MCH EPI Conference

Time Trend Analysis for MCH Outcomes in Large and Small Populations

December 8 – 11, 2008

JEFFREY SHAFFER: Okay. So we're ready to begin today's training. Like Henry said in the other room, my name is Jeffrey Shaffer. And this is my colleague Dr. Francis Mather. And we will be teaching the class, and we're going to be working with time trend analysis for maternal and child health outcomes.

Okay. An overview of time trend analysis. What we're interested in doing is identifying changes for some indicator over time. And when I say that, although we'll be talking about time trend analyses for maternal and child health outcomes, the analyses, the statistical techniques and descriptive techniques that we'll talk about today and tomorrow also apply to many other types of outcomes as well.

We'll talk about this time trend analysis in the context of maternal and child health outcomes, in particular, infant mortality. Some potential indicators that you might be interested in analyzing time trends for include for maternal and child health, prenatal care, initiated in the first trimester. Infant mortality rates, you might be interested gestational age, and seeing how gestational age may affect infant mortality rates over time.

So it's one thing to just describe the time trend, it's another thing to figure out why the time trend is there.

Okay. And like I said before, for this course, although the techniques are in the context of maternal and child health outcomes you could also apply these to other outcomes as well for environmental health, one may apply these techniques to perhaps ozone or something like that, measuring ozone, particulate matter and over time.

It's often of interest to monitor these things over time to detect changes so that you can try to control for these things and establish policy and things like that that might diminish the particular matter.

Okay. But like I said before, we're going to be talking about infant mortality. Okay. Now, the changes that time trend analysis is able to describe, some of the types of changes you might want to consider are increasing and decreasing trends, time trends, okay.

Abrupt changes. Okay. Those are things that you want to look at, too, in a time trend analysis. Abrupt changes are often called spikes. That's an abrupt change.

And you also have oscillating changes. Some type of seasonal change or something like this with an oscillating pattern over time. And these are all things that we want to consider over time.

Sometimes, when there's no trend at all, that says something in itself, too. So when there's no significant trend, as we're going to see for some of the techniques in here, when there's no significant trend, that says something, too.

Time can be considered and time in the sense that you guys typically report these types of things in months, days, years, weeks, et cetera. That's what we're talking about there.

Time trend analyses, you can have visual approaches. That's reported a lot in the literature. And you can also have statistical approaches, which we're really going to work a lot with in this course.

By visual approaches, I mean scatter plots, time trend plots, things like this. Statistical approaches could be in the form of purely prescriptive such as means, rates, things like this. Or they could be inferential, such as some type of hypothesis test or confidence interval.

Some analytical issues, the data quality, particularly for these types of things. Maternal and child health, the data reporting isn't always accurate. So these are things we always want to consider.

The distribution of the indicators. We're going to talk a lot about this as we go through the course. And the appropriateness of data for inferential statistical techniques.

Okay. It may not be appropriate to use inferential statistical techniques. For instance, most inferential statistical techniques require a random sample. So if we don't randomly sample, then we may not be able to use inferential statistical techniques.

Okay. Targeted audiences for time trend analysis can include the general public, the media, policymakers and researchers. And it's important to note that each of these four audiences are interrelated.

The influences are interrelated. For instance, the policymakers may influence the media, or vice versa. The media, probably the media is going to influence the policymakers.

If the media reports something quite a bit, okay, the policymakers might want to make a change. Researchers, researchers might call attention to something to the media. And then the media can influence policy.

So these are the types of target audiences that you're typically dealing with. Okay, now a little bit about the course structure. Now, that's just an overview of

trend analysis. Now, for this course, you guys have a copy of the manual, and in that manual, the workshop manual, I think the first part of that is going to be the agenda.

Okay. We're going to try to stick to that agenda as closely as we can. Now, in your book I've also provided you with the manual for the course. And that is going to be starting, it looks to be on page -- that's page 1. They have it as page 1 in your workshop book. So this is the main manual for the course.

Now, what this presentation that you guys see on the screen is basically a summary of that. So when you guys are going back and -- and you guys can have a copy of this presentation, too. I can make that available on the desk tops, if you like.

All the materials in this case will be available through AMCHP's website at the conclusion of this workshop. So you guys will have digital copies at your disposal, too. But the manual is here and we're going to be doing a lot of exercises out of this manual.

Now, we have a series of data sets. The data sets are posted on the desktops of each of these computers. So we're going to take a look at those in a little bit. Presentation, and at the end of the course there will be an evaluation.

And if we have any revisions to these materials, like I said, all these materials will be posted on AMCHP's website at the end of the workshop. After the workshop's finished, and I've given you guys the URL here if you want to visit that URL.

You can see last year we also taught this class so all the materials are up there right now. Okay. The course structure, okay, I'll begin with a brief lecture of the material given in the workshop manual.

And then we will do a set of in-class exercises using the SAS software. What I've done here is I've given you a code, a SAS code to do the things that I'm going to ask you to do.

And that's given in the workshop manual. And there's one exception to that because I've added a few exercises and I'll talk about that when we get to it. But most of the exercises that you guys see, the SAS code for these exercises is going to be given in the manual.

Okay. Like I said before, the course data sets are loaded on the desktop for each computer in a folder called Student Data Sets. So we'll look at those in a little bit.

So the contents for this workshop. Okay. Well, this is the introduction. Next we're going to talk about the data that we looked at.

Now, like I said, you guys can apply this to other types of data as well. We'll talk about the data sets we use to perform these analyses. And then we'll talk about the problem. That will be Chapter 3. And in Chapter 4 we're going to look at some exploratory trend analysis techniques.

And that's typically how you begin a thorough analysis is with a real strong bivariate analysis of all of the factors of interest. And we will look at that in Chapter 4.

Chapter 5 we'll get into some traditional regression models and see how those apply to infant mortality.

Chapter 6, we'll talk about join point regression, which is essentially an extension to the regression models we'll look at in Chapter 5. Chapter 7 we'll talk about Poisson regression. Poisson regression, it's very appropriate for count data. The Poisson distribution is perhaps the most important distribution for this type of count data that we'll look at today.

Chapter 8 we're going to talk about some spatial considerations. For the most part we're going to be dealing with time trends. But there are also other types of trends that exist, spatial trends are also there. Spatiotemporal trends, things differing over space and time. By spatial differences, I mean things that differ by region.

We'll be talking about that in Chapter 8 so that's tomorrow. We're probably going to cut it off today after the join point regression. Chapter 9 we'll talk about some more advanced time series techniques. We'll talk about some AREMA models tomorrow.

The time series techniques are a little more involved, but they essentially account properly for the time dependence as we're going to be seeing.

Chapter 10, we'll talk about some survival analysis techniques, some time to event stuff. And I've given you guys a flow chart of the analytical techniques we used here. We'll probably see Chapter 11 a little bit today.

So that's how the course is set up. And we're going to go ahead and jump right in. Any questions before I start, about the structure?

Okay. Trends. Some uses for trend analysis. Public health policies. Expenditures. You can justify expenditures by looking at time trends. Prices. Okay. Prices of many different things can be influenced by trends.

Any type of trend can influence price of just about anything. Treatment programs can be influenced by trends. Research priorities can also be influenced by

trends. Trends can be used to, as surrogate measures of community health or economic development.

Okay, for instance, one might say decreasing IMR, decreasing infant mortality rate, is saying that your state is doing better. Okay. That's an example of what you see here. It can be used as a measure of community health.

So a trend in infant mortality rates, looking at that over time and seeing that decline, is saying something good about your state. So that can certainly be used as a surrogate measure.

Some types of trends. I briefly mentioned this earlier. We'll be looking for the most part at temporal trends, time trends. We'll also be looking at a little bit like I said, tomorrow, at spatial trends. We'll do a little bit of what we call disease mapping.

Disease mapping is mapping some health event, some health outcome. And we'll also be looking a little bit at spatiotemporal trends, meaning trends over time and space.

Okay. Some classifications for trends. Immediate trends. You can have immediate trends. All kinds of different trends. Some classifications for trends, first one, immediate. This can indicate perhaps an underlying epidemic. So once

you determine that there is a trend, the next question is why is that there. Immediate trends are often indicative of underlying epidemics.

Short-term trends may indicate seasonal effects. If the trend is there and then it goes away and then it comes back, that may be indicative of seasonal effects. Trends in influenza or something like that, are, perhaps, seasonal.

Long-term trends. Long-term trends can be indicative of things like effective treatment plans, if you see a monotonically decreasing trend over time, that can be indicative of an effective treatment program.

Or educational campaign or technological advances or vice versa, if it's going the other way. If the trend is increasing in a bad way.

Okay. Trend analysis is formally defined as the analysis of change in some outcome over time, space or other quantitative factor.

And it may be used to establish causation. Okay? So you can certainly use it to do that. Exploratory techniques involve investigating bivariate relationships without controlling for other factors.

And this is typically where you start. Like I said, a thorough statistical analysis is with exploratory techniques and a good bivariate analysis should always be at the beginning of your final analysis.

Some exploratory techniques include descriptive tabulations, displays, scatter plots, they can be very simple displays. These are usually good things to do at the beginning. Hypothesis test, two sample T tests things like that, those are still important.

Ordinal measures of association. Okay. We're going to talk a little bit about this later. What do you think of association, probably the first thing you probably think of is like an R square or something like this.

We're going to look at some methods of looking at association for infant mortality data. And annual percent changes. We'll be talking a lot about annual percent changes because they have a lot of appeal for MCH data. Data often reported in this manner.

Regression techniques. Okay. Now when we get into regression techniques, we're usually interested in doing a little bit more than what we did using the exploratory techniques.

The difference between the exploratory techniques and the regression techniques, is with the regression techniques, usually you have a predictive model. We're going to see one case when you don't. But for the most part you mostly have a predictive model, so you're able to predict.

Further, you're able to easily control for confounders or determine whether a variable is a confounder or is related to the outcome.

Some of the regression techniques we'll look at in this class include linear regression. Now, linear regression is perhaps the simplest type of regression there is. But the trend must be linear in nature.

Okay. To use linear regression, you have to be looking at a linear trend. We'll also be looking at non-linear regression, in particular quadratic regression. We'll also be looking at low S smoothing, which is a nonperimetric regression model that requires, doesn't require as many assumptions as the previous two.

Then we'll spend a lot of time on join point regression. Join point regression is essentially modeling multiple lines to data. So linear regression says you want to fit one line to a set of data.

Join point regression says you want to simultaneously fit many lines to a set of data. So we'll spend a lot of time on join point regression. It's very useful for MCH data.

Then we'll talk about Poisson regression. Poisson regression is appropriate for when you have an outcome that is rare and an outcome that is discrete.

So a lot of MCH data does involve a rare discrete outcome. We'll also be looking at spatial regression and this should only be used if spatial dependence exists.

So thinking about spatial regression, it's another type of regression there is. The question is do you need it? Do you need to use it? It's a more complex model.

We're going to see how to determine if spatial dependence exists and whether a spatial model is appropriate. And we'll also look at some cluster analysis briefly. But it's very useful for detecting regions or small areas of similar or dissimilar outcomes.

So if you were interested in identifying an epidemic in a particular area, okay, clustering would be useful, because that depends on the spatial nature of the data.

If you're not considering the spatial aspect of the data you can say maybe there is a trend or there's something there, there is an epidemic, but you can't say where.

So we'll be looking at some cluster analysis as well. We'll also be looking at time series analysis. Now, time series is like I said a little more complex, a formal time series analysis is a little more complex.

But it's very powerful when the objective is to forecast new events. Okay. If you're interested in talking about something that's going to happen 10 years from now, the best of any of these techniques is going to be right here. Time series analysis, using a formal time series model.

Now, you may not always have to do that. These other regression techniques are very powerful when you're not interested in forecasting beyond the range of the data. When you're interested in seeing something beyond the range of data, something in the future, predicting a future event, this one's very appropriate.

And, once again, we want to focus on which one of these is appropriate. So the idea isn't just to go through and say okay I have 10 different techniques, I can use any of them.

We're going to talk about when to use which one. And we'll also talk a little bit about survival analysis. Now, this is a different type of question. We're going to be analyzing time to event, time to some event.

So this is a little bit different. It's a different question, let's put it this way for now. We're going to be talking about a different type of question for survival analysis stuff.

Okay. Now, let's get into a little bit of the details here. Like I said, we're going to be for the most part we'll be looking at infant mortality.

And infant mortality, infant mortality is usually expressed in the form of rates. Now, a rate is formally defined as the event frequency for a specific time divided by the population at risk for the same specific time times some multiplier. I just called that multiplier K. This may be per thousand persons or per 100,000 persons or something like that, that's what K would be, the thousand or 100,000.

Now, this differs from a ratio in that those observations in the numerator are also included in the denominator. So that's what makes this a rate.

And it's very common to express trend data in the form of rates over time. For instance, prevalence incidence and, like I said, for the most part we're going to be talking about mortality.

Okay. Now, some things to notice here about this definition of a rate. We essentially have two major components here. There's a numerator and the denominator. Now, there are different ways to model rates, and we're going to look at these different techniques.

One way to model a rate is to consider the rate as continuous. And that means consider this whole thing as one component. But we actually have more than that. We have a numerator here, a denominator here, and a multiplier here.

Now, if we were to consider that as one piece, just as one component, do you guys see the ramifications for doing that? Well, let's say I have a rate defined just like this of, a rate of five, I don't know, a rate of five for something. Let's say this number's five. And let's say the event count for that, let's disregard the multiplier for now. Let's say the event count for that was based on a population, an associated population that was very small.

Now, let's say that I have another rate of five, but this population is very big. Now, if you compare those two rates one is going to be very reliable and one is not. The one based on a very small population or very small denominator is not going to be very reliable.

But the one based on the very large population is going to be much more reliable. If you have a population of 10. Let's say you have a population of 10, and you have a rate and you calculate a rate based on that population of 10, you're probably not going to be too confident in the results.

But if you have another rate that's calculated based on a population of a thousand, you're probably be very confident in the results. So I'm saying that it's best to consider these separately.

Make sure you make use of both the numerator and denominator in your modeling. Now, rates are modeled as continuous all the time. So if they're done like that, the one thing you lose is the ability to say anything about the population of interest. Okay? In other words, you're not able to put -- to use a risk type of interpretation for your results as we're going to see later. So just think about this as we go along.

Remember that formula there. Okay. Rate precision is directly affected by the size of the risk population, the denominator. And it's often unstable at the local level. For instance, rates are often unstable at the census track level the zip code level and things like that.

So typically what we do is we like to aggregate the results. We may like to report a rate at the state level. Well, sometimes that's okay. But for the most part the

local officials want local results. Okay. So this raises the question how small can this be before it's too small and unreliable.

So we're going to look at that as we go along, too. But like I said we're always going to consider that denominator the whole way through. And I say here rates are desirable for local public health officials. So aggregation may not be appealing. They don't want to hear what's going on in the entire state. They want to know what's going on in their zip code or their census track, something like that.

Okay. Motivation for this analysis, everything we're going to do in this course. We will explore the aforementioned issues for infant mortality rates in the state of Louisiana.

In particular, is there an observed time or spatial trend in the IMRs, infant mortality rates, in Louisiana? What are some of the data reporting issues? What are some of the IMR risk factors? What influences IMRs?

If we see a trend, then why is it there? And can observed IMR trends be explained with a set of predictors? This is really what you want to do. Okay. The objective here is to come up with a set of predictors -- now, it isn't easy to do this. But the objective and the ultimate motivation is to come up with a set of predictors that completely explains the IMR trends.

Once you explain everything the problem's solved. So if you can identify what causes these trends, then that's what you want to do. That's the objective of this problem, is to identify a set of predictors that explains the trends and removes all trends.

If you come up with a set of predictors and you essentially have no trend left, you've explained the trend. So that's going to be our motivation. Although this is a very complex problem to completely explain the trend in IMRs, I don't know that it's ever even been done, but that's always the motivation, is to come up with new explanations for why the IMRs could be changing over time.

And also how can observed IMR trends be effectively communicated? Okay? Perhaps to public health officials. Perhaps to your boss. Perhaps for some type of publication or reporting or something like that.

Okay. I guess are we going to go right into Chapter 2? We're going to go right into Chapter 2 and talk about the data. These are the data we looked at. Now, it may be similar for your state, for infant mortality.

But even though we're talking about infant mortality once again, keep in mind these techniques are going to apply to any type of health outcome, any type of

mortality data, these could easily be applied to and many other types of data as well.

Okay. The data that we have were abstracted from the Louisiana vital records department. And we have all live births for all live birth data for Louisiana residents from 1990 to 2005.

We also have death certificate data for Louisiana infants dying prior to one year of age. And these are the three data sets that you guys see here that we're going to be working with quite a bit.

The first one is the birth data set. That's the birth data that I mentioned up here. Notice the time period, 1990 to 2005. So what we're interested in doing is seeing what happens in terms of time and a little bit in terms of space later from 1990 to 2005 with respect to infant mortality.

So the births data says is here. That's its name, births data set. And this is N equals, that's the number of observations we have in this data set.

So what this is saying is I have over a million live births in the Louisiana for this entire time period. N is over a million here. For the deaths data set, we have a little over 10,000 for the death data. Now, obviously many -- most of these infants

aren't dying, so that's why you guys see such a small number here. Now, you have information both in the births data set and the deaths data set.

The idea is to link these together. And that's often done for many, many reports that you guys see, NVSS has the infant mortality link data out there. There are a lot of different reports that go through and link these two data sets.

I'm going to show you guys a way today of linking these two data sets and some of the factors to consider, but keep in mind there are probably other ways for linking these data sets that we'll look at today.

Now, births/deaths is going to be the link file. That's the file that's going to allow us to link births to deaths. For instance, if you have information on the birth certificate, and you have information on the death certificate, the idea is to link those together, but you need an ID here, a child ID here and a child ID here. And then the idea is to link those together and talk about some of the things that are reported in births and deaths and see which of those factors could be influencing infant mortality. That's what you have at your disposal.

Now, this isn't to say that you couldn't create other variables as well. They may not be -- you wouldn't be able to link them to the children, but you may be able to link these things to a certain area or region or something like that and report it in some type of aggregate form.

That could be done as well. But the first step is to link these two data sets so that you can see what's causing the deaths. And this linked data set is going to allow us to do that. Now I'll talk a little bit about the structure in the linked file in a little bit. For the linked file we had a little over 10,000 observations as well.

Now, notice here, this number's 10,245. This one's 10,609. If this data were perfect, I think this one, 10,609 would be the same as this. It's not.

So we have some data reporting as well. This is one of the problems you have with linking. That's going to be called an unmatched observation, and we're going to talk a little more about that in this chapter.

So instantly when you see this number doesn't match this number, there are data reporting issues. Because you have some deaths that you're not going to be able to link. Some of these deaths are not reported in the linked file. So you're not going to be able to link them to births.

And we'll talk a little more about why that is in a second. Here are some of -- here's the coding guide for births, the births data set and here are some of the variables in this data set. You have the birth ID. Okay. Birth identification number. And you also have the birth year. The gender of the child. These are the variable names you guys see here. Typically I type my variable names in capital letters so

when you see in the manual that something's capitalized that means it's either going to be a data set name or a variable name.

Multiple. The number of children born. Perhaps a woman had twins, then multiple would be two. Mother's race. It was given like this. Given in terms of a lot of different races. Now, what you're typically going to want to do is categorize race into say black, white and other. And for the other categories there aren't many of them. So you're not going to be able to say much about them.

If you were interested in talking about infant mortality rates for Vietnamese, that would be difficult, because there aren't going to be that many of them, and if we go back and think about that denominator and that rate, that's going to be very small. So we probably need to aggregate over a pretty large region to do that. Although we could do it if we had enough observations.

Okay. Mother's age. Okay. Obviously this is going to be an influential thing when it comes to infant mortality. Once the mother reaches a certain age, there's a lot higher risk of infant mortality. So this is something that we may want to control for it. We may want to say something about it. But we at least want to account for it.

Married. Mother's marital status is there. The weight of the child. That's going to be important, too. That's certainly very much related to infant mortality, the weight

of the child. And we're going to talk about some cut points for that in a little bit. Meaning where do you break it up. What's risky?

And like I said these are my variable names. The number of prenatal visits. That can be influential as well.

The mother's education level is there. The month prenatal care began. Gestational age. And that's reported in weeks. Residence, zip code, this will be important for spatial stuff. If we want to consider this spatial distribution of the infant mortality rates, this is where we're going to get that. And we also have the resident's county.

So if we're interested in spatial stuff this is what we have. Now, this is what we were presented with. Okay. For the spatial stuff we can instantly see we can go no longer than the zip code level, at least from what I've shown you on the screen.

If you had census track level reported here, then you could say something about it. But we know right now instantly zip code is as low as we go.

Here's the coding guide for the deaths data set. So we're given essentially two main data sets. We're also given a linked data set, but we're interested in these

covariates, all these potential predictors, and these ones associated with the deaths.

Now, here's the coding guide for deaths. We were given a death identification number. The infant's gender. Now, notice infant's gender is also here. Infant's race. The unit of infant's age at death. It could be months, days, hours, minutes and so on.

The actual infant's age at death. I think what I -- measure of units in age code. Oh. One is the unit. So the way this was reported is this variable told you whether it was a month, day, hour, minute and so on. And this one told you the actual measurement.

So let's say this number here for age unit was 2. Okay. You would need to know if that was two days or two months or so on. So you'd have to use these two in conjunction with one another to figure out the actual age.

You'd use that in some type of calculation or something or create a new variable would probably be the first thing to do to get everything in weeks, which I think we do a little bit later and we'll talk about tomorrow.

The residence of the infant, I mean we had the residence of the mother. So we would think for the most part those are going to be the same. We have the residence of the mother given here, the location of the mother.

So that might not be very important, residence of infant. And similarly for the residence, the zip code for the infant. Probably not very important for us. I'd probably prefer to use the one from the birth certificate, from the mother.

The year of death, okay. This will be important for us. We'll talk a little bit about that later and the period of death occurrence, okay, number one, neonatal. Number two, post neonatal.

Okay. Now those are the two data sets we have. Now, they're separate data sets. Okay. We were also given a coding guide. Here's the coding guide for births deaths. This is the data set we're going to use to link this data set to this data set.

And this is how we're going to calculate things like infant mortality rates. And well not necessarily -- there are different types of infant mortality rates that we're going to look at. For cross sectional infant mortality rates I'll show you guys how to calculate those that involves using these data sets independently.

But if I want to talk about, let's say, well, any of these, really, if I want to talk about the year of death and some type of cohort fashion, then I'm going to have

to link these data sets, because that involves variables from this data set and this data set based on the same infant.

Okay. So if I want to look at this from a cohort perspective, which we'll discuss a little more, too, then I'm going to need to link these data sets. Now, to do that I'm going to do need something that allows me to link these data sets and this is the data set called births/deaths, and this is the Louisiana MCH linked birth and death data. So this was given to us as well.

Now, here are the variables in it. Okay, only four variables. Birth ID. Birth year. Death ID and death year. See, that will allow us to go from here, birth ID here, death ID here, from here to here and link these two together.

See, death ID is not going to be the same number as it is for birth ID. You're going to get a different death identification number. So there's no way to go from here to here directly. This allows us -- see, here we have birth ID and death ID in the same data set.

So this is going to allow us to link the two files. Now, one thing I can say when you're linking files together like this, when you're linking birth and death data, you're not only going to want to join these, link these based on the birth ID and death ID, it's very important to consider birth ID, birth year combinations,

because most state reporting is going to involve assigning the same birth ID for different years.

For instance, in 1990 you may have the first child born in 1990 may be given a 1, may be assigned a birth ID of 1. Well, in 1991 that first child born in 1991 may be given a birth ID of 1. They're not the same infant. The way around that is to consider not just that, but consider the birth ID/birth year combination.

So if I say, yes, child 1 was born in 1990, but if I consider the 1 and the 1990 together, then that number is unique. Then that combination is unique. So these are things that you want to consider when you're interested in joining data sets together.

Okay. We're going to go through a little exercise in joining these together. Like I said, the technique we use, it certainly could vary from state to state. It's probably pretty similar no matter where you go as far as linking these together, but there are some issues that you need to consider when you do it.

Okay. Some variable indicators for births. So one of the first things we did here for this, for the births data set, if you back up the births, you can see a lot of these variables can be categorized. And you want to -- that's one of the first things you want to do is categorize these into meaningful things that you can work with.

One of the first things that I see is race. I'd instantly want to categorize that as white, black and other, or something like that. Probably mother's age as well. You'd want to think about how to categorize that. Should you categorize it. Well, you can model -- when you're interested in modeling later on, you can certainly model mother's age as continuous without categorizing that or breaking that into different pieces. But it's usually more appealing to talk about it in terms of categorical predictor, to say something about all women, let's say, over 38 years of age or something like that, as opposed to given a particular year, you know, it's a little harder to talk about each individual year for a woman.

And we also want to consider risk, too. Women may be under 35 years of age aren't at risk at all. So they may not -- that may not be appealing for this study.

Okay. So here are the indicators that we came up with. Okay. We categorized multiple or mult cat. The variable is multiple. We categorized multiple into mult cat. Meaning mult cat is a categorical analog for multiple, its continuous form.

We started with this variable. And that was given in the births data set. That was given right here. Multiple. So we were given this, number of children born, and we categorized that into singleton, one child per multiple, meaning you have more than one.

We also categorized the race of the mother into white, black and other. We categorized the age of the mother into these two categories, less than 38 and greater than or equal to 38 years. We categorized the weight of the child. Very important predictor for infant mortality into these three categories.

Well, these are open to debate. Some of these predictors you guys may want to include on a finer scale, meaning you may want more categories than what we have up here. Mother's education, we categorized that into these two categories, high school or beyond and less than high school, less than the high school diploma.

Mon prenatal, we categorized as prenatal. And we said one to three months or none or greater than or equal to four months. What this means is one to three months means they received prenatal care between one to three months after conception.

And the second one says that they received either no prenatal care, zero, or they started their prenatal care at least four months after conception.

Okay. So this is the one you see here is going to be the risk group. Gestational age, we categorized as less than 37 weeks and greater than or equal to 37 weeks.

Okay. Now what I want to do is we're going to go ahead and go through our first set of exercises. Now, these exercises are not going to be in the manual. I added these exercises recently just to get, so you guys can get acclimated with the data sets and everything on the computers.

I've given you these exercises in the presentation. So that's where I'm getting these from. You won't see them in the manual. Now, what I want to do is I just want to print these -- we'll print them to the screen and run [Inaudible] contents. I want you to look at some of the descriptive information for each of these data sets. That will allow us to get acclimated to the data sets on the computer and you guys will get an understanding of how we're going to complete the exercises as we go through the course.

Exercise 2.1 says, I'll do the first one up here. You can follow along on your PCs, some of you guys are paired up, too, that's fine, you guys.

If you're not -- if you need any help, just raise your hand and Dr. Mather will come around and help you. And I'm going to mention a few things before we start about how I write my SAS programs. What you guys see here are my, some of my global statements I like to indicate at the top of the program. Of course, a title, optional, of course. This one here, you guys may not have seen that, but that's a nice little statement to have in your SAS programs.

And I'm assuming everyone here has at least a basic understanding of SAS. This is called the DM option, and it's useful for clearing your output and log files simultaneously after each submission. So if you guys haven't seen that before, I would include it at the top of every SAS program you write. It's very useful.

Even if you don't know completely what it does now, put that at the top of your program. It's a very important option. These are the options I like to indicate my global options. I like to set the line space to 80. Suppress the date and the number from the output. Okay. And this is the one that's going to be important for us. The lib name statement. This is where we're going to indicate where the SAS data sets are.

So since I did this back in New Orleans, what we're going to do is we're probably going to have to change -- this is called the lib ref. This is the directory in where the SAS data sets lie. We're going to change this directory based on the directory on your desktop. And it's probably going to be different from the one I have up here.

Okay. And after that, okay, so we'll do this for each exercise. You guys can copy and paste from exercise to exercise, if you want. We'll need to write those statements essentially once for all the exercises.

Now what I do in the exercises is I give you guys the SAS code to do what I ask. So this one says print to the screen. Actually, I think I changed it to generate the contents. But we might as well print it to the screen, too, using a proc print or proc contents. I just want you guys to look at it in some fashion.

So I say print to the screen or like I did right here, write proc contents procedure step to view the descriptive information.

The birth data, 1990 to 2005 for indicated yeah county Louisiana, county No. 001. I'll give you the data set name. For Acadia County, what I've done is I've filtered my data so that it only includes for the data for Acadia County.

The reason I did that is because I'm going to have you guys link these data together. Now I had a lot of birth data, over a million observations. If I would have given you guys everything, it was over 200 megs. So that's the reason I went to only Acadia County, is because that's what I was able to do based on the size of the data set.

Okay. So what we're going to do is we're going to view the data for Acadia County. Now what I do here is simply read the data set in. Once the lib ref is defined, write a small data set to read these data in. Now, it's very important, when you're dealing with linking files, any time you link files together, you should write your SAS programs like this.

Indicate the number of observations you start with. Actually, indicate the number of observations is a comment next to each statement like this. I say data X 21 give me a blank statement to work in, place births 001 in there.

But here I say N equals 1504. That's important because you guys are going to need to keep track of these observations and make sure it makes sense when observations are getting removed or not showing up and things like this. When you're linking data, that's essential. Make sure the resulting number of observations is accurate and understand if you see some that aren't there that you think should be there, understand what happened in the SAS program.

You don't want that to be the reason that you lost observations. So I say here that I started with 15054 now, there are a few ways of figuring that number out. I could just double click the data set and take a look and see how many observations I had. Or I can run the proc contents here.

I'm pretty confident in what I've done here. I haven't done any fancy linking or anything yet. So I think it's pretty safe to say whatever number of observations I see for this contents procedure is going to be accurate.

Okay. So let's go ahead and go through that. What I did for Exercises 2 and 3 is basically repeat this process for the death data and the linking file.

And what you guys see at the top here is just a comment, okay, to describe what it is. What I want to do I'll do the first one up here on the screen -- actually I'll do all of them up on the screen if you like. Follow along.

Like I said, it's in your presentations.

UNKNOWN SPEAKER: I have a question [Inaudible] there's this year but I think they were born last year. So you have to link two years together. When you have children for this year but actually died for December that can die in January next year. So we also have this problem so the -- it's not perfect ratio.

What we do is the number is always this year's births, but when we need to do research, we link this year's and last year's births and this numerator is actually this year's deaths but the denominator is this year's births but kids included was actually post years.

JEFFREY SHAFFER: Right. You're talking about cohort data.

UNKNOWN SPEAKER: If it's actually when we're doing research here because we want to link births and deaths together. So numerator and denominator actually involve two years. Is that just cohort -- [Inaudible].

JEFFREY SHAFFER: Can I get back to your question after we link, after we link the data set so you can see what I did here, and then you can run it by me again? Because we're going to link the data sets in our next set of exercises.

And I'm sure there are competing ways for linking these together. Now, we have one way of linking these together to account for the cohort effect that I think you're talking about.

UNKNOWN SPEAKER: It's not just cohort. When you link two files together, even deaths died this year was born last year.

UNKNOWN SPEAKER: We pooled all of the births and pooled all of the deaths and then linked. So we're not doing it year by year. In that case we're doing it so that for every --

UNKNOWN SPEAKER: When we report year by year, MCH require every year.

UNKNOWN SPEAKER: No, this is different. We're pooling them all over -- to do a long-term trend to get all the births.

UNKNOWN SPEAKER: You do research we do it that way. But when we actually report individual years.

JEFFREY SHAFFER: I'm not sure I quite understand your question. But I will come back to -- let me address that again, though, after we link so you can see exactly what we can. We'll go through the whole process.

Okay. Let's go ahead and do this exercise. You guys can go ahead and open up SAS. Now, on your desktops -- now, I have a lot of stuff here that -- let me set this up so it looks just like yours.

Okay, you're going to have SAS 913 is going to be all these icons I set up on your desktop. You're going to have SAS 913, join point, GEODA sat scan. And right underneath you should have a folder called student sets. On your desktop you'll see a little group of icons like I have highlighted there. That's what we're going to be working with.

Now, the first thing, these three little packages, the join point and GEODA and sat scan we'll talk about later. Those are little tiny programs that we're going to use to perform analysis.

For now let's go ahead and double click SAS 913 and go ahead and get into SAS. And this is what you guys see. Okay. What I'd like to do is the first thing I'd like to do is maximize the program editor window. And like I said you guys at this point can go ahead and just type in what you guys see in the presentation. So I'm going to go ahead and type Title 1 at the top. MCH Workshop. Maybe 2008.

These statements can go in any order. DM account clear. Log clear. How about a lib ref? Let's define the lib ref. That will depend on your computer. And I can come around and figure out what it should be for your computer, but I'm going to figure it out for mine up here.

Now the way you can figure out your lib ref is click one of the files on your desktop. If you guys want, double click student data sets. Go into these and right click any of the data sets in here. I'll choose Births 001. I'll right click that file and go down to properties, and here's where that file lies, under where it says location. This is telling me -- this is giving me the lib ref or the library reference for all of the data sets we're going to use in this course.

I'm going to Control C that directory and that's going to be my lib ref for the entire course. That's where my SAS data sets lie.

Now, I'll come back into SAS. I like to call my lib ref path, my first one, short for directory or whatever. Doesn't have to be called path. Yours is going to be different. It's not going to say C:userJShaffer. I can't guarantee they'll be the same among you guys. But whatever that is, after you've found the data set and right clicked it, put it there as your directory.

Now what else am I missing? Options statement. Options LS equals 80. No date. No number. Okay. Now we're ready to begin.

What I wanted you guys to do, okay, we're going to say make a little comment here. Birth data. And you guys can just copy this. But I want you to think about it as you're going through and try to understand what we're doing here. Birth data. And this is just some kind of comment. Data exercise 2 underscore 1. You don't have to use the same temporary names I do, just be consistent.

Everybody has a preference for their temporary names. Some people like to use New 1, 1, 2, 3. It doesn't matter to me, whichever you're comfortable with.

Okay. Exercise 2.1. Let's go ahead and say set path.births001run, that reads it in. Then you can do whatever you want to do with it. You can run proc print or proc content. Since it has so many variables and so many observations I think I'm just going to run a proc contents.

Data equals EX 2 underscore 1. You guys want to get fancy and put all the titles in, you can. Problem No. 1 or how about Exercise 2.1. Run. That will be Problem 1 go ahead and hit the F8 key or click the running man to submit it.

You guys should see what it has. This is essentially a coding guide. If you run content, you're going to see the output look something like this. This is essentially a coding guide for the births data set.

Like I said, I subsetted this, I filtered this because the main births data set was so large.

Here are the variables. Here are their names, their types, their lengths and so on. But -- and you guys can repeat. Go ahead and repeat this process for the death data and the link data. The link file. And I'll give you guys a few minutes to finish that.

How are we doing on time? Do you want to take a break after this or do you want me to go into Chapter 3. We'll take a small break after you guys are finished. How many minutes are we going? 15? How many minutes for the break?

UNKNOWN SPEAKER: 15.

JEFFREY SHAFFER: We'll take a 15-minute break when we're finished with this, you guys, we'll come around and help you. For now the only thing different for your program is the lib ref. By the lib ref or the lib ref directory, by that I mean what I have highlighted up here. That won't be the same.

Find your data sets. Right click one of them and figure out which one that is. And if there's anyone who needs help, just go ahead and raise your hand.

UNKNOWN SPEAKER: I can't figure this out.

JEFFREY SHAFFER: Let's take a look at it. That print is real small. Did you guys set it that way? I wonder if we can figure out how to change that.

UNKNOWN SPEAKER: Copied it.

JEFFREY SHAFFER: You copied it. Was there an underscore there? Because I see an underscore here. I think that's the problem. See -- you can get it there, too, that's fine. But if you go to properties and go to end, there's no underscores in there. But you can get it up here, too, two places up here. Not all computers have this showing. That's why I didn't see it up there.

Now, what I'm going to do -- and it is a good idea to copy and paste these, because it's space sensitive, and it's very easy to make a mistake on that.

So this is your lib ref for the entire course. Okay. So every one of your programs will start just like that.

UNKNOWN SPEAKER: So just copy it?

JEFFREY SHAFFER: Yeah, these are called global statements. I just copy and paste those. Okay. And your comment here with a semicolon.

UNKNOWN SPEAKER: It's going to round --

JEFFREY SHAFFER: The rest of it is fine. Click that. Click the running man at the top and you're good to go. You can save that if you want. You guys can save to the desktop if you want. Okay. You guys have the permission to do that.

If you guys want to save everything to the desktop and then at the end of the course go ahead and drag everything to your thumb drive or USB drive you're welcome to do that.

If you want to save everything to your thumb drives right now you are welcome to do that. But you guys can save stuff to the desktop. You have permissions to do that.

The next thing we're going to talk about later is actually linking the data sets. We're going to go through some exercises and see how one approach for linking these data together.

Okay. Any other questions? I'll go ahead and do the last two up on the screen. Like I said, my exercises are pretty much set up like this the whole way through. Okay. So I'm going to say Data 22.

You don't have as many variables in the death data sets so you guys can probably run a proc print and make something out of it if you want. So it can be contents or print for the second one. After you're finished, you can take about 15. We'll say 20 minutes from now. 20 minutes. So after you're finished we'll take a short break.

Let's say 20 minutes from now, because I have a few people -- has everyone finished this?

UNKNOWN SPEAKER: Almost, yes.

JEFFREY SHAFFER: We'll say 20 minutes from this time. That's going to be -- I'm showing 8:44, which would be, what, 9:44. How about we say 10:00. Is my clock right on this computer?

UNKNOWN SPEAKER: No.

JEFFREY SHAFFER: It's not. What time are you guys showing?

UNKNOWN SPEAKER: Quarter to 10.

JEFFREY SHAFFER: Okay. Let's go ahead and make it 10.