Homework 1 Solutions

1) For this question, I used the "csi" command with the "or" option so an odds ratio would also be calculated. The csi command is appropriate for cohort or cross-sectional studies. (Note that the cci command is for case-control studies and only gives odds ratios.)

```
. csi 61 27 75 312, or

                 |   Exposed   Unexposed  |      Total
-----------------+------------------------+------------
           Cases |        61          27  |         88
        Noncases |        75         312  |        387
-----------------+------------------------+------------
           Total |       136         339  |        475
                 |                        |
            Risk |  .4485294     .079646  |   .1852632
                 |                        |
                 |      Point estimate    |    [95% Conf. Interval]
                 |------------------------+------------------------
 Risk difference |         .3688834       |    .2804678     .457299
      Risk ratio |         5.631536       |    3.748488    8.460531
      Odds ratio |         9.398519       |    5.610924    15.73917
                 +-----------------------------------------------
                              chi2(1) =    87.50  Pr>chi2 = 0.0000
```

a) This is a cross-sectional sample so we use a chi-square test of independence. We may write Ho: P(HIV | IVDU) = P(HIV | not IVDU). One issue to recall with this design is that since we have a cross-sectional sample, we can talk about the probability, or risk, of "having" the disease (HIV+) rather than the risk of "acquiring" the disease. Stated equivalently, we talk about disease prevalence, not incidence.

We see that the chi-square statistic for testing this hypothesis is 87.50 and the p-value for this test is very small ($p < 0.001$). Therefore, we conclude the risk of HIV infection in this population of women is associated with a history of IVDU use.

b) The risk of HIV infection in women in this sample who do not report a history of IVDU use is 27/339 = 0.08 while the risk of HIV infection among women in this sample who report a history of IVDU use is 61/136 = 0.45. Therefore, the estimated difference in risk (the increased risk among women reporting a history of IVDU use compared to those who do not report a history of IVDU use) is 0.37. We are highly (95%) confident that the true difference in risk between these populations is between 0.28 and 0.46. A summary sentence may be:

"Among women that report IV drug use, we find 45 percent seropositive while the risk among women that do not report IV drug use we find 8 percent seropositive. The increased risk associated with IV drug use can be summarized by the risk difference of 37 percent (95% confidence interval: 28%, 46%). These results suggest a relatively high prevalence of HIV among incarcerated women that do not report IV drug use (8 percent), but an additional 37 percent are seropositive among the women that do report IV drug use."

c) The risk ratio (relative risk) is the probability of HIV+ among the IVDU women relative to the probability of HIV+ among the non-IVDU women. The risk if HIV infection in women in this sample who do not report a history of IVDU use is 27/339

= 0.08 while the risk of HIV infection among women in this sample who report a history of IVDU use is 61/136 = 0.45. Therefore, the estimated relative risk is 0.45/0.08 = 5.6 with a 95% confidence interval of (3.7, 8.5). A summary sentence might look quite similar to the previous:

"Among women that report IV drug use, we find 45 percent seropositive while the risk among women that do not report IV drug use we find 8 percent seropositive. The increased risk associated with IV drug use can be summarized by the risk ratio of 5.6 (95% confidence interval: 3.7, 8.5). These results suggest a relatively high prevalence of HIV among incarcerated women that do not report IV drug use (8 percent), but more than a 5.5-fold increase in the likelihood of being seropositive among the women that do report IV drug use."

d) The odds ratio is similar to the risk ratio with the difference being that the odds (i.e., p/(1-p) are compared for the two groups. The odds ratio is estimated as follows:

The estimated odds among women that report IV drug use was 0.45/(1-0.45)=0.82 seropositive while among women that do not report IV drug use, the odds was 0.08/0.92=0.09 seropositive. The increased risk associated with IV drug use can be summarized by the odds ratio 9.4 (95% confidence interval 5.6, 15.7) which compares the odds of seropositivity among IDVU women relative to the odds among non-IVDU women. These results suggest a relatively high prevalence of HIV among incarcerated women that do not report IV drug use (8%), but more than a 9-fold increase in the odds of being seropositive among women that also report IV drug use.

2) I read the data in using the HivnetWide.do file.

a) Q: Did randomization work?

To answer this question, we compare baseline measurements of the ICgroup==0 and the ICgroup==1 groups. Specifically, for the nurse0 item we can use the tabulate command: "`tab ICgroup nurse0, row chi`"

```
  Informed | allocation knowledge
   Consent |        at t=0
     group |         0          1 |     Total
-----------+----------------------+----------
         0 |       192        308 |       500
           |     38.40      61.60 |    100.00
-----------+----------------------+----------
         1 |       191        309 |       500
           |     38.20      61.80 |    100.00
-----------+----------------------+----------
     Total |       383        617 |     1,000
           |     38.30      61.70 |    100.00

          Pearson chi2(1) =   0.0042   Pr = 0.948
```

From this summary, we find that 61.6 percent of the control group answered correctly at baseline and 61.8 percent of the intervention group answered correctly. The groups are comparable at baseline (as would be expected by randomization).

b) Comparing the 6 month response for the two groups, our hypotheses are

Ho: $P_{ICgroup=0} = P_{ICgroup=1}$
Ha: $P_{ICgroup=0} \neq P_{ICgroup=1}$

where P is the probability estimated at the six month visit (i.e., nurse6 = 1). We use the "cs" command to generate a risk difference (RD); you may also present results in terms of a relative risk (RR) or an odds ratio (OR).

```
. cs nurse6 ICgroup, or

                | Informed Consent group |
                |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
          Cases |       417         326  |        743
       Noncases |        83         174  |        257
----------------+------------------------+-----------
          Total |       500         500  |       1000
                |                        |
           Risk |      .834        .652  |       .743
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
                |------------------------+------------------------
Risk difference |              .182      |      .12902      .23498
     Risk ratio |          1.279141      |    1.186676    1.378811
     Odds ratio |          2.681573      |    1.989742    3.613795
                +-------------------------------------------------
                              chi2(1) =    43.37  Pr>chi2 = 0.0000
```

From these displays, we see that among the ICgroup==1 subjects, 83.4 percent correctly answered the nurse item while only 65.2 percent of the control group answered correctly. The 18.2 percent improvement (95% CI 12.9%, 23.5%) attributable to the intervention is statistically significant (p-value < 0.001). We conclude that the informed consent significantly increased understanding of the randomization process (as measured by nurse6).

c) My hypotheses are

Ho: $P_0 = P_6$
Ha: $P_0 \neq P_6$

where P_t is the probability that the nurse variable is equal to 1 at time t. The basic question here is, "Did those in the intervention group improve between baseline and 6 months?". Since these data are paired (repeated binary responses for each subject) we will use "mcc" command:

```
. mcc nurse6 nurse0 if ICgroup == 1

                | Controls               |
Cases           |   Exposed   Unexposed  |      Total
----------------+------------------------+-----------
        Exposed |       271         146  |        417
      Unexposed |        38          45  |         83
----------------+------------------------+-----------
          Total |       309         191  |        500

McNemar's chi2(1) =     63.39    Prob > chi2 = 0.0000
```

```
Exact McNemar significance probability       = 0.0000

Proportion with factor
        Cases           .834
        Controls        .618     [95% Conf. Interval]
                    ---------    --------------------
        difference      .216      .1643124   .2676876
        ratio       1.349515      1.253175    1.45326
        rel. diff.   .565445      .4736866   .6572035

        odds ratio  3.842105      2.672886   5.645147
```

Note that the labels “cases” and “controls” are not appropriate for our analysis. Here, “cases” would be the observations taken at 6 months and the “controls” would be the observations taken at baseline.

McNemar’s test allows us to assess the null hypothesis: Among the ICgroup==1 subjects, the probability of answering correctly at month 6 is the same as the probability of answering correctly at baseline. We obtain a chi-square statistic of 63.4 on one degree of freedom. The p-value for this test statistic is less than 0.001 and we therefore reject the null hypothesis and conclude the probability of answering correctly at month 6 differs from the probability of answering correctly at baseline. I interpret these results to indicate that there was significant improvement in knowledge among those receiving the mock informed consent between baseline and 6 months. The odds of answering correctly at month 6 is 3.8 times higher than the odds of answering correctly at baseline for those subjects.

d) Let’s focus on who understood the knowledge item at follow-up. Did the intervention “correct” the subjects that answered incorrectly at baseline and/or “reinforce” those subjects that answered correctly at baseline. First, let’s consider the subjects that answered incorrectly at baseline:

```
. cs nurse6 ICgroup if nurse0 == 0, or

                | Informed Consent group |
                |   Exposed   Unexposed  |     Total
----------------+------------------------+----------
          Cases |       146          91  |       237
       Noncases |        45         101  |       146
----------------+------------------------+----------
          Total |       191         192  |       383
                |                        |
           Risk |  .7643979    .4739583  |   .618799
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+----------------------
 Risk difference |         .2904396       |  .1976471     .383232
      Risk ratio |         1.612796       |  1.362649    1.908863
      Odds ratio |         3.600977       |  2.326159    5.573911
                +-------------------------------------------------
                            chi2(1) =    34.24  Pr>chi2 = 0.0000
```

From this summary, we see that there was a very large improvement in knowledge (29 percent (95% CI: 20 – 38)) at the 6 month follow-up in the intervention group compared to the control group among those that answered incorrectly at baseline. Therefore, the intervention improved the “correction” of incorrect understanding. Interestingly, among the men that answered incorrectly

at baseline, the percentage responding correctly at 6 months in the control group was 47%, which is consistent with randomly guessing.

e) Among the subjects that answered correctly at baseline, we have:

```
. cs nurse6 ICgroup if nurse0 == 1, or

                | Informed Consent group |
                |   Exposed   Unexposed  |     Total
----------------+------------------------+----------
          Cases |       271         235  |       506
       Noncases |        38          73  |       111
----------------+------------------------+----------
          Total |       309         308  |       617
                |                        |
           Risk |  .8770227     .762987  |  .8200972
                |                        |
                |      Point estimate    |  [95% Conf. Interval]
                |------------------------+----------------------
Risk difference |         .1140356       |  .0540666    .1740047
     Risk ratio |         1.149459       |  1.066456    1.238923
     Odds ratio |         2.215342       |  1.444976    3.395817
                +-----------------------------------------------
                            chi2(1) =    13.60  Pr>chi2 = 0.0002
```

Here, too, we find an effect of the intervention. Among the 309 intervention subjects that answered correctly at baseline, 271/309 = 87.7% answered correctly at follow-up, compared to only 235/308=76.2% of the control subjects. We see that there was a more modest (but still statistically significant result: p = .0002) improvement in knowledge (11 percent (95% CI: 5 – 17)) at 6 months in the intervention group compared to the control group among those that answered correctly at baseline. Thus, we may conclude that the intervention either reinforced the baseline knowledge among these men or improved knowledge among those that may have simply guessed correctly at baseline.

3) Regression thinking

a) In question 2(d), we can write the regression model:

$P[\text{nurse6} = 1] = E[\text{nurse6}] = \beta_0 + \beta_1 \text{ Icgroup}$

keep in mind that we are restricting our analyses (and interpretation!) to those subjects that answered incorrectly at baseline (i.e., nurse0==0). In this model, we have:

β_0 = the probability answering correctly at month 6 in the control group (among subjects that answered incorrectly at baseline),
β_1 = the probability answering correctly at month 6 in the intervention group MINUS the probability answering correctly at month 6 in the control group (among subjects that answered incorrectly at baseline). Thus, β_1 is the "risk difference" among nurse0==0 subjects.

b) In question 2(e), we write a regression model similar to that in part 3(a):

$P[nurse6 = 1] = E[nurse6] = \beta_0 + \beta_1$ Icgroup

keep in mind that we are restricting our analyses to subjects that answered "correctly" at baseline (i.e., nurse0==1). In this model, we have:

β_0 = the probability answering correctly at month 6 in the control group (among subjects that answered correctly at baseline),
β_1 = the probability answering correctly at month 6 in the intervention group MINUS the probability answering correctly at month 6 in the control group (among subjects that answered correctly at baseline).
Thus, β_1 is the "risk difference" among nurse0==1 subjects.

c) We can combine these models (i.e., include all subjects regardless of their baseline response to the nurse question) as:

$P[nurse6 = 1] = \beta_0 + \beta_1$ Icgroup $+ \beta_2$ nurse0 $+ \beta_3$ Icgroup*nurse0

In this model, we have:

β_0 = the probability answering correctly at month 6 in the control group (among subjects that answered incorrectly at baseline),
β_1 = the probability answering correctly at month 6 in the intervention group MINUS the probability answering correctly at month 6 in the control group (among subjects that answered incorrectly at baseline).
$\beta_0 + \beta_2$ = the probability answering correctly at month 6 in the control group among subjects that answered correctly at baseline,
$\beta_1 + \beta_3$ = the probability answering correctly at month 6 in the intervention group MINUS the probability answering correctly at month 6 in the control group among subjects that answered correctly at baseline.

We find β_1 to be the "treatment effect" among subjects with nurse0==0, and $\beta_1 + \beta_3$ to be the "treatment effect" among subjects with nurse0==1. (Here, "treatment effect" refers to the difference in the average response among intervention subjects minus the average response among control subjects.)

Additionally, we have:

β_2 = for the control group, the probability of answering correctly at month 6 between subjects who answered correctly at baseline MINUS the probability of answering correctly at month 6 who answered incorrectly at baseline.
β_3 = the difference in the effect of treatment comparing subjects that answered correctly at baseline (nurse0==1), to the effect of treatment among subjects that answered incorrectly at baseline (nurse0==0). Again,

the “effect of treatment” specifically means the difference between the probability answering correctly for the intervention subjects (ICgroup==1) and the probability answering correctly for the control subjects (ICgroup==0).

d) The standard linear model assumptions are:

   i. Linearity – not a concern since we are using predictor variables that are categorical (binary).

   ii. Normality – clearly the data (i.e., the binary outcome variable of answering correctly at month 6) are not normal. However, this assumption is not necessary for application to a large dataset (more than 100 observations).

   iii. Equal variance – the model errors, e = y (0/1 data) – fitted model, do not have equal variance for binary data.

   Therefore, we clearly violate two of these assumptions, so fitting a linear model yields estimated coefficients with useful interpretations, but not without a notable cost (i.e., statistical inference will be suspect). We will develop logistic regression methods to allow regression analysis on binary outcomes without such deficiencies.