SESSION 8: PRACTICAL

STRATIFICATION AND ANALYSIS OF 2X2 TABLES

Question 1.

To start with go back to the `bwmal` data we used before.

Generate new variables:

- **lbw**, for Inter Uterine Growth Retardation (IUGR) if bweight<2.5 kgs.
- **pargrp**, to categorise parity into nulliparous (parity=0) and parous (parity=1 or more).
- **matage2**, for mothers age 22 years or less (0), or 23 years or more (1)

We will restrict the analysis to babies born with at least 37 weeks gestation (gestwks>=37).

Show that both parity and mother's age are associated with IUGR (ie low birth weight in term babies). This suggests that, when considering the relationship between parity and IUGR, we should take into account their association with mother's age. That is, we take account of possible confounding.

We will do this by conducting a stratified analysis

Stratified analysis

Before you start, check the data.

which is the response variable?
which is the explanatory variable?
which is the confounding variable?

The first step is to consider the relationship between parity and IUGR, for each group (stratum) of mother's age.

To see the stratified tables we need to use the table command as follows:

```
table lbw pargrp matage2 if gestwks>=37
```

You can also use the following command:

```
bysort matage2 : tab pargrp lbw if gestwks>=37, col chi
```

To obtain the statistical analysis we need to use the cs command, by adding a third variable as follows:

```
cs lbw pargrp if gestwks>=37, by(matage2)
```

Note that the first two variables named, *pargrp* and *lbw*, are the ones which form the table - the rows and columns - for each value of the by variable, *matage2*. The summary statistics which we saw before are now produced for **each** table.

For stratum 1 (*matage2*=0),

what is the value of the risk ratio? ____

and for stratum 2 (*matage2*=1)? ____

The output also a summary of the results from the two separate strata. What are the values of the following:

the Crude Risk Ratio? ____
the Summary (Mantel-Haenszel) Risk Ratio? ____

Statistics
Let us first look at the results on risk ratios (RR's). Note that the crude RR above is the same as that from the original table of parity and low birth weight. How does this crude RR compare with the RR's from each strata of *matage2*? What do you conclude about the confounding effect of mother's age? The summary RR can be considered as the risk ratio of low birth weight, adjusted for the confounding effect of mother's age.

The final part of the output tests whether the risk ratios of low birth weight vary between the two age groups (ie if there is an interaction). In this example, there is no evidence of a significant difference between the two risk ratios. If there *were* a significant difference in RR's, it would not be sensible to present the summary RR. Instead, the RR's for each maternal age group should be reported separately.

It should be emphasised that statistical packages will often not directly give you the best table to present your data. For instance, to tabulate low birth weight by parity and mother's age, we might set out a table as follows:

		Birth weight		Total
		Low	Normal	
Mother aged 22 Years or less	parity 0	__(__%)	__(__%)	__
	parity 1+	__(__%)	__(__%)	__
Mother aged 23 Years or more	parity 0	__(__%)	__(__%)	__
	parity 1+	__(__%)	__(__%)	__

Action
Use the output from the previous table command to derive the numbers for this table. Either work out the percentages with a calculator or use the tabulate command.

Final notes
We have seen how to investigate the association between two binary variables while controlling for a confounding factor. We did this by conducting stratified analyses of the original data file. We will now look at another method for conducting stratified analyses using data that are already tabulated, so that we do not need the original data file. This method uses the *Statcalc* module of Epi Info. (There is a command in STATA — csi — which will do this, but only for a single table, not a stratified analysis. If you really want to be able to do this kind of analysis in STATA, see the optional material at the end of the session.)

Question 2. Use of *Stata* to Conduct Stratified Analyses for Pre-Tabulated Data

To illustrate this method, we will use the following data from two different villages (A and B) on the use of bed-nets and presence of an enlarged spleen. In this example, the presence of an enlarged spleen may be regarded as an indicator of malaria. The data suggest that those persons who use bed-nets are less likely to have an enlarged spleen (27/91 = 30%) than those persons who do not use bed-nets (46/87 = 53%). In other words, bed-nets seem to protect against having an enlarged spleen (crude RR = 30% / 53% = 0.57).

Both Villages Combined

	Spleen Enlarged		Total
	Yes	No	
With nets	27 (30%)	64	91
Without nets	46 (53%)	41	87
Total	73 (41%)	105	178

When the effect of bed-net use on having an enlarged spleen is looked at separately for the two villages, a different picture emerges. In village A, those persons who use bed-nets are less likely to have an enlarged spleen (12/24 = 50%) than those persons who do not use bed-nets (42/71 = 59%), but the crude risk ratio is much closer to one (crude RR = 50% /59% = 0.85), suggesting only moderate protection of bed-nets. In village B, those persons who use bed-nets are less likely to have an enlarged spleen (15/67 = 22%) than those persons who do not use bed-nets (4/16 = 25%), but again the crude risk ratio is much closer to one (crude RR = 22% / 25% = 0.88), again suggesting only moderate protection of bed-nets.

	Village A			**Village B**		
	Spleen Enlarged		Total	Spleen Enlarged		Total
	Yes	No		Yes	No	
With nets	12 (50%)	12	24	15 (22%)	52	67
Without nets	42 (59%)	29	71	4 (25%)	12	16
Total	54 (57%)	41	95	19 (23%)	64	83

A stratified analysis is necessary since village is a confounding factor: it is related to both the response variable i.e. enlarged spleen (57% of village A and only 23% of village B have an enlarged spleen); and it is related to the explanatory variable i.e. bed-net use (24/95=25% of village A and 67/83=81% of village B use nets).

To conduct this analysis we need to enter the data into Stata.

We can do this in 2 ways:

By entering the data using an editor.

```
edit
```

After entering the data (see below), then you will have to change the names of the variables

```
rename var1 Village
```

By using the input command and entering the data at the keyboard (this can be used in a do file)

```
input Village Spleen Nets Count , automatic label
0 1 1 12
0 1 0 42
0 0 1 12
0 0 0 29
1 1 1 15
1 1 0 4
1 0 1 52
1 0 0 12
end
```

You can then define the labels for these values, and apply them to the variables

```
** Define labels
label define village 0 "A" 1 "B"
label define nets 0 "No nets" 1 "Bednet"
label define spleen 0 "Normal" 1 "Enlarged"

** And apply the labels to the variables
label val Village village
label val Nets nets
label val Spleen spleen
```

Note we only enter the cell totals in the variable Count. We use the variable Count to tell us the frequency of each observation. Get the overall tabulation of nets by spleen using the tab command, with the instruction to use the numbers in Count variable as frequency weights for each cell.

```
tab Nets Spleen [fweight=Count] , row
```

Get the same data for each village by using the command with if Village==0, and if Village==1

Now get the estimates of the Risk Ratio, and Odds ratio, by using the cs command.

```
cs Nets Spleen [fw=Count]
```

and repeat for each Village separately by using ‘if Village==0’ and ‘if Village==1’. To obtain the stratified estimate of the risk ratio and odds ratio, we need to include Village as strata. We do this using the by(village) option.

```
cs Nets Spleen [fw=Count] , by(Village)
```

What are the values of

the crude RR? ____

the RR in Village 0? ____
the RR in Village 1? ____

the M-H summary RR? ____
the M-H summary chi-square? ____

What are your conclusions about the relationship between bed-net use and presence of an enlarged spleen?

What was the effect of controlling for the confounding variable, village?

Question 3 Oral contraceptives and risk of myocardial infarction

Data were collected from a case-control study on the use of oral contraceptives (OC) by women with a myocardial infarction (MI, heart attack cases) and controls.

		MI	Control
OC use	Yes	29	135
	No	205	1,607

OR = 1.68

a. Use the Stata cci command to obtain the odds ratio of the association.

b. Enter the data into Stata and obtain the same result

c. Obtain the same estimate from StatCalc. Do they agree?

d. Use Stata to stratify the data by age (using the same technique as for Question 2).

Stratified by age: (MI= cases, Ctl=controls)

		Age									
		25-29		30-34		35-39		40-44		45-49	
		MI	Ctl	MI	Ctl	MI	Ctl	MI	Ctl	MI	Ctl
OC	+	4	62	9	33	4	26	6	9	6	5
	-	2	244	12	390	33	330	65	362	93	301
OR		7.9		8.9		1.5		3.7		3.9	

a) Is the association between oral contraceptive use and MI confounded by age?

b) Is the association between oral contraceptive use and MI modified by age?

Summary - statistics introduced in this session

Stratified analyses for 2x2 tables are used to investigate the association between two binary variables while controlling for the presence of one or more confounding factors. The Mantel-Haenszel risk ratio measures the association between the two binary variables controlled for the confounding factor(s).

The final test provided in a stratified analysis in STATA looks at whether the Risk Ratio in each stratum (i.e. for each value of the confounding factor(s)) vary. If there *were* a significant difference in RR's, it would not be sensible to present the summary RR. Instead, the RR's for each maternal age group should be reported separately.

OPTIONAL EXERCISE

An alternative way to make STATA analyse an aggregate (i.e., pre-tabulated) dataset.

We can do this by making a dataset with the relevant number of records. For the above stratified analysis of spleen ‘rates’ by village, try the following. (You may be able to find a more efficient way!) The commands prefixed by ‘*’ are just comments and don’t have to be typed.

```
* clear existing data
clear

* tell STATA how many records the new dataset will have,
* ie the number of people in the study
set obs 178

* we need variables of length 178 to contain the data on
* spleen, net & village

* the following command is a quick way to make a variable
* with values 1, 2, 3, ... 178
range net 1 178 178

* recode so that the first 91 people used a net
recode net 1/91=1 92/178=0


* make a new variable for spleen
range spleen 1 178 178

* out of the 91 net users, 27 had enlarged spleen, etc
recode spleen 1/27=1 28/91=0 92/137=1 138/178=0

range village 1 178 178

* NB the following command should all be on one line
recode village 1/12=1 13/27=0 28/39=1 40/91=0 92/133=1 134/137=0 138/166=1
        167/178=0

* we are now ready to do the analysis.
* you might like to ‘browse’ first
* the results from this be the same as from STATCALC
tabulate spleen net

table net spleen village

cs spleen net, by(village)
```