Part 1, Question 1: Exploratory analysis

In this series of questions, we examine data from a study of 158 infants who visited Northbay Healthcare in Solano County, California for a Vitamin K shot. Assume that the infants in the study are a representative sample from all infants in Northbay Healthcare.
Nurses administered a Vitamin K shot to each infant. Infants were randomized to two different protocols to study how to reduce pain experienced by the infants due to the shot. The infants were divided into two groups – the control group, where standard protocol for handling the infants was used; and an intervention group, where mothers held their infants prior to, during, and after administration of the shot. Pain was measured using the Neonatal Infant Pain Score (NIPS) (Lawrence et. al 1993). The variables in the dataset are described below:

- id – unique identifier for each infant
- group – 1 if intervention group, 0 if control
- pain0 – NIPS score 0 seconds after shot
- pain30 – NIPS score 30 seconds after shot
- pain60 – NIPS score 60 seconds after shot
- pain120 – NIPS score 120 seconds after shot
- crytime – total time that the infant cried in seconds

Use the babies.dta dataset to answer the questions below.
These data were made available through SOCR (http://www.socr.ucla.edu/).
Source: Lawrence J, Alcock D, McGrath P, Kay J, MacMurray SB, Dulberg C. (1993) The development of a tool to assess neonatal pain, Neonatal Network, 12:59-66

Before jumping into analyzing the babies.dta dataset, first explore the dataset using summary statistics and graphical analyses.

Calculate the average cry time in each group

```
. mean crytime, over(group)

Mean estimation                     Number of obs    =     158

      control: group = control
 intervention: group = intervention

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
crytime      |
     control |   39.20253   2.624219       34.0192    44.38586
intervention |   29.60759   2.430496      24.80691    34.40828
--------------------------------------------------------------
```

Control: 39.20253
Intervention: 29.60759

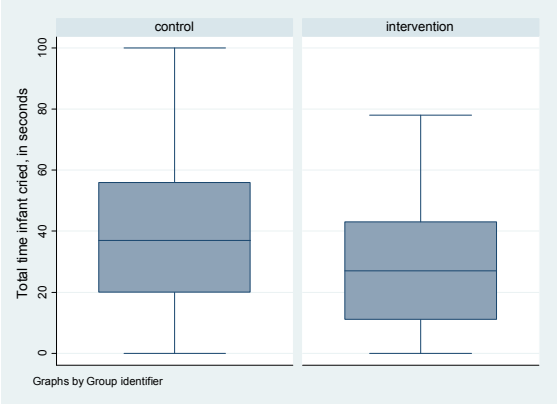
Calculate the median cry time in each group.

```
. by group, sort : summarize crytime, detail
-------------------------------------------------------------------------------
-> group = control
             Total time infant cried, in seconds
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            2              0
10%           11              0       Obs                  79
25%           20              2       Sum of Wgt.          79
50%           37                      Mean           39.20253
                      Largest       Std. Dev.      23.32457
75%           56             81
90%           73             84       Variance       544.0354
95%           81             86       Skewness       .3022815
99%          100            100       Kurtosis       2.394307
-------------------------------------------------------------------------------
-> group = intervention
             Total time infant cried, in seconds
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            5              0       Obs                  79
25%           11              0       Sum of Wgt.          79
50%           27                      Mean           29.60759
                      Largest       Std. Dev.      21.60272
75%           43             72
90%           64             73       Variance       466.6774
95%           72             73       Skewness         .53556
99%           78             78       Kurtosis       2.238552
```

Control: 37
Intervention: 27

Make a boxplot of cry time by group. According to the boxplot, which group has more variability in cry time? Control

```
graph box crytime, by(group)
```



Using the central limit theorem, construct a 95% confidence interval for the average total cry time for infants in the control group and infants in the intervention group. For this question only, assume that the standard deviation of cry time within each group is known and is equal to 22 seconds.

Control		Intervention	
Lower Bound:	34.35124	Lower Bound:	24.7563
Upper Bound:	44.05382	Upper Bound:	34.45888

Consider a population with mean μ and standard deviation σ. According to the CLT, for large sample sizes, the sample mean approximately follows a normal distribution with mean μ and standard deviation σ/√n. Then, a 95% confidence interval for μ is: $\bar{x} \mp Z_{0.025}\sigma/\sqrt{n}$.

Repeat this calculation for each group.

```
by group, sort : summarize crytime
-------------------------------------------------------------------------------
-> group = control

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     crytime |        79    39.20253    23.32457          0        100

-> group = intervention
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     crytime |        79    29.60759    21.60272          0         78

di 39.20253 - invnormal(0.975)*22/sqrt(79)
34.35124

. di 39.20253+invnormal(0.975)*22/sqrt(79)
44.05382

. di 29.60759 - invnormal(0.975)*22/sqrt(79)
24.7563

. di 29.60759 + invnormal(0.975)*22/sqrt(79)
34.45888
```

Part 1, Question 2: Two-sample Non-parametric Test

Now, we examine the relationship between cry time and group among infants at Northbay.

1. Suppose we wish to perform a two-sample test, but we do not want to make any normality (or other strong parametric) assumptions. Conduct an appropriate non-parametric test to test whether the distribution of cry time is the same in both groups at the 0.05 level of significance.

Note that the correct test to use is the Wilcoxon rank sum test.

```
. ranksum crytime, by(group)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       group |      obs    rank sum    expected
-------------+---------------------------------
     control |       79      7043.5      6280.5
intervention |       79      5517.5      6280.5
-------------+---------------------------------
    combined |      158       12561       12561

unadjusted variance    82693.25
adjustment for ties      -41.39
                     ----------
adjusted variance      82651.86

Ho: crytime(group==control) = crytime(group==intervention)
             z =   2.654
    Prob > |z| =   0.0080
```

What is your p-value? 0.0080

Your conclusion from the test?

a) There is evidence that the means of the two groups are different (specifically, there is evidence that the mean is higher in the control group)
b) There is not evidence that the means of the two groups are different is higher in the control group)
c) None of the above

The nonparametric test tells us about differences in medians, not means.

2. Assuming randomization was successful, which of the following should we be concerned about:

a) Confounding by sex of the infant
b) Confounding by the amount of pain experienced by the infant
c) Effect modification by sex of the infant
d) Misclassification of the exposure status of the infant

Part 1, Question 3: Linear Regression

In the babies.dta full dataset, generate a covariate called painind defined as 1 if the infant experienced severe pain upon receiving the shot (pain0 = 7) and as 0 otherwise. In Stata, you can use the commands:

```
generate painind = 0
replace painind = 1 if pain0 == 7
```

Fit a linear regression model with total cry time as the outcome; and with painind and group as covariates.

The regression model is: $Y_i = \beta_0 + \beta_1\, group_i + \beta_2\, painind_i + \epsilon_i$

where $\varepsilon_i \sim N(0, \sigma^2)$.

1. Using the notation from the model above, what are your estimates of the regression coefficients and residual standard deviation?

```
. regress crytime group painind

      Source |       SS       df       MS              Number of obs =     158
-------------+------------------------------           F(  2,   155) =    9.54
       Model |  9040.70577     2  4520.35289            Prob > F      =  0.0001
    Residual |  73431.3702   155  473.750775            R-squared     =  0.1096
-------------+------------------------------           Adj R-squared =  0.0981
       Total |  82472.0759   157  525.299847            Root MSE      =  21.766

------------------------------------------------------------------------------
     crytime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |  -7.679168   3.509335    -2.19   0.030    -14.61146   -.7468734
     painind |   12.61215   3.734197     3.38   0.001     5.235661    19.98863
       _cons |   29.78333   3.711391     8.02   0.000      22.4519    37.11477
```

β_0: 29.78333
β_1: -7.679168
β_2: 12.61215
σ: 21.766

2.Estimate the mean change in cry time for infants with severe pain versus those without severe pain, holding group constant. Provide a 95% confidence interval for this estimate.

Estimate: 12.61215
95% Confidence interval Lower Bound: 5.235661
95% Confidence interval Upper Bound: 19.98863

This is simply the estimate of β_2 and the confidence interval for β_2.

3.Again, use the notation above for the regression model. The correct interpretation for β_1 is:

a) Infants in the intervention group have β_1 times the risk of experiencing an increase in cry time compared to infants in the control group
b) Infants in the intervention group have β_1 times the risk of experiencing an increase in cry time compared to infants in the control group after controlling for pain experienced by the infant
c) Infants in the intervention group an average have β_1 increase in cry time.
d) Infants in the intervention group on average have β_1 increase in cry time after controlling for pain experienced by the infant.

4.Using the regression model, estimate the average cry time in the following groups:

Control group infants with severe pain upon receiving the shot: 42.39548

$$E(Y_i|group_i = 0 \; and \; painind_i = 1) = \beta_0 + \beta_2$$

```
lincom _cons  + painind

 ( 1)  painind + _cons = 0
------------------------------------------------------------------------------
     crytime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   42.39548   2.624988    16.15   0.000     37.21011    47.58085
------------------------------------------------------------------------------
```

Control group infants without severe pain upon receiving the shot: 29.78333

$$E(Y_i|group_i = 0 \; and \; painind_i = 0) = \beta_0$$

```
. lincom _cons

 ( 1)  _cons = 0
------------------------------------------------------------------------------
     crytime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   29.78333   3.711391     8.02   0.000      22.4519    37.11477
------------------------------------------------------------------------------
```

Intervention group infants with severe pain upon receiving the shot: 34.71631

$$E(Y_i|group_i = 1 \; and \; painind_i = 1) = \beta_0 + \beta_1 + \beta_2$$

```
. lincom _cons+group+painind

 ( 1)  group + painind + _cons = 0
------------------------------------------------------------------------------
     crytime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   34.71631   2.878326    12.06   0.000      29.0305    40.40212
------------------------------------------------------------------------------
```

Intervention group infants without severe pain upon receiving the shot: 22.10417

$$E(Y_i|group_i = 1 \; and \; painind_i = 0) = \beta_0 + \beta_1$$

```
lincom _cons + group

 ( 1)  group + _cons = 0
------------------------------------------------------------------------------
     crytime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   22.10417   3.306418     6.69   0.000     15.57271    28.63562
------------------------------------------------------------------------------
```

5. Without using the regression model, estimate the mean cry time in the following groups:

```
. mean crytime, over(group painind)

Mean estimation                   Number of obs    =     158
         Over: group painind
    _subpop_1: control 0
    _subpop_2: control 1
    _subpop_3: intervention 0
    _subpop_4: intervention 1

--------------------------------------------------------------
        Over |       Mean   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
crytime      |
   _subpop_1 |      34.95   5.902486      23.29147    46.60853
   _subpop_2 |   40.64407   2.896597      34.92274     46.3654
   _subpop_3 |     18.875   3.278642      12.39906    25.35094
   _subpop_4 |   36.91489   3.006442       30.9766    42.85319
--------------------------------------------------------------
```

Control group infants with severe pain upon receiving the shot: 40.64407
Control group infants without severe pain upon receiving the shot: 34.95
Intervention group infants with severe pain upon receiving the shot: 36.91489
Intervention group infants without severe pain upon receiving the shot: 18.875

6. Compare your estimates from the regression model to the “non-parametric” estimates above. In large sample sizes, would you expect the "non-parametric" estimates or the regression based estimates to have less bias?

a) non-parametric
b) regression

7. When we have continuous covariates we cannot estimate the means using the non-parametric method as above due to the “curse of dimensionality”. This is because:

a) some continuous variables have skewed distributions
b) there is typically only one observation per continuous variable hence the mean cannot be estimated well

8. Suppose that sex is an effect modifier of the association between group and cry time. How should you analyze your results?

a) Construct a linear regression model with sex as a covariate.
b) Construct two separate linear regression models: one among male infants and one among female infants.
c) Construct a linear regression model, but do not control for sex as a covariate because this is a randomized clinical trial.

Part 2

The following tables show the crude and sex-specific results from a Prospective Cohort Study that examines the association between a binary exposure (E) and the development of a disease (D) during 20 years of follow-up.

Since there were several versions of the two-by-two tables, different people will have different results. Here are the answers for one version of the exam:

Full Data	**D+**	**D-**	**Total**
E+	48	252	300
E-	48	252	300
Total	96	504	600

Males	**D+**	**D-**	**Total**
E+	36	144	180
E-	36	144	180
Total	72	288	360

Females	**D+**	**D-**	**Total**
E+	12	108	120
E-	12	108	120
Total	24	216	240

1. Assume that this cohort is a simple random sample from a broader population of interest. Model the number of disease positive individuals among all exposed individuals in the sample using the binomial distribution with probability of disease p{e+}; and model the number of disease positive individuals among the unexposed in the sample using a binomial distribution, with probability of disease p{e-}.
Estimate p{e+}, the proportion of exposed individuals who are disease positive, and provide an exact 95% confidence interval.

```
cii 300 48
                                                        -- Binomial Exact --
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
             |        300         .16     .021166        .1203891    .2064788
```

Estimated Proportion: 0.16
Confidence Interval:
Lower Bound: .1203891
Upper Bound: .2064788

2. Would you expect the large-sample Wilson confidence interval to provide similar results to the exact confidence intervals in question 1?

a) Yes
b) No

3. Consider the following hypothetical scenario. Suppose that the data generating mechanism was different, and the data were generated from a stratified random sample of the population, where the probability of disease varies by stratum and the sampling probabilities vary by stratum. For instance, suppose the sampling was stratified by gender, where males were oversampled.

Would the binomial model described in question 1 still be appropriate for estimating the proportion of diseased positive individuals in the population within exposure groups? (Model the number of disease positive individuals among all exposed individuals in the sample using the binomial distribution; and model the number of disease positive individuals among the unexposed in the sample using a binomial distribution).

No, the binomial model described in question 1 would not be appropriate. We would need more parameters other than p{e+} and p{e-}. This is because the probability of disease varies by stratum.

4. Now, we examine the risk difference between the exposed and unexposed populations. Estimate the risk difference for the disease and construct a corresponding large-sample 95% confidence interval. Calculate the risk difference as the proportion of diseased individuals in the exposed minus the proportion of diseased individuals in the unexposed.

Since there were several versions of the two-by-two tables, different people will have different results. In the example version above, the correct answer is

```
. csi 48 48 252 252

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        48          48  |         96
        Noncases |       252         252  |        504
-----------------+------------------------+------------
           Total |       300         300  |        600
                 |                        |
            Risk |       .16         .16  |        .16
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |                 0      |   -.0586681    .0586681
      Risk ratio |                 1      |    .6930344     1.44293
 Attr. frac. ex. |                 0      |   -.4429299    .3069656
 Attr. frac. pop |                 0      |
                 +-------------------------------------------------
                               chi2(1) =     0.00  Pr>chi2 = 1.0000
```

Risk Difference: 0
Confidence Interval:
Lower Bound: -.0586681
Upper Bound: .0586681

5.Conduct a two-sample proportion test that the risk difference is equal to zero (versus the alternative that the risk difference is not equal to zero) at the 0.05 level of significance.

What is the absolute value of the test statistic?

Since there were several versions of the two-by-two tables, different people will have different results. In the example version above, the correct answer is Test Statistic = 0.0

```
. prtesti 300 48 300 48, count

Two-sample test of proportions                     x: Number of obs =      300
                                                   y: Number of obs =      300
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |        .16    .021166                      .1185154    .2014846
           y |        .16    .021166                      .1185154    .2014846
-------------+----------------------------------------------------------------
        diff |          0   .0299333                     -.0586681    .0586681
             |  under Ho:   .0299333     0.00   1.000
------------------------------------------------------------------------------
        diff = prop(x) - prop(y)                                  z =   0.0000
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.5000         Pr(|Z| < |z|) = 1.0000          Pr(Z > z) = 0.5000
```

What is the distribution of the test statistic under the null hypothesis?

a) Standard Normal
b) t-distribution
c) Binomial

What is the p-value?
Since there were several versions of the two-by-two tables, different people will have different results. In the example version above, the correct answer is 1.000

What is your conclusion? (enter the letter of your best answer from the options listed below)

a) We have evidence that the risk difference is not equal to 0.
b) We do not have evidence that the risk difference is different from zero.
c) None of the above.

Since there were several versions of the two-by-two tables, different people will have different results. Depending on your version of the exam, the correct answer was either (a) or (b). In the example version above, the correct answer is (b) We do not have evidence that the risk difference is different from zero.

6. Rather than testing that the risk difference is equal to 0 (as in question 5), could you have conducted a Pearson-chi square test to test for an association between disease and exposure?

a) Yes
b) No

7. What is the value for the Crude Risk Ratio, comparing exposed subjects to non-exposed subjects?
Since there were several versions of the two-by-two tables, different people will have different results. In the example version above, the correct answer is
Risk Difference= (48/300) - (48/300) = 0.

8. Using the Mantel-Haenszel formula, what is the value for the sex-adjusted Risk Ratio, comparing exposed subjects to non-exposed subjects?
Since there were several versions of the two-by-two tables, different people will have different results. In the example version above, the correct answer is
RR_{MH} =((36*288/360)+(12*216/240)) / ((144*72/360)+(108*24/240))=1.0

9.Using the total data as a standard population, what is the value for the Standardized Risk Ratio?
Since there were several versions of the two-by-two tables, different people will have different results. In the example version above, the correct answer is Risk Ratio=1.

10. Is sex a confounder in this study? (enter the letter of your best answer from the options listed below)

a) Yes, because the crude RR equals the sex-adjusted RR
b) No, because the crude RR equals the sex-adjusted RR
c) Yes, because the crude RR does not equal the sex-adjusted RR
d) No, because the crude RR does not equal the sex-adjusted RR
e) Yes, because the RR among the males equals the RR among the females
f) No, because the RR among the males equals the RR among the females

Since the RR is the same for males and females, we conclude that there is no confounding by sex.

11.Using the Risk Ratio as a measure of association, is sex an effect modifier in this study?

a) Yes, because the crude RR equals the sex-adjusted RR
b) No, because the crude RR equals the sex-adjusted RR
c) Yes, because the crude RR does not equal the sex-adjusted RR
d) No, because the crude RR does not equal the sex-adjusted RR
e) Yes, because the RR among males equals the RR among females
f) No, because the RR among males equals the RR among females

Part 3: Study Design

Select the most appropriate study design for each of the following questions. (Note: All study design options may not be used and each design option can be used more than once.)

1.A study is done to examine the association between a mother's education and risk of a congenital heart defect in her offspring. The investigator enrolls a group of mothers of babies with birth defects and a group of mothers of babies without birth defects. The mothers are then asked a series of questions about their education.
Case-control study

2. A study on the association of coffee consumption and performance on a memory test randomly assigns half of the enrolled subjects to drink coffee one hour before taking the memory test and the other half to not drink coffee one hour before taking the memory test.</p>
Randomized clinical trial

3. A study examining the association between meat consumption and heart disease compares the average number of kilograms of meat consumed per person for 50 different countries to the incidence rate of heart disease in the same 50 countries.
Ecological study

4. An investigator enrolls a group of healthy individuals and distributes questionnaires to collect information on sex and blood type. The investigator then examines the association between sex and blood type
Cross-sectional study: The exposure and outcome were measured simultaneously.

5. A study describes a group of hospital patients all of whom suffer from migraine with aura and experienced an ischemic stroke.
Case series: Everyone included in the study has both the exposure and the outcome of interest.

6. A study recruits a group of college graduates. At the time of recruitment, participants provide a blood sample that is immediately stored and they complete a questionnaire about lifestyle behaviors. The participants are followed over time to see who develops Parkinson's Disease. Two studies are performed. Classify each study.

a) In one study, the researchers compare people reporting regular physical activity to those who are sedentary.
Prospective cohort study

b) In another study, the stored blood sample for each case is compared to the blood sample for a non-case that was alive and at risk on the day that the case was diagnosed with Parkinson's disease. The blood samples are analyzed to see if serum levels of Vitamin D are associated with the risk of developing Parkinson's disease.
Nested case-control study

7. A researcher uses a database of medical records to identify a group of retired factory workers. He reviews each person's medical records to follow their factory exposures over time and see which of these subjects has developed skin cancer in the past 25 years.
Retrospective cohort study