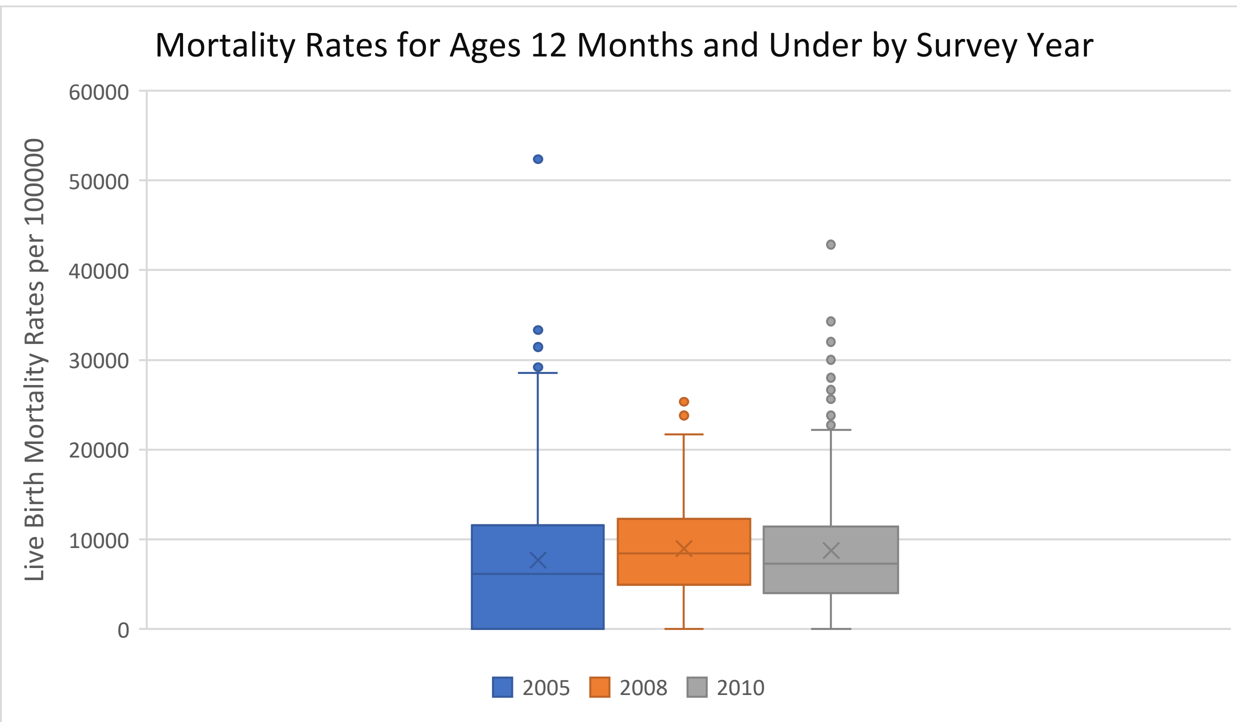
First, I created pivot tables of my cluster village IDs and their corresponding counts of age at deaths, birth years(to be a substitute for total births), and the calculated each villages live births by subtracting the deaths away from the births totals. This newly created live birth is used to find the mortality rates of the cluster villages through the quotient of total deaths over total live births,

Leo Do
Professor Teevrat Garg
Quantitative Methods 1
6 December 2019

QM Project Writeup: Questions 1-2

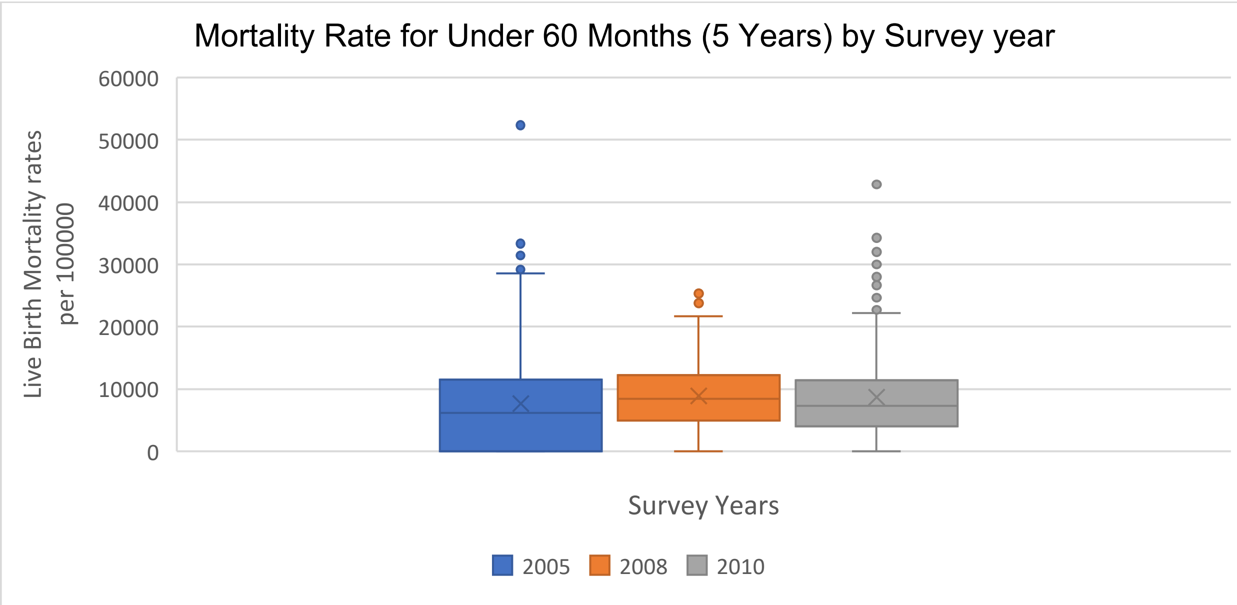
then multiplying this rate by 100,000 to find the expected number of deaths per 100,000 livebirths. With these steps I filtered out ages 12 months and under to find their rates and did the same steps for 60 months and under rates. Afterwards, I inputted the rates of these age clusters by survey year in one figure holding all three box and whisker plot per age cluster I wanted. These two figures can be seen below. Additionally, to find the mean and medians mortality rates of each survey year was calculated through the "average" and "median" functions on excel as shown in the table below.

Survey Year	Mean	Median
2005	7668.7124	6155.30303
2008	8934.31811	8435.87843
2010	8707.82707	7272.72727



Leo Do
Professor Teevrat Garg
Quantitative Methods 1
6 December 2019

QM Project Writeup: Questions 1-2



In conclusion, we can see that for both 60 and 12 months and under, the interquartile range within 2005 is the largest compared to the other two survey years. 2005 can be seen to have the outlier that is the farthest from the box for both age clusters.

2b)

To calculate the mortality differences between the cluster year ranges of "00-04" and "06-10" I used pivot tables to found the total counts of births and deaths of each "**Cluster(Village) ID**" then I used a "Sum" function to add up the deaths or births of each cluster year grouping. From there I calculated the mortality rates of the villages and their corresponding "norm.dist" value. This norm.dist value gives me the values to create a Normal Probability Density plot shown below. Before finding the normal distribution probability values, I needed to find the sample mean and sample standard deviations of the mortality rates between the two-year range clusters. The density plot below illustrates the differences between mortality rate distributions of the two-year clusters.

A statistical hypothesis test was also needed to see if we can conclude with a significance of .05 to see if the differences between the true population means of the years are different. This was done through a two tailed test with unequal variance between the populations. The null hypothesis I was looking to negate, and reject was that the true population mortality rate means were the same between the year ranges. The alternative hypothesis was that there was a difference of mortality rates between the year ranges. These hypothesis will be written as, $H_o\colon \mu_{00-04} = \mu_{06-10}$ and $H_A\colon \mu_{00-04} \neq \mu_{06-10}$ with $\mu = mortality\ rates$ of each category. From these two test parameters, I utilized the "t.test" function to obtain a P value of "2.32178E-12" which is clearly lower than my significance value of .05 allowing me to reject the hypothesis that the true means of the year range clusters are the same. Within the "t.test" function, the arrays of data that was used did not matter in which order it was inputted, whether it was mortality rates of each village cluster in either time ranges. I used a two tailed test to see simply if there was a

Leo Do
Professor Teevrat Garg
Quantitative Methods 1
6 December 2019

QM Project Writeup: Questions 1-2

difference between the two true population means and an unequal variance of the two samples. In conclusion we can now also statistically see that there is a difference between 2000-2004 and 2006-2010.