Epi 202 SAS Code Documentation: Crude and Stratified Tabular Analyses

This document describes the mechanics of and how to use the following SAS code files:

* epi202\_count\_data.sas
* epi202\_person\_time.sas
* epi202\_case\_control.sas

As researchers, our data is usually not given to us in tabular format. Instead, we work with datasets that have row(s) of observations for each of the participants in our study and columns with different variables we have collected information on from our participants. An example of such a dataset, evans\_example\_dat.csv has been provided to you on the course website. This dataset has been adapted from the Evans County dataset provided by the R package lbreg. In the version provided to you, we have kept the three variables of interest and additionally generated a variable for person-time and case-control selection so that you can will be able to conduct closed cohort, open cohort, and case-control analyses all on the same example data. Briefly, the Evans County study followed white males for incident coronary heart disease for 7 years. For the purposes of these examples, we are going to assume no loss to follow-up or competing risks.

For your reference, the variables in the dataset are coded as follows:

* CHD: a binary indicator of outcome status where 1=incident CHD case and 0=non-case
* HTN: a binary indicator of exposure status where 1=hypertensive and 0=not hypertensive at baseline
* SMK: a binary of smoking status where 1=smoker and 0=non-smoker at baseline
* person-time: a continuous variable representing amount of person-time under follow-up prior to a CHD event or administrative censoring at the end of follow-up (maximum=7)
* caco: a binary indicator of case status for participants selected into a nested case-control study; ’case’=case and ’control’=control. For participants not selected to the case-control study, it is set to ‘NA’

1) Analyses for count data

*Crude Analyses*

The SAS code to conduct analyses, both crude and stratified, for count data, such as might arise in a closed-cohort setting, is contained in the file epi202\_count\_data.sas. To begin these analyses, we have to import our data using proc import. When doing this, we create a dataset in the working directory called ‘dat’ that is used for all of our subsequent analyses. In order for proc import to work properly, we must be careful to give the correct file location (in this example, the ‘evans\_example\_data.csv’ file is stored in the folder ‘Epi 202 2018’ in my P drive) as well as the file type (here, a csv file).

/\*Call in data\*/

proc import out= WORK.dat DATAFILE= "P:\Epi 202 2018\evans\_example\_data.csv"

DBMS=csv REPLACE;

GETNAMES=YES;

RUN;

In the next step, we create indicator variables to represent the following types of individuals we observe: exposed cases, exposed non-cases, unexposed cases, and unexposed non-cases. To do this, we use ‘if’ statements and the values of our exposure and outcome variables. In the example below, I first drop individuals with a missing value for the exposure variable (HTN) and then categorize individuals based on the exposure variable (HTN) and the outcome variable (CHD).

**To use this code with a different exposure and/or outcome, make sure that your variables are coded as 0 (unexposed/non-case) or 1 (exposed/case) and simply change ‘HTN’ to the name of your exposure variable and/or ‘CHD’ to the name of your outcome variable. After this, you should not need to make any further changes to the rest of the crude analysis section in order for the code to run correctly.**

data dat;

set dat;

if HTN^=.;

if HTN=1 and CHD=1 then exposed\_case=1;

if HTN=1 and CHD=0 then exposed\_noncase=1;

if HTN=0 and CHD=1 then unexposed\_case=1;

if HTN=0 and CHD=0 then unexposed\_noncase=1;

run;

Next, we need to determine how many individuals we have in each of our exposure and outcome combinations. To do this, we can use the sum option in proc means and sum over the indicators we created in the previous data step. We can use the ods output command to store these sums in a new dataset called crude\_closed.

proc means data=dat sum maxdec=2;

ods output Summary=crude\_closed;

var exposed\_case exposed\_noncase unexposed\_case unexposed\_noncase;

run;

If we check the ‘Results Viewer’ window, our procs means output will look like:

A screenshot of a cell phone

Description generated with very high confidence

Before we do our crude analyses, we can clean up our dataset to make it easier to use. To do this, we’ll rename and label our count variables as ‘a’, ‘b’, ‘c’, and ‘d’ to match up with the formulas on the Epi 202 Roadmap. We’ll also drop the variables we don’t need (all of which start with ‘VNAME\_’).

data crude\_closed;

set crude\_closed;

rename exposed\_case\_Sum=a

exposed\_noncase\_Sum=c

unexposed\_case\_Sum=b

unexposed\_noncase\_Sum=d;

drop VNAME\_exposed\_case VNAME\_unexposed\_case VNAME\_exposed\_noncase VNAME\_unexposed\_noncase;

run;

data crude\_closed;

set crude\_closed;

label a="a"

b="b"

c="c"

d="d";

run;

Now that we have our dataset cleaned up, we can begin the calculations for our crude analyses using another data step. First, we begin by creating variables for N1, N0, M1, M0, and T. Next, we calculate the cumulative incidence ratio (CIR), variance of the CIR, and corresponding 95% confidence interval. To do this, we create a series of new variables. First, we calculate the CIR using a, N1, b, and N0. Then we similarly calculate the variance of the CIR and each the lower (CIR\_confint\_lower) and upper (CIR\_confint\_upper) bounds of the confidence interval as their own variables. Lastly, we concatenate the CI bounds to a single variable (CIR\_confint). This process in then repeated using the appropriate formulas for the cumulative incidence difference and a comparable set of variables are produced.

Additionally, we calculate the Z2 statistic (‘Z\_squared’) and corresponding p-value for the test of no association (‘p\_testassoc’). To do this, we store the expected value under the null and the variance of the test statistic as new variables, which are then used to produce the Z\_squared variable representing our test statistic. Lastly, we use the value of test statistic to calculate the corresponding p-value based on the X2 distribution with 1 degree of freedom.

data crude\_closed;

set crude\_closed;

/\*Specify count related variables\*/

N1=a+c;

N0=b+d;

M1=a+b;

M0=c+d;

T=N1+N0;

/\*Calculate CIR & 95% CI\*/

CIR=(a/N1)/(b/N0);

Variance\_CIR\_confint=(c/(a\*N1))+(d/(b\*N0));

CIR\_confint\_lower=round(exp(log(CIR)-1.96\*sqrt(Variance\_CIR\_confint)),0.001);

CIR\_confint\_upper=round(exp(log(CIR)+1.96\*sqrt(Variance\_CIR\_confint)),0.001);

CIR\_confint=catx(',', CIR\_confint\_lower, CIR\_confint\_upper);

/\*Calculate CID & 95% CI\*/

CID=(a/N1)-(b/N0);

Variance\_CID\_confint=((a\*c)/(N1\*\*3))+((b\*d)/(N0\*\*3));

CID\_confint\_lower=round(CID-1.96\*sqrt(Variance\_CID\_confint), 0.001);

CID\_confint\_upper=round(CID+1.96\*sqrt(Variance\_CID\_confint), 0.001);

CID\_confint=catx(',', CID\_confint\_lower, CID\_confint\_upper);

/\*Test of Association\*/

Expected\_null=(N1\*M1)/T;

Variance\_testassoc=(M1\*M0\*N1\*N0)/(T\*\*3);

Z\_squared=((a-Expected\_null)\*\*2)/Variance\_testassoc;

p\_testassoc=1-probchi(Z\_squared, 1);

run;

In the last portion of this section of code, we print the CIR, the CID, and their 95% confidence intervals, which will then appear in your Results Viewer window. Similarly, we print the test statistic and p-value for the test of no association.

proc print data=crude\_closed;

var CIR CIR\_confint CID CID\_confint;

run;

We can view our printed CIR, CID, and corresponding confidence intervals in the Results Viewer:

A screenshot of a cell phone

Description generated with very high confidence

proc print data=crude\_closed;

var Z\_squared p\_testassoc;

run;

Which we can also view in the Results Viewer window:

A screenshot of a cell phone

Description generated with very high confidence

*Stratified Analyses*

We can do stratified analyses in much the same as we just did our crude analyses. In order to conduct stratified analyses, it is important to make sure to appropriately account for our stratification factor(s) before we get the counts of individuals in each combination of exposure and outcome. To do this, we can use the same dataset ‘dat’ that we produced in the crude analysis code and sort our data by the stratification factor(s) and before using a similar proc means step as above, but this time we must add the additional ‘by’ command to get sums within each level of our stratification factor. In the example below, I am stratifying by smoking status, which is represented by the variable ‘SMK’. Note that this step is using the same dataset, ‘dat’, that we created above & which already contains the variables ‘exposed\_case’, ‘exposed\_noncase’, ‘unexposed\_case’, and ‘unexposed\_noncase’. In the proc means step, we create a new outputted dataset called ‘stratified\_closed’ that we will use for our stratified analyses after renaming and labelling our variables (this time with the suffix ‘\_i’ to represent that the counts are stratum-specific). **As above, the only thing that you need to change for this code to run smoothly with different data is the name of your stratification variable, which should replace each instance of ‘SMK’.**

/\*Sort by stratification factor(s)\*/

proc sort data=dat;

by SMK;

run;

/\*Get counts within each stratum\*/

proc means data=dat sum maxdec=2;

ods output Summary=stratified\_closed;

var exposed\_case exposed\_noncase unexposed\_case unexposed\_noncase;

by SMK;

run;

/\*Rename variables to a\_i, b\_i, c\_i, and d\_i\*/

data stratified\_closed;

set stratified\_closed;

rename exposed\_case\_Sum=a\_i

exposed\_noncase\_Sum=c\_i

unexposed\_case\_Sum=b\_i

unexposed\_noncase\_Sum=d\_i;

drop VNAME\_exposed\_case VNAME\_unexposed\_case VNAME\_exposed\_noncase VNAME\_unexposed\_noncase;

run;

/\*Label variables appropriately\*/

data stratified\_closed;

set stratified\_closed;

label a\_i="a\_i"

b\_i="b\_i"

c\_i="c\_i"

d\_i="d\_i";

run;

If we check the Results Viewer window, we can see the results of the proc means procedure stratified by smoking status:

A screenshot of a cell phone

Description generated with very high confidence

Now that we have a cleaned-up dataset with 1 row per stratum, we can start to set up the calculations for our stratified analyses. As with the crude analyses above, we will do this by using a data step to create a series of new variables. Once we have each of these terms calculated, we can use a few simple proc sql commands to sum over each of the strata to get our summary statistics and test statistics.

In this data step, we begin by creating variables for N1i, N0i, M1i, M0i, and Ti and then calculating the CIR, variance, and 95% confidence interval for each stratum. We then set up the **future** calculation of our Maentel-Hanszel summary CIR and 95% confidence interval by creating variables representing each stratum’s contribution to the MH weight (MH\_w\_i), variance numerator (MH\_variance\_numerator\_term\_i), and denominator (MH\_variance\_denom1\_i & MH\_variance\_denom2\_i). We will **later** use these values to get the MH CIR and confidence interval.

In the next portion of this data step, we create a comparable set of variables to get the stratum-specific CIDs, variances, and 95% confidence intervals. We also set up the future calculation of the inverse variance (IV) weighted summary CID and confidence interval. To do this, we create a variable representing each stratum’s weight (IV\_w\_i) and contribution to the numerator of the summary statistic (IV\_numerator\_term\_i). We will **later** use these values to get the IV CID and confidence interval.

Lastly, we set up variable that will be used in the calculation of the test of no association. These variables included the expected number of cases in each stratum under the null (expected\_value\_i) and each stratum’s contribution to the variance (variance\_term\_i).

data stratified\_closed;

set stratified\_closed;

/\*Specify count related variables\*/

N1\_i=a\_i+c\_i;

N0\_i=b\_i+d\_i;

M1\_i=a\_i+b\_i;

M0\_i=c\_i+d\_i;

T\_i=N1\_i+N0\_i;

/\*Stratum-specific: Calculate CIR & 95% CI\*/

CIR\_i=(a\_i/N1\_i)/(b\_i/N0\_i);

Variance\_CIR\_confint\_i=(c\_i/(a\_i\*N1\_i))+(d\_i/(b\_i\*N0\_i));

CIR\_confint\_lower\_i=round(exp(log(CIR\_i)-1.96\*sqrt(Variance\_CIR\_confint\_i)),0.001);

CIR\_confint\_upper\_i=round(exp(log(CIR\_i)+1.96\*sqrt(Variance\_CIR\_confint\_i)),0.001);

CIR\_confint\_i=catx(',', CIR\_confint\_lower\_i, CIR\_confint\_upper\_i);

/\*Summary: Calculate MH CIR & 95% CI\*/

MH\_w\_i=(b\_i\*N1\_i)/T\_i;

MH\_numerator\_term\_i=(a\_i\*N0\_i)/T\_i;

MH\_variance\_numerator\_i=((M1\_i\*N1\_i\*N0\_i)-(a\_i\*b\_i\*T\_i))/(T\_i\*\*2);

MH\_variance\_denom1\_i=(a\_i\*N0\_i)/T\_i;

MH\_variance\_denom2\_i=(b\_i\*N1\_i)/T\_i;

/\*Stratum-specific: Calculate CID & 95% CI\*/

CID\_i=(a\_i/N1\_i)-(b\_i/N0\_i);

Variance\_CID\_confint\_i=((a\_i\*c\_i)/(N1\_i\*\*3))+((b\_i\*d\_i)/(N0\_i\*\*3));

CID\_confint\_lower\_i=round(CID\_i-1.96\*sqrt(Variance\_CID\_confint\_i), 0.001);

CID\_confint\_upper\_i=round(CID\_i+1.96\*sqrt(Variance\_CID\_confint\_i), 0.001);

CID\_confint\_i=catx(',', CID\_confint\_lower\_i, CID\_confint\_upper\_i);

/\*Summary: Calculate Inverse Variance CID & 95% CI\*/

IV\_w\_i=((N1\_i\*\*3)\*(N0\_i\*\*3))/(((N0\_i\*\*3)\*a\_i\*c\_i)+((N1\_i\*\*3)\*b\_i\*d\_i));

IV\_numerator\_term\_i=IV\_w\_i\*CID\_i;

drop CIR\_confint\_lower\_i CIR\_confint\_upper\_i CID\_confint\_lower\_i CID\_confint\_upper\_i;

/\*Test of No association\*/

expected\_value\_i=(N1\_i\*M1\_i)/T\_i;

variance\_term\_i=(M1\_i\*M0\_i\*N1\_i\*N0\_i)/(T\_i\*\*3);

run;

Once the above data step is complete, we can use our data set and a few simple ‘proc sql’ calls to calculate our summary statistics and test statistics. We start by calculating the MH summary CIR and corresponding confidence interval. The first portion of the ‘select’ command calculates the MH CIR as the weighted average of the stratum-specific CIRs using the stratum-specific CIRs and weights stored in our dataset. The next two lines of respectively calculate the lower and upper bounds of the 95% CI for the MH CIR by summing over the variables we created to represent each stratum’s contribution to the MH CIR itself as well as to the variance of the MH CIR. After running this chunk of code, the MH CIR & 95% CI will be outputted to your SAS output window. **You should not need to edit any of these formulas.**

proc sql;

title 'MH CIR';

select sum(MH\_numerator\_term\_i)/sum(MH\_w\_i) as MH\_CIR,

exp(log(sum(MH\_numerator\_term\_i)/sum(MH\_w\_i))-1.96\*sqrt(sum(MH\_variance\_numerator\_i)/(sum(MH\_variance\_denom1\_i)\*sum(MH\_variance\_denom2\_i)))) as MH\_CIR\_lower,

exp(log(sum(MH\_numerator\_term\_i)/sum(MH\_w\_i))+1.96\*sqrt(sum(MH\_variance\_numerator\_i)/(sum(MH\_variance\_denom1\_i)\*sum(MH\_variance\_denom2\_i)))) as MH\_CIR\_upper

from stratified\_closed;

quit;

We can see the MH CIR in the Results Viewer:

A screenshot of a cell phone

Description generated with very high confidence

Similarly, we can next calculate the inverse-variance weighted CID and corresponding 95% CI. The first line of the proc sql select command calculates the summary CID while the next two obtain the lower and upper bounds of the 95% CI respectively.

/\*Calculate IV CID & 95% CI\*/

proc sql;

title 'Inverse Variance CID';

select sum(IV\_numerator\_term\_i)/sum(IV\_w\_i) as IV\_CID,

sum(IV\_numerator\_term\_i)/sum(IV\_w\_i)-1.96\*sqrt(1/sum(IV\_w\_i)) as IV\_CID\_lower,

sum(IV\_numerator\_term\_i)/sum(IV\_w\_i)+1.96\*sqrt(1/sum(IV\_w\_i)) as IV\_CID\_lower

from stratified\_closed;

quit;

We can see the summary CID in the Results Viewer window:

A screenshot of a cell phone

Description generated with very high confidence

Similarly, we can use our stored variables to calculate the Z2 statistic for the test of association & corresponding p-value.

/\*Test of Association\*/

proc sql;

title 'Test of No Exposure-Disease Association';

select ((sum(a\_i)-sum(expected\_value\_i))\*\*2)/sum(variance\_term\_i) as Z\_squared,

1-probchi(((sum(a\_i)-sum(expected\_value\_i))\*\*2)/sum(variance\_term\_i),1) as p,

1 as DF

from stratified\_closed;

quit;

We can see the results of the test of association in the Results Viewer:

A screenshot of a cell phone

Description generated with very high confidence

Lastly, we can use proc sql to store our summary statistics & use these to calculate the test of heterogeneity for the CIR:

/\*Test of Heterogeneity: CIR\*/

/\*Store MH CIR for H statistic calculation\*/

proc sql noprint;

select sum(MH\_numerator\_term\_i)/sum(MH\_w\_i) into :MH\_CIR

from stratified\_closed;

quit;

proc sql;

title 'Tests of Homogeneity: CIR';

select sum(((log(CIR\_i)-log(&MH\_CIR))\*\*2)/Variance\_CIR\_confint\_i) as H\_CIR,

1-probchi(sum(((log(CIR\_i)-log(&MH\_CIR))\*\*2)/Variance\_CIR\_confint\_i), (count(\*)-1)) as p,

count(\*)-1 as DF

from stratified\_closed;

quit;

We can see the results in the Result Viewer:

A screenshot of a cell phone

Description generated with very high confidence

And for the CID:

/\*Test of Heterogeneity: CID\*/

/\*Store IV CID for H statistic calculation\*/

proc sql noprint;

select sum(IV\_numerator\_term\_i)/sum(IV\_w\_i) into :IV\_CID

from stratified\_closed;

quit;

proc sql;

title 'Tests of Homogeneity: CID';

select sum(((CID\_i-&IV\_CID)\*\*2)/Variance\_CID\_confint\_i) as H\_CID,

1-probchi(sum(((CID\_i-&IV\_CID)\*\*2)/Variance\_CID\_confint\_i), (count(\*)-1)) as p,

count(\*)-1 as DF

from stratified\_closed;

quit;

We can see the results in the Results Viewer window:

A screenshot of a cell phone

Description generated with very high confidence

2) Analyses for person-time data

*Crude Analyses*

The SAS code to conduct analyses, both crude and stratified, for person-time data, such as might arise in an open-cohort setting, is contained in the file epi202\_person\_time.sas. To begin these analyses, we have to import our data using proc import. When doing this, we create a dataset in the working directory called ‘dat’ that is used for all of our subsequent analyses. In order for proc import to work properly, we must be careful to give the correct file location (in this example, the file is stored in the folder ‘Epi 202 2018’ in my P drive) as well as the file type (here, a csv file).

/\*Call in data\*/

proc import out= WORK.dat DATAFILE= "P:\Epi 202 2018\evans\_example\_data.csv"

DBMS=csv REPLACE;

GETNAMES=YES;

RUN;

Next, we use a data step to drop observations where the exposure status is missing & to create indicators for exposed cases, unexposed cases, exposed person-time, and unexposed person-time. In this example, I will be looking at the association between hypertension (HTN) and cardiovascular disease (CHD). **If you would like to use this code in a different dataset, you only need to make sure your variables (exposure, outcome, and person-time) are similarly coded (1=exposed/event; 0=unexposed/non-event) and write over the names of the variables I have used here – after that, the rest of code should work on the same.**

data dat;

set dat;

if HTN^=.;

if HTN=1 and CHD=1 then exposed\_case=1;

if HTN=0 and CHD=1 then unexposed\_case=1;

if HTN=1 then exposed\_PT=person\_time;

if HTN=0 then unexposed\_PT=person\_time;

run;

Next, we need to determine how many cases have occurred for each value of exposure and how much exposed and unexposed person-time was accumulated during follow-up. To do this, we can use the sum option in proc means and sum over the indicators we created in the previous data step. We can use the ods output command to store these sums in a new dataset called crude\_open.

proc means data=dat sum maxdec=2;

ods output Summary=crude\_open;

var exposed\_case unexposed\_case exposed\_PT unexposed\_PT;

run;

Which will give us the following output in the Results Viewer window:

A screenshot of a cell phone

Description generated with very high confidence

Next, we use two data steps to rename our variables & add labels. These variables match with the variables used on the Epi 202 Roadmap, so you will be able easily see how those formulas are coded in the subsequent analyses.

data crude\_open;

set crude\_open;

rename exposed\_case\_Sum=a

unexposed\_case\_Sum=b

exposed\_PT\_Sum=N1

unexposed\_PT\_Sum=N0;

drop VNAME\_exposed\_case VNAME\_unexposed\_case VNAME\_exposed\_PT VNAME\_unexposed\_PT;

run;

data crude\_open;

set crude\_open;

label a="a"

b="b"

N1="N1"

N0="N0";

run;

We are now ready to conduct the crude analyses. We begin by creating variables for M1 and T, the other two variables we will need for our calculations. Then we calculate the incidence rate ratio (IRR), the variance of the IRR, and the corresponding 95% CI. The IRR is stored in the variable ‘IRR’ and the variance of the IRR is stored in ‘Variance\_IRR\_confint’. We then calculate the lower (‘IRR\_confint\_lower’) and upper (‘IRR\_confint\_upper’) bounds of the 95% confidence interval and concatenate these to a single variable (‘IRR\_confint’).

In the next section of this data step, we calculate a similar set of variables in order to get the incidence rate difference (IRD), it’s variance, and it’s 95% confidence interval.

\*Note: if the incidence rates are particularly small, you may adjust the number of digits to round to (currently set to round to the closest thousandth place) by adding additional 0’s prior to the 1 in the lower and upper bound calculations of the IRR and IRD.

Lastly, we can conduct the test of no association by calculating the expected value under the null (‘expected\_null’) and the variance under the null (‘variance\_testassoc’). Using these, we can then get the Z2 statistic and corresponding p-value.

data crude\_open;

set crude\_open;

label a="a"

b="b"

N1="N1"

N0="N0";

run;

data crude\_open;

set crude\_open;

/\*Specify count related variables\*/

M1=a+b;

T=N1+N0;

/\*Calculate IRR & 95% CI\*/

IRR=(a/N1)/(b/N0);

Variance\_IRR\_confint=(1/a)+(1/b);

IRR\_confint\_lower=round(exp(log(IRR)-1.96\*sqrt(Variance\_IRR\_confint)),0.001);

IRR\_confint\_upper=round(exp(log(IRR)+1.96\*sqrt(Variance\_IRR\_confint)),0.001);

IRR\_confint=catx(',', IRR\_confint\_lower, IRR\_confint\_upper);

/\*Calculate IRD & 95% CI\*/

IRD=(a/N1)-(b/N0);

Variance\_IRD\_confint=(a/(N1\*\*2))+(b/(N0\*\*2));

IRD\_confint\_lower=round(IRD-1.96\*sqrt(Variance\_IRD\_confint), 0.001);

IRD\_confint\_upper=round(IRD+1.96\*sqrt(Variance\_IRD\_confint), 0.001);

IRD\_confint=catx(',', IRD\_confint\_lower, IRD\_confint\_upper);

/\*Test of Association\*/

Expected\_null=(N1\*M1)/T;

Variance\_testassoc=(M1\*N1\*N0)/(T\*\*2);

Z\_squared=((a-Expected\_null)\*\*2)/Variance\_testassoc;

p\_testassoc=1-probchi(Z\_squared, 1);

run;

With this data step complete, we can use the proc print to print the IRR, IRD, and their confidence intervals to our SAS Results Viewer window.

proc print data=crude\_open;

var IRR IRR\_confint IRD IRD\_confint;

run;

Which should look like:

A screenshot of a cell phone

Description generated with very high confidence

Similarly, we can print the test statistic and p-value for our test of no association:

proc print data=crude\_open;

var Z\_squared p\_testassoc;

run;

Which should look like:

A screenshot of a cell phone

Description generated with very high confidence

*Stratified Analyses*

We can do stratified analyses in much the same as we just did our crude analyses. In order to conduct stratified analyses, it is important to make sure to appropriate account for our stratification factor(s) before we get the counts of cases and total person-time for each level of exposure. To do this, we first need to sort our data by the stratification factor(s) and then use a similar proc means step as above, but this time we must add the additional ‘by’ command to get sums within each level of our stratification factor. In the example below, I am stratifying by smoking status, which is represented by the variable ‘SMK’. Note that this step is using the same dataset, ‘dat’, that we created above & which contains the variables ‘exposed\_case’, ‘unexposed\_case’, ‘exposed\_PT’, and ‘unexposed\_PT’. In the proc means step, we create a new outputted datataset called ‘stratified\_open’ that we will use for our stratified analyses after renaming and labelling our variables (this time with the suffix ‘\_i’ to represent that the sums are stratum-specific). As above, the only thing that you need to change for this code to run smoothly with different data is the name of your stratification variable, which should replace each instance of ‘SMK’.

/\*Sort by stratification factor(s)\*/

proc sort data=dat;

by SMK;

run;

proc means data=dat sum maxdec=2;

ods output Summary=stratified\_open;

var exposed\_case unexposed\_case exposed\_PT unexposed\_PT;

by SMK;

run;

data stratified\_open;

set stratified\_open;

rename exposed\_case\_Sum=a\_i

unexposed\_case\_Sum=b\_i

exposed\_PT\_Sum=N1\_i

unexposed\_PT\_Sum=N0\_i;

drop VNAME\_exposed\_case VNAME\_unexposed\_case VNAME\_exposed\_PT VNAME\_unexposed\_PT;

run;

data stratified\_open;

set stratified\_open;

label a="a\_i"

b="b\_i"

N1="N1\_i"

N0="N0\_i";

run;

The MEANS procedure output in the Results Viewer window should look like:

A screenshot of a cell phone

Description generated with very high confidence

Now that we have a cleaned-up dataset with 1 row per stratum, we can start to set up the calculations for our stratified analyses. As with the crude analyses above, we will do this by creating a series of new variables. Once we have each of these terms calculated, we can use a few simple proc sql commands below to sum over each of the strata to get our summary statistics and test statistics.

In this data step, we begin by creating variables for M1i and Ti and then calculating the IRR, variance of the IRR, and 95% confidence interval of the IRR *for each stratum*. We then set up the **future** calculation of our Maentel-Hanszel (MH) summary IRR and 95% confidence interval by creating variables representing each stratum’s MH weight (MH\_w\_i) and its contribution to the variance numerator (MH\_variance\_numerator\_term\_i), and denominator (MH\_variance\_denom1\_i & MH\_variance\_denom2\_i). We will **later** use these values to get the MH IRR and confidence interval.

In the next portion of this data step, we create a comparable set of variables to get the *stratum-specific* IRDs, variances, and 95% confidence intervals. We also set up the future calculation of the inverse variance (IV) weighted summary IRD and confidence interval. To do this, we create a variable representing each stratum’s weight (IV\_w\_i) and contribution to the numerator of the summary statistic (IV\_numerator\_term\_i). We will **later** use these values to get the IV IRD and confidence interval.

Lastly, we set up the variables that will be used in the calculation of the test of no association. These variables included the expected number of cases in each stratum under the null (expected\_value\_i) and each stratum’s contribution to the variance (variance\_term\_i).

data stratified\_open;

set stratified\_open;

label a="a\_i"

b="b\_i"

N1="N1\_i"

N0="N0\_i";

run;

data stratified\_open;

set stratified\_open;

/\*Specify count related variables\*/

M1\_i=a\_i+b\_i;

T\_i=N1\_i+N0\_i;

/\*Calculate IRR & 95% CI\*/

IRR\_i=(a\_i/N1\_i)/(b\_i/N0\_i);

Variance\_IRR\_confint\_i=(1/a\_i)+(1/b\_i);

IRR\_confint\_lower\_i=round(exp(log(IRR\_i)-1.96\*sqrt(Variance\_IRR\_confint\_i)),0.001);

IRR\_confint\_upper\_i=round(exp(log(IRR\_i)+1.96\*sqrt(Variance\_IRR\_confint\_i)),0.001);

IRR\_confint\_i=catx(',', IRR\_confint\_lower\_i, IRR\_confint\_upper\_i);

/\*Summary: Calculate MH IRR & 95% CI\*/

MH\_w\_i=(b\_i\*N1\_i)/T\_i;

MH\_numerator\_term\_i=(a\_i\*N0\_i)/T\_i;

MH\_variance\_numerator\_i=(M1\_i\*N1\_i\*N0\_i)/(T\_i\*\*2);

MH\_variance\_denom1\_i=(a\_i\*N0\_i)/T\_i;

MH\_variance\_denom2\_i=(b\_i\*N1\_i)/T\_i;

/\*Calculate IRD & 95% CI\*/

IRD\_i=(a\_i/N1\_i)-(b\_i/N0\_i);

Variance\_IRD\_confint\_i=(a\_i/(N1\_i\*\*2))+(b\_i/(N0\_i\*\*2));

IRD\_confint\_lower\_i=round(IRD\_i-1.96\*sqrt(Variance\_IRD\_confint\_i), 0.001);

IRD\_confint\_upper\_i=round(IRD\_i+1.96\*sqrt(Variance\_IRD\_confint\_i), 0.001);

IRD\_confint\_i=catx(',', IRD\_confint\_lower\_i, IRD\_confint\_upper\_i);

/\*Summary: Calculate Inverse Variance IRD & 95% CI\*/

IV\_w\_i=((N1\_i\*\*2)\*(N0\_i\*\*2))/((a\_i\*(N0\_i\*\*2))+(b\_i\*(N1\_i\*\*2)));

IV\_numerator\_term\_i=IV\_w\_i\*IRD\_i;

drop IRR\_confint\_lower\_i IRR\_confint\_upper\_i IRR\_confint\_lower\_i IRR\_confint\_upper\_i;

/\*Test of No association\*/

expected\_value\_i=(N1\_i\*M1\_i)/T\_i;

variance\_term\_i=(M1\_i\*N1\_i\*N0\_i)/(T\_i\*\*2);

run;

Once the above data step is complete, we can use our data set and a few simple ‘proc sql’ calls to calculate our summary statistics and test statistics. We start by calculating the MH summary IRR and corresponding confidence interval. The first portion of the ‘select’ command calculates the MH IRR as the weighted average of the stratum-specific IRRs using the stratum-specific IRRs and weights stored in our dataset. The next two lines of respectively calculate the lower and upper bounds of the 95% CI for the MH IRR by summing over the variables we created to represent each stratum’s contribution to the MH IRR itself as well as to the variance of the MH IRR. After running this chunk of code, the MH IRR & 95% CI will be outputted to your SAS output window. **You should not need to edit any of these formulas.**

/\*Calculate MH IRR & 95% CI\*/

proc sql;

title 'MH IRR';

select sum(MH\_numerator\_term\_i)/sum(MH\_w\_i) as MH\_IRR,

exp(log(sum(MH\_numerator\_term\_i)/sum(MH\_w\_i))-1.96\*sqrt(sum(MH\_variance\_numerator\_i)/(sum(MH\_variance\_denom1\_i)\*sum(MH\_variance\_denom2\_i)))) as MH\_IRR\_lower,

exp(log(sum(MH\_numerator\_term\_i)/sum(MH\_w\_i))+1.96\*sqrt(sum(MH\_variance\_numerator\_i)/(sum(MH\_variance\_denom1\_i)\*sum(MH\_variance\_denom2\_i)))) as MH\_IRR\_upper

from stratified\_open;

quit;

The results of which should look like:

A screenshot of a cell phone

Description generated with very high confidence

Similarly, we can next calculate the inverse-variance weighted IRD and corresponding 95% CI. The first line of the proc sql select command calculates the summary IRD while the next two obtain the lower and upper bounds of the 95% CI respectively.

/\*Calculate IV IRD & 95% CI\*/

proc sql;

title 'Inverse Variance IRD';

select sum(IV\_numerator\_term\_i)/sum(IV\_w\_i) as IV\_IRD,

sum(IV\_numerator\_term\_i)/sum(IV\_w\_i)-1.96\*sqrt(1/sum(IV\_w\_i)) as IV\_IRD\_lower,

sum(IV\_numerator\_term\_i)/sum(IV\_w\_i)+1.96\*sqrt(1/sum(IV\_w\_i)) as IV\_IRD\_upper

from stratified\_open;

quit;

The results of which should look like:

A screenshot of a cell phone

Description generated with very high confidence

Similarly, we can use our stored variables to calculate the Z2 statistic for the test of association & corresponding p-value.

/\*Test of Association\*/

proc sql;

title 'Test of No Exposure-Disease Association';

select ((sum(a\_i)-sum(expected\_value\_i))\*\*2)/sum(variance\_term\_i) as Z\_squared,

1-probchi(((sum(a\_i)-sum(expected\_value\_i))\*\*2)/sum(variance\_term\_i),1) as p,

1 as DF

from stratified\_open;

quit;

The test of association results should look like:

A screenshot of a cell phone

Description generated with very high confidence

Lastly, we can use proc sql to store our summary statistics & use these to calculate the test of heterogeneity for the IRR:

/\*Test of Heterogeneity: IRR\*/

/\*Store MH IRR for H statistic calculation\*/

proc sql noprint;

select sum(MH\_numerator\_term\_i)/sum(MH\_w\_i) into :MH\_IRR

from stratified\_open;

quit;

/\*Calculate test statistic & p-value\*/

proc sql;

title 'Test of Homogeneity: IRR';

select sum(((log(IRR\_i)-log(&MH\_IRR))\*\*2)/Variance\_IRR\_confint\_i) as H\_IRR,

1-probchi(sum(((log(IRR\_i)-log(&MH\_IRR))\*\*2)/Variance\_IRR\_confint\_i), (count(\*)-1)) as p,

count(\*)-1 as DF

from stratified\_open;

quit;

Which should look like:

A screenshot of a cell phone

Description generated with very high confidence

And for the IRD:

/\*Test of Heterogeneity: IRD\*/

/\*Store IV IRD for H statistic calculation\*/

proc sql noprint;

select sum(IV\_numerator\_term\_i)/sum(IV\_w\_i) into :IV\_IRD

from stratified\_open;

quit;

/\*Calculate test statistic & p-value\*/

proc sql;

title 'Test of Homogeneity: IRD';

select sum(((IRD\_i-&IV\_IRD)\*\*2)/Variance\_IRD\_confint\_i) as H\_IRD,

1-probchi(sum(((IRD\_i-&IV\_IRD)\*\*2)/Variance\_IRD\_confint\_i), (count(\*)-1)) as p,

count(\*)-1 as DF

from stratified\_open;

quit;

Which should look like:

A screenshot of a cell phone

Description generated with very high confidence

3) Unmatched case-control data

*Crude Analyses*

The SAS code to conduct analyses, both crude and stratified, for unmatched case-control data is contained in the file epi202\_case\_control.sas. To begin these analyses, we have to import our data using proc import. When doing this, we create a dataset in the working directory called ‘dat’ that is used for all of our subsequent analyses. In order for proc import to work properly, we must be careful to give the correct file location (in this example, the file is stored in the folder ‘Epi 202 2018’ in my P drive) as well as the file type (here, a csv file).

/\*Call in data\*/

proc import out= WORK.dat DATAFILE= "P:\Epi 202 2018\evans\_example\_data.csv"

DBMS=csv REPLACE;

GETNAMES=YES;

RUN;

Next, we must restrict the dataset to those cohort members that were selected into the case-control study. We can do this by including an ‘if’ command specifying to keep those observations where caco status is not set to ‘NA’. We can also use this data step drop observations where the exposure status is missing & to create indicators for exposed cases, unexposed cases, exposed controls, and unexposed controls. In this example, I will be looking at the association between marital status (exposure) and case (outcome). If you would like to use this code in a different dataset, you only need to make sure your variables (exposure, outcome) are similarly coded (1=exposed/event; 0=unexposed/non-event) and write over the names of the variables I have used here – after that, the rest of code should work on the same.

data dat;

set dat;

if caco^=NA;\*Restricts to observations selected for case-control study;

if HTN^=.;\*Restricts to observations with non-missing exposure data;

if HTN=1 and caco=’case’ then exposed\_case=1;

if HTN=0 and caco=’case’ then unexposed\_case=1;

if HTN=1 and caco=’control’ then exposed\_control=1;

if HTN=0 and caco=’control’ then unexposed\_control=1;

run;

Next, we need to determine how many cases and controls have occurred for each level of exposure. To do this, we can use the sum option in proc means and sum over the indicators we created in the previous data step. We can use the ods output command to store these sums in a new dataset called crude\_cc.

proc means data=dat sum maxdec=2;

ods output Summary=crude\_cc;

var exposed\_case unexposed\_case exposed\_control unexposed\_control;

run;

The MEANS procedure output will look like:

A screenshot of a cell phone

Description generated with very high confidence

Next, we use two data steps to rename our variables & add labels. These variables match with the quantities used on the Epi 202 Roadmap, so you will be able easily see how those formulas are coded in the subsequent analyses.

data crude\_cc;

set crude\_cc;

rename exposed\_case\_Sum=a

unexposed\_case\_Sum=b

exposed\_control\_Sum=c

unexposed\_control\_Sum=d;

drop VNAME\_exposed\_case VNAME\_unexposed\_case VNAME\_exposed\_control VNAME\_unexposed\_control;

run;

data crude\_cc;

set crude\_cc;

label a="a"

b="b"

c="c"

d="d";

run;

We are now ready to conduct the crude analyses. We begin by creating variables for N1, N0, M0, M1, and T, the other variables we will need for our calculations. Then we calculate the odds ratio (OR), the variance of the OR, and the corresponding 95% CI. The OR is stored in the variable ‘OR’ and the variance of the OR is stored in ‘Variance\_OR\_confint’. We then calculate the lower (‘OR\_confint\_lower’) and upper (‘OR\_confint\_upper’) bounds of the 95% confidence interval and concatenate these to a single variable (‘OR\_confint’).

Additionally, we can conduct the test of no association by calculating the expected value under the null (‘expected\_null’) and the variance under the null (‘variance\_testassoc’). Using these, we can then get the Z2 statistic and corresponding p-value.

data crude\_cc;

set crude\_cc;

/\*Specify count related variables\*/

M1=a+b;

M0=c+d;

N1=a+c;

N0=b+d;

T=N1+N0;

/\*Calculate OR & 95% CI\*/

OR=(a\*d)/(b\*c);

Variance\_OR\_confint=(1/a)+(1/b)+(1/c)+(1/d);

OR\_confint\_lower=round(exp(log(OR)-1.96\*sqrt(Variance\_OR\_confint)),0.001);

OR\_confint\_upper=round(exp(log(OR)+1.96\*sqrt(Variance\_OR\_confint)),0.001);

OR\_confint=catx(',', OR\_confint\_lower, OR\_confint\_upper);

/\*Test of Association\*/

Expected\_null=(N1\*M1)/T;

Variance\_testassoc=(M1\*M0\*N1\*N0)/((T\*\*2)\*(T-1));

Z\_squared=((a-Expected\_null)\*\*2)/Variance\_testassoc;

p\_testassoc=1-probchi(Z\_squared, 1);

run;

With this data step complete, we can use the proc print to print the OR and corresponding confidence interval to our SAS output file.

proc print data=crude\_cc;

var OR OR\_confint;

run;

Which will look like:

A screenshot of a cell phone

Description generated with very high confidence

Similarly, we can print the test statistic and p-value for our test of no association:

proc print data=crude\_cc;

var Z\_squared p\_testassoc;

run;

Which will look like:

A screenshot of a cell phone

Description generated with very high confidence

*Stratified Analyses*

We can do stratified analyses in much the same way as we just did our crude analyses. In order to conduct stratified analyses, it is important to make sure to appropriate account for our stratification factor(s) before we get the counts of individuals in each combination of exposure and outcome. To do this, we first need to sort our data by the stratification factor(s) and then use a similar proc means step as above, but this time we must add the additional ‘by’ command to get sums within each level of our stratification factor. In the example below, I am stratifying by smoking status, which is represented by the variable ‘SMK’. Note that this step is using the same dataset, ‘dat’, that we created above & which contains the variables ‘exposed\_case’, ‘exposed\_control’, ‘unexposed\_case’, and ‘unexposed\_control’. In the proc means step, we create a new outputted datataset called ‘stratified\_cc’ that we will use for our stratified analyses after renaming and labelling our variables (this time with the suffix ‘\_i’ to represent that the counts are stratum-specific). As above, the only thing that you need to change for this code to run smoothly with different data is the name of your stratification variable, which should replace each instance of ‘SMK’.

proc sort data=dat;

by SMK;

run;

proc means data=dat sum maxdec=2;

ods output Summary=stratified\_cc;

var exposed\_case unexposed\_case exposed\_control unexposed\_control;

by SMK;

run;

data stratified\_cc;

set stratified\_cc;

rename exposed\_case\_Sum=a\_i

unexposed\_case\_Sum=b\_i

exposed\_control\_Sum=c\_i

unexposed\_control\_Sum=d\_i;

drop VNAME\_exposed\_case VNAME\_unexposed\_case VNAME\_exposed\_control VNAME\_unexposed\_control;

run;

data stratified\_cc;

set stratified\_cc;

label a\_i="a\_i"

b\_i="b\_i"

c\_i="c\_i"

d\_i="d\_i";

run;

The output of the above MEANS procedure will look like:

A screenshot of a cell phone

Description generated with very high confidence

Now that we have a cleaned-up dataset with 1 row per stratum, we can start to set up the calculations for our stratified analyses. As with the crude analyses above, we will do this by creating a series of new variables. Once we have each of these terms calculated, we can use a few simple proc sql commands below to sum over each of the strata to get our summary statistics and test statistics.

In this data step, we begin by creating variables for N1i, N0i, M1i, M0i, and Ti and then calculating the OR, variance, and 95% confidence interval for each stratum. We then set up the **future** calculation of our Maentel-Hanszel summary OR and 95% confidence interval by creating variables representing each stratum’s contribution to the MH weight (MH\_w\_i), variance numerator (MH\_variance\_numerator\_term\_i), and denominator (MH\_variance\_denom1\_i & MH\_variance\_denom2\_i). We will **later** use these values to get the MH OR and confidence interval.

Lastly, we set up variable that will be used in the calculation of the test of no association. These variables included the expected number of cases in each stratum under the null (expected\_value\_i) and each stratum’s contribution to the variance (variance\_term\_i).

data stratified\_cc;

set stratified\_cc;

/\*Specify count related variables\*/

M1\_i=a\_i+b\_i;

M0\_i=c\_i+d\_i;

N1\_i=a\_i+c\_i;

N0\_i=b\_i+d\_i;

T\_i=N1\_i+N0\_i;

/\*Calculate OR & 95% CI\*/

OR\_i=(a\_i\*d\_i)/(b\_i\*c\_i);

Variance\_OR\_confint\_i=(1/a\_i)+(1/b\_i)+(1/c\_i)+(1/d\_i);

OR\_confint\_lower\_i=round(exp(log(OR\_i)-1.96\*sqrt(Variance\_OR\_confint\_i)),0.001);

OR\_confint\_upper\_i=round(exp(log(OR\_i)+1.96\*sqrt(Variance\_OR\_confint\_i)),0.001);

OR\_confint\_i=catx(',', OR\_confint\_lower\_i, OR\_confint\_upper\_i);

/\*Summary: Calculate MH IRR & 95% CI\*/

MH\_w\_i=(b\_i\*c\_i)/T\_i;

MH\_numerator\_term\_i=(a\_i\*d\_i)/T\_i;

/\*Set up RGB Variance calculation\*/

RGB\_A\_i=(a\_i\*d\_i)/T\_i;

RGB\_B\_i=(b\_i\*c\_i)/T\_i;

RGB\_C\_i=(a\_i+d\_i)/T\_i;

RGB\_D\_i=(b\_i+c\_i)/T\_i;

drop OR\_confint\_lower\_i OR\_confint\_upper\_i;

/\*Test of No association\*/

expected\_value\_i=(N1\_i\*M1\_i)/T\_i;

variance\_term\_i=(M1\_i\*M0\_i\*N1\_i\*N0\_i)/((T\_i\*\*2)\*(T\_i-1));

run;

Before we can get the confidence interval for the MH summary OR, we need to calculate the RGB variance and store its value to be used in subsequent steps. **You should not need to edit this formula.**

/\*Calculate RGB Variance\*/

proc sql noprint;

select sum(RGB\_A\_i\*RGB\_C\_i)/(2\*sum(RGB\_A\_i)\*sum(RGB\_A\_i))+(sum((RGB\_A\_i\*RGB\_D\_i)+(RGB\_B\_i\*RGB\_C\_i))/(2\*(sum(RGB\_A\_i)\*sum(RGB\_B\_i))))+(sum(RGB\_B\_i\*RGB\_D\_i)/(2\*(sum(RGB\_B\_i)\*\*2))) into :RGB

from stratified\_cc;

quit;

Once the above calculation is complete, we can use our data set and a few simple ‘proc sql’ calls to calculate our summary statistic and test statistics. We start by calculating the MH summary OR and corresponding confidence interval. The first portion of the ‘select’ command calculates the MH OR as the weighted average of the stratum-specific ORs using the stratum-specific ORs and weights stored in our dataset. The next two lines of respectively calculate the lower and upper bounds of the 95% CI for the MH OR by summing over the variables we created to represent each stratum’s contribution to the MH OR and the RGB variance that we previously calculated. After running this chunk of code, the MH OR & 95% CI will be outputted to your SAS output window. **You should not need to edit any of these formulas.**

proc sql;

title 'MH OR';

select sum(MH\_numerator\_term\_i)/sum(MH\_w\_i) as MH\_OR,

exp(log(sum(MH\_numerator\_term\_i)/sum(MH\_w\_i))-1.96\*sqrt(&RGB)) as MH\_OR\_lower,

exp(log(sum(MH\_numerator\_term\_i)/sum(MH\_w\_i))+1.96\*sqrt(&RGB)) as MH\_OR\_upper

from stratified\_cc;

quit;

The output should look like:

A screenshot of a cell phone

Description generated with very high confidence

Similarly, we can use our stored variables to calculate the Z2 statistic for the test of association & corresponding p-value.

/\*Test of Association\*/

proc sql;

title 'Test of No Exposure-Disease Association';

select ((sum(a\_i)-sum(expected\_value\_i))\*\*2)/sum(variance\_term\_i) as Z\_squared,

1-probchi(((sum(a\_i)-sum(expected\_value\_i))\*\*2)/sum(variance\_term\_i),1) as p,

1 as DF

from stratified\_cc;

quit;

Which should look like:

A screenshot of a cell phone

Description generated with very high confidence

Lastly, we can use proc sql to store our summary statistics & use these to calculate the test of heterogeneity for the OR:

/\*Test of Heterogeneity: IRR\*/

/\*Store MH IRR for H statistic calculation\*/

proc sql noprint;

select sum(MH\_numerator\_term\_i)/sum(MH\_w\_i) into :MH\_OR

from stratified\_cc;

quit;

proc sql;

title 'Tests of Homogeneity: OR';

select sum(((log(OR\_i)-log(&MH\_OR))\*\*2)/Variance\_OR\_confint\_i) as H\_OR,

1-probchi(sum(((log(OR\_i)-log(&MH\_OR))\*\*2)/Variance\_OR\_confint\_i), (count(\*)-1)) as p,

count(\*)-1 as DF

from stratified\_cc;

quit;

Which will look like:

A screenshot of a cell phone

Description generated with very high confidence