MCH EPI Conference

Time Trend Analysis for MCH Outcomes in Large and Small Populations

December 8 – 11, 2008

JEFFREY SHAFFER: Okay. Let's go ahead and continue in Chapter 3. We're going to talk a little more about some linking issues and the actual problem in Chapter 3 so a problem. A live birth is defined as a newborn child exhibiting any evidence of life.

Okay. So that's the live birth in the United States. That may change from country to country. You want to always consider the definition of a live birth before you compare the number of live births to that for another region or another country or something like that. So that's important. But in the United States it's pretty universal. It's a pretty universal definition.

An annual IMR, or annual infant mortality rate, per 1,000 live births is defined as you see here. The convention is to use per 1,000 live births for infant mortality. It may be more for other types of mortality per 100,000 is very common. Makes it a little easier to read when you're talking about per thousand as opposed to a smaller number. You kind of get integers that are easier to understand, using the thousand-person multiplier.

So an annual IMR, or annual infant mortality, rate per 1,000 live births is as you see here. I've used the subscript I to denote the Ith year. This says IMRI equals

the number of deaths less than one year of age in the Ith year. I could be something like 1990, '91, '92, 2008.

And different statisticians have different ways of using, indicating a subscript. And we'll talk more about that later. Okay. The multiplier again is 1,000. That's the K that we talked about for the rate earlier, divided by the total number of live births in year I. So the population of interest is going to be the number of live births in the Ith year. Now, once again, the denominator, we like these denominators to be large.

Rates based on large denominators are always better than those based on small denominators. This definition you see of an IMR is referred to as a cross-sectional IMR. And the reason it is, is because we're considering year of birth and survivorship status simultaneously.

Some of the infants dying, let's say, in 1990. Or let's say you have a live birth in 1990. Well, if you have an infant death arising from that live birth that infant may not have died until 1991. That's very possible. Now, we're only considering less than one year of age here. So it couldn't be passed 1991, but it could certainly be 1991.

We're considering only those in the numerator who died in the same year. So we wouldn't be really considering that, that infant could have died in 1991.

We'll look at that later, look at that definition of an IMR later and those will be defined as cohort IMRs. But for now they're cross-sectional IMRs. And it's one way of just taking a snapshot of the number of deaths in a certain year divided by the number of births in a certain year.

And IMRs, in general, are a very common measure of state or nation failure. That's one of the most universal measures there are when you're talking about healthcare and things like this, you like to talk about IMRs. Keep in mind that these sometimes can be compared from region to region, country to country, if they have the same definition of IMR.

Now, if a country, if another country doesn't have the same definition of a live birth, or something like that, then the IMRs are not comparable. And I mention that here. Cross-sectional IMRs direct comparisons should not be made between IMRs with different live birth definitions, because it's going to -- the disparities are going to be a function of the definitions.

So they aren't comparable at all. Intercountry comparisons may be inappropriate. It's very common in other countries to have different definitions of live birth. But in the U.S. there's one definition of live birth, so the results can usually be compared from state to state.

Crude results are often misleading due to confounding factors. So we're going to talk about how to control for confounding factors later on. This is a crude rate, which is what this could be called, too, because you haven't controlled for anything here.

Establish risk factors for IMRs. Low birth weight. Obviously that's definitely a risk factor. Multiple gestational births. That's certainly a risk factor, too. Extreme gestational age. Smoking status and race, and we're going to talk a little bit about the race predictor with a lot of our analyses. In particular, blacks and whites. We're going to filter the data so we only have blacks and whites and we'll talk about some comparisons.

The next type of IMR that we'll talk about are called or referred to as cohort IMRs. Now, cohort IMRs, with cohort IMRs, the population of infants are born in the same year. Cross-sectional IMRs, based on the definition, may be confounded by certain exposures.

Like I said before, if we back up to a definition of a cross-sectional IMR, let's say you have a birth in 1990 and that child dies in 1991. The reason for the death may change based on the year.

Okay, for instance, a certain cohort may be exposed to a certain disease in a particular year. Well, maybe that disease went away the next year. Okay. So it's essentially confounded by the year of birth.

Okay. And what you're essentially doing there is accounting for that using cohort IMRs, and this is the definition. I have the notation here, but the concept is pretty straightforward. Now, in my notation, the subscript I is going to refer to the year, the lower subscript I. And for cohort IMRs, the superscript is going to be, correspond to the year of the birth.

So let's read through this definition. The denominator's pretty straightforward. NI is going to be the number of live births in the Ith year. So if we were interested in the number of live births in 1990 this would say N 1990, maybe N 90 something like that. The thousand, we've seen that multiplier before. Let's look at the first one. Why sub I? The subscript means year.

So why I? And Ys going to be the deaths. Ys are the deaths and the Ns are the number of live births. So Y sub I says the number of deaths in a certain year.

Now, the superscript is going to be the year they were born. So this says if I had Y 90 with a subscript up here of 90, notice this I is the same as that I, so they have to be the same. That says that I have an infant that was born in 1990 and died in 1990.

Now, notice this one. This YI, it goes to I plus 1 as the superscript. So with the same group, if I was talking about those dying in 1990, this would say Y 90, meaning those who were born in 1990, the lower subscript is those born in 1990. The upper subscript are going to be those who died in the next year, I plus 1.

And the crux of this is you read this as the sum of these same births, the sum, the numerator is going to be the sum of those dying in that year or the next year.

And that defines a cohort IMR, infant mortality rate. Okay. Once again, Y sub I superscript I is the number of deaths in the year I for those born in year I. And NI is the number of live births in year I.

So Y sub I with superscript I plus 1 is the number of deaths in the next year.

UNKNOWN SPEAKER: Does that mean the range of possible ages of the infant is from 0 to 2?

JEFFREY SHAFFER: No, 0 to 1. We're only considering live births. So a live birth, an infant live birth is going to be those under one year of age exhibiting any signs of life. So we don't even have any of those in our data set. These are only --

UNKNOWN SPEAKER: 23 months?

JEFFREY SHAFFER: Excuse me?

UNKNOWN SPEAKER: That's what I was thinking.

UNKNOWN SPEAKER: Never mind. I'm confused. Never mind.

JEFFREY SHAFFER: They're all under one year. All the infants we considered are under one year in all of the data sets we didn't have any outside of that.

UNKNOWN SPEAKER: When they die it will be the next year.

UNKNOWN SPEAKER: Okay. All right.

JEFFREY SHAFFER: You can have an infant born in 1990 and die in 1991.

UNKNOWN SPEAKER: Born in January, let's say, 2000, died December of 2001.

JEFFREY SHAFFER: You're born when?

UNKNOWN SPEAKER: 2000.

UNKNOWN SPEAKER: 2000, what month.

UNKNOWN SPEAKER: January 2000.

UNKNOWN SPEAKER: Died December 2001.

JEFFREY SHAFFER: That would be here. You were born here and decided 2001.

UNKNOWN SPEAKER: There's no visible --

UNKNOWN SPEAKER: It would be more than 12 months of age, if it was under -- if it was born in say December, 2000, and they died in January 2001, if you didn't account for that extra year you would miss that data. So by going into 2001, it's still under 12 months, you're going to capture that data, does that make more sense?

UNKNOWN SPEAKER: I think there's an invisible step here you subset. We're just not seeing that.

JEFFREY SHAFFER: Okay. We can take a closer look at that here. Here are the death totals by year for Louisiana. Year of death is what you see here, going across the top. And down the year of birth column you see the year of birth. Okay

of the so in 1991, the number of cross-sectional deaths would be 604. You had 604 total infant.deaths.

95 were born in 1990 and 1991, 509. You add those two together. You disregard the year of birth for cross-sectional births.

For cohort deaths, though, it's going to be 613. It's going to be this plus this for 613. So cross-sectional, down the column. Cohort across the row. Is that clear now?

All right. So that defines a cohort IMR. Now, the denominator is just going to be the number of births in that year. That hasn't changed. So let's take a look at an example. Okay. 71,910 live births were reported in Louisiana for 1991. Okay. So based on that, based on these data here, this is where I'm getting the death totals from, based on these data here, what is the cohort IMR for 1991? Well, that's going to be IMR 91 is going to be Y sub 91 superscript 91. That says those who were born in '91 and died in 1991, plus those who were born in '91 and died in '92, divided by the total number of those born in 1991 times a thousand. Times a thousand, substitution, I get 80.5 per thousand.

Okay. So this is a cohort IMR. It's not going to be the same. It's not going to be the same as the cross-sectional one. But this is how it's calculated. So viewing the data from a cohort perspective, it is estimated that there were 8.52 infant

deaths per thousand births in Louisiana for 1991. Does everyone follow this substitution?

Okay. So let's talk a little more about the problem. We're interested in analyzing IMR trends in Louisiana over this time period. We want to generate and test hypotheses related to these trends using cross-sectional IMRs, cohort IMRs and we also want to talk about some spatial trends as we talked about earlier.

So these are the types of -- something about their trends and say something about what they're related to what affects the trends and things like that. So the data linkage. Back to this. We printed out the data sets in the first exercise. We ran a proc contents, looked at what they had, talked about the coding guides.

Here's what you have -- here's what you're presented with and here's what you want to do. We're given the births data set. We're given the deaths data set and the idea is to link these together.

And we'll use the link file that you guys ran a proc contents on earlier to link these together, to obtain one linked data set, that way you can associate the birth with the death and you can talk about predictors and things like that that appear in both the births and deaths data sets.

Some reasons for linkage? Number one you can calculate the number of cohort deaths. You can't calculate the cohort deaths unless these are linked.

And you have access to the covariates. You have access to the covariates in the deaths file. You really don't if you don't link these together. There's no way to associate the births with the deaths. So if you can't do that, you really can't talk about the covariates in the deaths file. The objective is to link these together.

It's common practice to link these together. But, like I said, there's different ways of doing this. We're going to talk about one way of doing it. Many of you may already be given the linked data set to work with.

I talked to a few people that are already presented with the linked data set and you may be working with that. But here's what's going on before you get it. Okay. So for the births data set, like I said, any type of data linkage you always want to keep track of your observations. And the births data set we had over a million observations.

In the linked data set, we have a little over 10,000, and in the deaths data set a little over 10,000 as well. 287 observations were excluded from births with no matches in births/deaths. So what that's saying is that when I link these together, I found 287 indicated in births, 287 IDs were not in here.

Okay. So what do you do about that? There's not much you can do about it. You can do some weighting. We didn't do that here, but that's probably be a good way to start. You want to at least think about why you lost this many observations relative to births it's not all that many.

But there's still certain patterns associated with these 287 observations excluded that are not probably not consistent with the remaining million you see here.

Why were they excluded? Some potential reasons include data reporting. That's probably one of the biggest ones. Other potential reasons might include moving. Moving the different locations. That really results in a lot of unmatching when people move from one region to another, that can be a little problematic.

Okay. But we ended up with 287 that didn't show up in the linked file. Also, there were 697 observations excluded from deaths with no matches in the linked file. Okay. So when we tried to match these together, there were 687 deaths that didn't show up in the linked file. So we couldn't link those together.

So you want to think about why these weren't linked together, who are very low birth weight. There are probably physicians that just don't consider that as much as they should -- just like they would a full-term baby. They just don't report the same way. I think it's safe to say that.

So you want to think about things like that. There's not much you can do at this point, like I said, other than weighting the data and trying to figure out what the pattern of these is.

There's a certain bias incorporated into the results based on excluding these unmatched observations. Okay. No birth data was excluded for the unlinked data.

So when you're not linking, you don't exclude anything. For the cross-sectional IMRs, you don't need to exclude anything. It's just the number of births -- number of deaths in one year divided by the number of births.

So this is problematic when you're linking the data files together, you run into these problems. For the linked data, we attempted to correct for some of this by including only those infants that were at least 500 grams or more of birth weight. Okay. We feel that the results are a little more stable there. We feel that a lot of the unmatched observations were probably low birth weight babies.

Okay. So this would correct for some of the unmatched deaths. That's where you're really going to affect that, is with the deaths, because if the deaths weren't reported properly, it could very well be that those are low birth weight babies.

There's no way to verify that, but we suspect that. And we'll talk about the last note I have here in a second when we actually go through the calculation.

I can do that now, because we're going to do some more exercises. Let me back up and say one thing about removing year 2005 for the cohort data. If I back up to the definition of a cohort IMR, our data range from 1990 to 2005. So if I'm asked to calculate the IMR, the cohort IMR in terms of this definition, for 2005, notice here that I'll need the number of deaths in 2005, the number of those infants dying in 2005 -- born in 2005 and dying in 2006.

So the death is at the top. Your birth year is at the bottom. So I need the number dying -- born in 2005 and dying in 2006. Now, there's no way to get that for these data, because I was only given data ranging from 1990 to 2005.

So what you're going to have to do is you're going to have to consider one year less than your data range for cohort IMRs. There's no other way to do that.

Okay. So that's what I did. So if we talk about cross-sectional IMRs we can cover both end points of the range. Talk about cohort IMRs you'll lose a year at the end, exactly one year.

So with that in mind, what I want to do is I want to go ahead and go through a linking process, and, like I said, this is not the only way to link these together.

There are some other issues you could consider. Certainly weighting -- okay, that's the one issue that could certainly be considered, and that's reported in a lot of the literature, how to account for the unmatched observations. That's one thing we need to consider. What we did is we chopped it off as those births or those deaths had to be at least 500 grams.

And what that does, when you do that, now, notice, we are eliminating even more observations than we started with. Let's back up to that.

We go ahead and we say we want to get rid of those less than 500 grams. Right here I say that. Now, if you think about this, you're eliminating more observations than you would have eliminated from just doing this. We lost 697 observations from deaths with no matches, just unmatched, 697.

So now we're going to lose even more because we're including only those less than 500 grams. But what you're doing here is you're eliminating the bias introduced because you're thinking that most of these, or a lot of these -- these are going to be disproportionately underweight babies. Does everyone see that? If you eliminate them all, then you're hoping to eliminate at least some of the bias. So that's the motivation for doing that. That's one of the motivations for doing that.

Okay. Now, what I want to do is I want you guys to go ahead and go through the linking process and the way we did it. What I've done here -- now, these exercises are not in the presentation. These ones are going to be in the manual. From here on out all exercises will be given in the manual.

So I'll call your attention to page number 11 of your manuals. That's where I've given you guys the code to do this. So it's lib ref directory, same all of that. This is the code to do that. I'm going to do this up on the screen and I'm going to kind of talk my way through it and explain how we went about matching these data.

So I'm going to do this with you guys. But this is where I'm getting the code. Okay. So I'm going to come back into SAS. You guys can save these as different program files if you want. I'll get rid of the stuff we just did. And I'll keep my same lib ref.

What I've asked you guys to do in Exercise 3.1 -- let's put that up here. It says link the birth and death data for County 001. Data set names are births 001, deaths 001 and births/deaths. And we've read those in.

If you guys have them from the previous exercise, that's fine. I'm going to start from the beginning and assume that this is just a completely independent exercise. Okay. So let me go ahead and do this. I'm going to say data X 3.1. Now

this is exactly what we did before. No difference. I'm just reading in the three data sets again.

Okay. Set, path.births deaths, births, 001, run. Now, what I'd like to do when, like I said, anytime you're linking data sets you want to keep track of the observations. So if you guys want to run the program now and look at the log file and determine the observations, you can.

You can put it in there later. At least put it in at some point because you want to see where these things are going and understand why they're being removed.

Okay. So I'm going to put, next to this, I'm going to say N equals 15054. Now this is something I would have gotten after running the program. Look at the log file, like I said. Run a proc contents. Do something like that, determine the number of observations.

That's reading in the births data set. If you guys want, you can add the comments I have as well. This is just a comment to describe what it is. Okay. Let's read in the deaths data set next. Okay.

Up here I give these data sets -- like I said, you guys don't have to use my same data set names. I call this X 3 New 1. I could have called it X 3.1-A or something like that. You want these to be unique, okay, is the idea. You guys can get

creative here. Exercise 3.1 Data Set A. I said Exercise 3.1 New 1. It's a new data set one. If you like that notation, you can use it. But it's not going to influence your results at all.

Okay. I'm going to say Exercise 3.1 New 2. I use this because I'm going to have multiple data sets for Exercise 3.1. It's the motivation for that. This is going to be the desk file. Like I said, we did all this in the first set of exercises.

N is 166 here. Why is N 166? Because I filtered these data because we only have one county out of Louisiana. That's why it's so small.

But it's exactly the same process for larger data sets. You just have more observations. So what I'd want to do is run a proc print or a proc content and see how many observations are in that data set so I can kind of manage things.

Next we'll read in the link file. Let's say link file or something like that. I know this is kind of tedious to type this all in, but the reason I'm having you guys do that is because it's a little easier to follow if you just go ahead and retype this stuff. It's a little easier to follow what I did.

It's kind of hard to articulate it without actually going through the process. Okay. So back to what we said before. Ideally, I would have 166 observations in births/deaths. Now, the reason that that is so much different, that number there, that 10245 from the 166 is because I gave you the linked file for the entire state

of Louisiana. And I had to do that. I couldn't filter this data based on a county. I couldn't do it because I didn't have the county code.

So that's why these -- these numbers look very different. But they won't always look that much different in practice. The problem is this one's for the entire state. These two aren't. But you're going to see it doesn't matter that I've given you the one for the entire state.

Redundant observations are just going to fall out. So that's just reading in the three files. Okay. Now let's go ahead and start our linking process.

Okay. So I'm going to say something like join births.

UNKNOWN SPEAKER: We just had a question. Who creates that link file?

JEFFREY SHAFFER: We got it from vital records. Where does it come from?

UNKNOWN SPEAKER: Vital records. In our place the folks in vital records do it.

JEFFREY SHAFFER: It was something we were presented with, from vital records.

UNKNOWN SPEAKER: We didn't always agree with what they did. But they set there and they link these and look -- probably looked at data that we were probably not privy to see which death went on with which birth, and they tried their best to link them as closely as they could.

I must say you may well have better reporting in other states than we do. Because we have lots of births that occur in emergency rooms and people just go home, and so there's absolutely nothing on them. If those were early deaths, then we know practically nothing about them.

So it's entirely possible that -- the well-run state you are probably going to do a lot better.

JEFFREY SHAFFER: Okay. I make a note here, and I say that note that there are 10,081 observations in births/deaths with no matching observations in births. And we're going to see that that isn't really going to matter. A large part of that, like I said, is because I've given you the link file for the entire state of Louisiana.

But I want you guys to get in the habit of keeping track of what observation -- what you're dealing with. So let's start the joining process and one way to join this, what you want to do first is I join births to births deaths. So you can start by joining births to the linked files or joining deaths to the link file. It doesn't matter where you start.

So the idea is to come here. And let me move up to the, right here. The idea is to join these together. Now, I can start by going from deaths to births, births/deaths, or births to births/deaths. But I can't go directly from births to deaths which is what I really want to do. I can't do it. The data doesn't allow for it.

I'm going to start by joining births and births/deaths.

When I join these together, since I'm going to be referencing variables in a bistatement. I'm going to go ahead and sort both data sets by what I'm going to include in the bistatement. So I'm going to say proc sort, data equals X 3.1.

So what is the purpose of this? SAS requires that any variables referenced in a bistatement need to be sorted beforehand. So you ask yourself, which variable combination are you going to be including in a bistatement. Well, we're going to not only be including verify D, which is one of the variables we're going to be joining on, we're going to be including birth ID, birth year combination.

Like I told you guys before, if you start in 1990 with birthday ID No. 1, that birth ID No. 1 could very well show up in '92 and those aren't the same infants. They're different.

So we'll go ahead and do that for the data sets we're joining together. We'll save the next one where it's going to be the births/deaths, data equals. I'm missing an equals here.

This is Exercise 3.1 New 3. No space there. I had a couple of typos up here, guys. Check what I have now. I think I fixed them all.

This is going to be right. Okay. Now we're ready to perform the merge. Okay or the link. Okay. So I'm going to come down here and I'm going to say data.

Okay. Time permitting we're going to go through all this. But I at least want to get you guys off and running on this first part and explain the things you're going to need to consider when you merge these together.

Okay. What I've done here -- now the sorting is no big deal. That's just to satisfy SAS. The merge actually takes place in the last data step you see on the screen. This says give me a new blank space to work in. And, of course, you want to merge the births with the linked data file, which is going to be the new 3 you see here. So you're just merging those together.

Then you're going to say that you want to merge them by combination of birth ID and birth year. You're saying the unique identifier is a combination of birth ID and birth year. We talked about that.

I use the IN equals operator so I'm able to keep track of that combination for each of the infant observations. Now, what I say here is I say if A equals 1, that says if A is true and that means only include those in the merge that contribute to the births data set.

In other words, if they didn't show up here -- I'm sorry, if they did show up, you must have a birth is what this is for. This says -- this combination must be included in births.

If you have that combination that shows up in the linked data file, let's say you have a combination of ID 1 in year 1990 that is not in births but it is in the linked file, which is possible. Why does that happen?

There are many things we could -- many things we could talk about there. Who knows why it happens. But it could happen that you could have a combination of that ID. Maybe it's a data entry error or a reporting error. Something like that. You don't want those observations that are in the linked data set and not in the births data set. They have no appeal whatsoever. So this is saying that this combination must appear in births to be considered.

Now, you're going to lose some observations from the linked file from that. Okay. You're going to lose some. This is how you want to go through and merge these.

Now, I'm just going to briefly discuss -- I think -- let's go ahead and finish this exercise. Okay. I don't know if we'll go ahead there you and type all of Exercise 3, too. Let's finish this one. Let's go ahead and join now this data set, this combined data set with deaths to complete the process.

We've linked now births to the linked data file. Let's link this big data set to the best data file. And that's what I do next. Down here I say join -- you guys should make comments like this. Join above data set above with deaths. And that will complete the join. A very crude join and we still have some issues to work out. But this will complete the process for the basic join.

Okay. So the question is, the first thing you want to ask yourself: What are you going to be joining on? Whatever you're going to be joining on needs to be sorted. And in any data set that references that combination, it's going to need to be sorted on.

Okay. So here we're going to join on death ID and death year. That's what we have the join on in the death data set. We have no birth ID in birth year. If we had that in the death data set we wouldn't need to do this. So what I'm going to do is I'm going to say I know I'm going to be referencing that combination in a by statement later. So I'm going to say proc sort data equals. The data set we just created -- these can be joined in any order. I'll choose the one we just created first.

Let's say sort X 3.1 New 4 by death ID or death -- year.

JEFFREY SHAFFER: Death year. You're right. Because that combination -- you're going to have duplicates of year but you're not going to have duplicates for year and within year. So it should be first sorted by death year.

And within that by death ID. Ascending is fine. Ascending, descending, it doesn't matter at this point. Let's do the same thing for the death data set, which is New 2. X 31 New 2. Like I said these don't have to be in any order. I'm missing a semicolon.

And this is just to satisfy SAS. Now, let's go ahead and complete the merge. The same process. We need to think about what we want to eliminate. I don't think I'd want to include any death ID death years that didn't appear in the combined data set. So these are things we're going to need to think about. That I'm going to say let's go ahead and merge these together. That should complete it.

And what this means, once again, think about what if B equals 1 means. The IN operator simply says keep track of what you're merging in both data sets. It's a boolean variable that's going to be 1 if true and 0 if false. It only has two values. Why did I call it A? I just chose it as a random variable name. That's commonly used in the literature.

But I could call that any name I like. What that says is I just want to keep track of the death year, death ID combinations for both of these data sets I'm merging.

Okay. So I go ahead and say if B equals 1, that says the combination of death year and death ID must be included in X 3.1 New 4 which means it must be included in the linked data set. If that death ID -- if that death ID/death year combination found in here is not also included in the linked data set, those observations are gone.

Those aren't ones you want. You can't link them anyhow. You're just going to end up with more unmatched records. Go ahead and get rid of them.

This is one way to link these data together. Now, like I said there are a lot of other issues that you could consider here. What about those observations, those unmatched observations? And like I said, we have to -- we could consider weighting these.

We went ahead and subsetted these to only include those births greater than or equal to 500 grams. Okay. Now, if you guys look at Exercise 3.2 that's a basic join. These data are joined at this point. Now we get into the other issues such as birth weight and things like that. What I do in Exercise 3.2 is I talk about the

cohort IMRs. Now where do we want to get rid of those observations that corresponded to low birth weight babies?

We could maybe do that right here, filter the data right here. Right now is when we started thinking about that, when to get rid of those. What I do in Exercise 3.2 is I guess I would like to go through -- I don't know if we're going to type all of this out, but I want to type the first couple of statements just to give you guys an idea how to calculate cohort IMRs. It isn't that hard in SAS. So what I'm going to do here, let's go ahead and go to Exercise 3. X 3.1 New 5. That linked together the births to the deaths, everything that's in there.

Now what I'm going to do is I'm going to say data. We're moving up to Exercise 3.2. I guess I called it X 3.1 New 6 but you can call, like I said, what you want to call it. I guess I'll stick with the notes so you guys can stick with me. But in practice you'd probably say Exercise 3.2. Maybe I should have done that.

Okay. New 2. We're going to say set. Not New 2. This is New 6 we're up to. Okay. We've made an exact copy. Now what we're going to do, just an exact copy of New 5 there. Now we'll get rid of the low birth weight babies to try to control for some of that bias for the unmatched observations.

So I'm going to say if birth -- if weight is greater than or equal to 500. And this is where the weighting would come in as well. Okay. If weight is greater or equal to

500, I further subset these as if birth year is not equal to 2005. Now for cross-sectional IMRs, I wouldn't do that. Like I said, but for cohort IMRs I have to because I can't calculate a cohort IMR for 2005. I don't know if they died in 2006 or not. It's impossible for me to say based on the range of my data. It's cut off at 2005, so I lose a year for cohort stuff.

Now let's go ahead and calculate the cohort, I think I've just asked you guys to calculate the cohort deaths. Well, I've asked you guys to calculate a variable called See Death Here that assigns the cohort death year for each observed infant death.

So with this variable we would be able to calculate the cohort IMRs. So let's go ahead and write one of these steps. Like I said, it's pretty easy to do. Looks harder than it is.

Based on the definition -- so where is this coming from? Okay. Let's go back to the definition of a cohort IMR. It's right here. Or just the cohort deaths. Maybe you guys find it easier to look at this table, the cohort deaths.

The cohort deaths in 1991 are going to be the number of deaths in 1991 plus the number of deaths in 1992. This is an aggregate thing you guys see here.

We're not talking about joining. We're just talking about here are the totals. We're going to do this for each infant death. We're going to link -- we're working with the linked data here.

Okay. So what we have to do is we have to get a hold of those in a certain year and if they died the next year, we also want to consider those deaths as well. Okay. So let's come back to SAS, because keep this definition in mind. Keep this table and this definition in mind.

Because this is what we're doing. We're essentially going to be doing part of this. We're calculating the cohort deaths. The numerator of that. Okay. So let's go ahead and say -- I say if birth year -- like I said, this is pretty straightforward. If birth year is 1990 and death year is in 1990, then death year. Okay. See this is one, see death year for cohort death year.

I say here so that I'll be able to calculate the cohort deaths, I say if your birth year is 1990 and you died in either 1990 or 1991, your cohort death year or C -- did I call it C birth -- I call it C death year. I say it's 1990.

And give me one variable, one variable that summarizes the deaths for these infants based on a cohort definition. Based on the cohort definition of the death, you could be born in one year. Okay. So if you're born in 1990, the corresponding deaths for the cohort IMR are going to include those dying in 1990 and 1991.

For those born in 1990 -- does everyone see what I mean? That's how you do that in SAS. Now, whether you wanted to call it C death year, that's up to you. That's a variable name I made up. There's probably a more descriptive variable for that. But that's what these are. These are going to be the cohort deaths.

And once you have it associated with a single year, you can calculate the cohort IMRs for 1990 now. The cohort IMRs in 1990 is going to be this number divided by the total number of live births in 1990. And that's the motivation for this.

The numerator is the hard part to calculate. The denominator is easy. Okay. So I go through and I did all of that. Now, there may be an easier way to write the code for this. This is pretty easy to understand the way I have it on here. But I repeated this process the whole way up through 2004.

And now you guys get an idea of why I could not use 2005, because if I put 2005 here for the birth year, then I don't have anything to put here. I can't put a 2006 there. I don't have that data.

So you're stuck going the whole way up to 2004 for the cohorts. Cohort deaths is what we calculated here. So this number resulting from this would be -- well, for a particular infant this infant, if we look at this from an observation to observation, this would be a -- this infant would either contribute to that or not contribute to

that, based on that infant's birth year. If that infant was born in 1990 and died in either of these two, then its cohort death year would be 1990.

What about those that don't die? Well they're not considered here. They're not considered at all. The numerator of the cohort IMRs only involves deaths. So that would go into the denominator. Okay. That would fall into the denominator, and that's what allows the denominator to differ greatly from the numerator, is those wouldn't go into the numerator.

Does everyone understand how to calculate cohort IMRs now? Now, I'm not going to type all this stuff out, and you guys don't have to either, but it's simply repeating that for every single year. That's really all that's being done there.

For the sake of time, I want to go ahead and just take a look at Exercise 2.3. We aren't going to type this in. I say --

UNKNOWN SPEAKER: 3.2.

JEFFREY SHAFFER: Exercise 3.3, I say continuing Exercise 3.2, create a categorical variable called race M that classifies mother's race as one as at white, two as black and zero as otherwise. I'm basically showing you guys-- I'm sure most of you probably already know how to calculate a categorical variable.

But I'm sure I have some beginners, some not beginners. If you haven't seen that before, there's how you do it.

That's how you create a categorical variable. And, like I said, we'd certainly want to collapse all the other race categories. The other, the Vietnamese, Filipino, there just aren't enough of them unless you talk about the U.S. as a whole. We really didn't have that many to get any kind of stable results from.

Exercise 3.4, I say here, continuing Exercise 3.3 generate a data set containing the number of cohort death births and deaths by year and race. I'm just going to briefly discuss what I did here.

The purpose, if you're asked to calculate the cohort IMRs, the cohort IMRs, this is how you'd do it. This is how you'd finish the process. Step one is link the data sets together. You guys can't calculate the cohort IMRs without doing that because you need to know when the infant died. It depends on birth year and death year. So there's no way to say if you're given an infant you don't know when it died, when the baby died, if you don't join the data.

So to calculate cohort IMRs you need the link first. After doing that, after calculating the numerators, which we just did, the cohort deaths, what you're going to be interested in doing is summing these up.

See, this one we just looked at, this is going to say whether a single infant. This is going to go observation to observation and assign if you happen to be -- if that baby happened to meet that criteria. It's going to say it. Then it's going to change this variable to 1990.

We're going to have a lot that aren't assigned anything because they didn't die. Now, what we want to do is we want to go in there and count these up. How many 1990s do we have? Do we have a thousand of them? Well, what we want to do is count those up, by year, and by race. I just added race in there arbitrarily to show you guys you're probably going to want to count them up by different predictors, too, by classifying based on a combination of year and something else.

So what I did here is I used the output statement for the means procedure to do that.

The N option is kind of convenient for actually keeping track. You don't want to add up 1990 plus 1990, 1990. You want that to contribute one. It's frequency data. You're counting how many do you have is what you're interested in. Not adding that number up.

So if you use the N option as I've done here in the output statement for the means procedure, you can easily go ahead and print those to the screen. And

once you have the cohort births and the cohort deaths then you can calculate cohort IMRs. It's just the number of deaths divided by the number of births times 1,000. It's a very simple calculation.

Most of the work is right here and in the linkage. That's almost all of the work. Okay. And I did say here where the mother's race is not equal to other. I don't know. Whether you wanted to see that in there you could easily eliminate that where statement to see all three categories of race. But what you're going to get here, if you run a proc print on this X 3.1 New 8 is you're going to get a listing of the cohort deaths in a certain year, 1990 to 2005, the number of cohort deaths the number of cohort births, and that's going to be separate for each race.

So with that information, I'd be able to calculate the cohort IMRs for whites and blacks separately in 1990, 1991, 1992 and so on. Okay. So that's the whole process that we went through to get the cohort data.

Like I said cross sectional data is easy. You don't have to do all this. You don't have to join it together. And like Dr. Mather was saying, sometimes you aren't, by law you aren't supposed to join it together to report the cross sectional data. There are ramifications for joining it together as we saw before. All those unmatched observations. How do you deal with those. You don't have that problem with the cross sectional IMRs.

But you lose the ability to talk about the cohort effects. You lose the ability to link the variables from the birth data set to the death data set. So there's -- you have to ask yourself which one you're interested in. But that is how you calculate the cohort IMRs.

Okay. For the sake of time, I guess we won't type all that in. What I want to do next is move on to the next chapter. Okay. Let's see. View slide show.

Okay. Now we're going to talk about some exploratory trend analysis techniques. I think it's up and running now. Okay. My computer's a little slow right now.

Okay. We talked about this. Okay. Now for some analysis. Okay. These are going to be exploratory results. Like I said, we get more into the modeling and more advanced techniques later today, this afternoon and also tomorrow. Okay. We're going to begin with the simplest type of analytical technique. We'll begin with some time trend plots.

And after this we'll talk about some methods of describing significance. Okay. The most basic technique for analyzing time trends is a time trend plot. It's very -- and much of what you guys do or are going to be doing in practice involves using simple things like this. There's certainly nothing wrong with that.

In practice, sometimes the best thing is a very simple thing. So these are very informative. Okay. This is a time trend plot. And it's simply a plot, of course, that has a numeric time scale as the X axis. And the outcome of interest as the vertical axis.

So what you guys see here is a time trend plot for the cross sectional IMRs for Louisiana from 1990 to 2005. So this is what you see. Now, what we're interested in doing is describing these trends. This is by race. So blacks are here. Whites are here.

It seems like they have the same trend. Okay. There's a very similar pattern. Of course, the blacks have a much higher IMR, infant mortality rate, about two-fold. But the pattern's very similar.

What we're interested in are the ups and downs and things like that. Okay. This is a time trend plot. And this is one of the first things you guys should do when you're working with data like this. And now you can't really include too many categories, because the plots get confusing.

A lot of plots like this is pretty informative. Now, we could have added a few more lines, but if you add too many different categories like other covariates and things like that, the plots get confusing really fast.

So simple is very good with these. Okay. These are cross-sectional IMRs. So what you guys are seeing here at each point is going to be the number of deaths divided by the number of births. And this one would be for blacks in 1991. This one would be for whites in 1991.

That doesn't require data linking. That's just how many births, how many deaths did you have. That isn't considering any cohort effects at all. Here's what you see for the cohort IMRs. Now, we want to pay attention to the difference between this -- let's take a race arbitrarily. How about the blacks? We're looking at IMRs between it looks like a little under 13, the whole way up to 17 here. That looks like the range of those IMRs for blacks.

Well, why is it that this one's so low. This is the cohort IMR here, but recall when I joined in data set I took out all those ones that were less than 500 grams. Well, obviously that's going to drive the IMR down. It has to.

Okay. Less than 500 grams is a very strong predictor for infant mortality. So when you take out the less than 500 grams, you're going to be driving the IMR down. There's no doubt.

And we see that from the cross-sectional, because we didn't take them out for the cross-sectional. We didn't have to. We didn't have any unmatched observations or anything like that. No reason to take them out here. But we did

have that issue with the cohort IMRs because we had to join the data sets together. We lost a bunch of observations based on the linking reporting and so on.

So we got rid of those less than 500 grams, and here's what we got. Okay. Now, what you see here, of course, these ones go down. Now, what I did is I kind of superimposed the all weights right above it. So this gives you an idea. This is a good time trend plot, too.

This isn't too confusing. Like I said, I see sometimes plots that include -- people try to include too much on plots sometimes, and it's actually counterproductive. This is pretty good here. Not too many lines. Pretty easy to follow. And if you're going to take things out, take things out of your data set, take observations out, filter the data set, this is a nice way in your work to illustrate what you did and the effects of doing that.

Because taking these out, now we may have considered that a good idea, but someone else may not. So you can illustrate what you did here and what effect that's going to have. Whether it's infant mortality or anything, any of your work that you do.

Okay. So here's what you see here. Like we said, you're driving the IMR down. Can't calculate 2005 because we talked about that issue. And here's what they

look like. Now, we're looking at about, no matter what we look at, it seems like we have about a two-fold difference between blacks and whites. So remember that the whole way through because we're going to talk about, when we model these data, we're going to talk about -- first of all, we're going to ask ourselves did we do it right? Second of all, if we did do it right then why is that gone? Why is that difference changed.

So right in your mind you should be thinking blacks about two times more than whites. And that's what it looks like in a crude fashion. We didn't -- no controlling or anything like that. Okay. I want to talk also -- we're going to begin this, like I said, with plots. I want to talk next about semi log plots. Semi log time trend plots.

With semi log plots the outcome variable is on a log scale. And natural log rhythm, and why would we do this? This facilitates interpretation. It makes the data scale more to linear form. We aren't so concerned with this thing here, not all data are even exponentially distributed.

Many forms are. I think you guys can see that survival, just survivorship of people over time as you get older, there's a much more higher likelihood of death at the end. That would be exponentially distributed at the end-of-life.

So one property of this is that it's going to convert that kind of data to linear form, but we're more interested in the one right before it. Now, a drawback of semi log

plots are they may fail to reveal outliers. Here's an example of one. It looks like I'm using the cross-sectional IMRs. I'll use the cross-sectional IMRs sometimes and I'll use the cohort IMRs sometimes.

I'll tell you guys when I'm using which. Obviously when we're interested in modeling more than just a few covariates and we're interested in the covariates in the death data set, we can't use the cross-sectional IMRs. So we'll reach a point where we have no use for the cross-sectional IMRs. But in the beginning we're going to use those quite a bit.

Here's a semi-log plot. Now I say semi-log because the outcome variable is on the log scale. Log, log plots are when the outcome variable and the variable along the X axis is on a log scale.

The way you look at these, this is on a log scale. So, like I said, it makes your data much more manageable for large data ranges. And you can talk about the differences between these. Now we can see that it still looks like about a two-fold difference. And like I said our scale here is .5 to 1.5. So if you had a lot of changes in the data covering a huge range, this would be appealing.

It's to make that range more manageable. Now, we could say quite a bit, I don't want to give you guys the impression that these plots weren't sufficient, because those aren't -- that data is pretty manageable already. And we can almost say

exactly the same thing we see there as we see here. But this may make the pattern more manageable if your scale's too large.

The first test we're going to look at -- now, those are our graphical techniques. We'll revisit some of those later. We'll use them in conjunction with some of the things we're going to do when we generate predicted outcomes and things like this. So we'll see stuff like that again.

Next I want to talk about some simple ways for testing for trend. Okay. Let's back up to the plot. It always starts with the time trend plot. That's a very good way to start this. You come here and you ask yourself, is there a trend. And the next thing you want to ask yourself is what type of trend, linear trend, quadratic trend, seasonal trend. One abrupt change. There are millions of different types of trends.

So what we're going to do first, and we're going to test for different types of trends. We're going to test the hypothesis of monotonicity, in other words, is there a increasing or decreasing trend in IMRs for Louisiana? Does IMR 1, when I use the subscript 1 I'm referring to the first year. So for 1 that means 1990 in this case.

Sometimes I'll do that. That means the first year. K is going to be the last year.

Sometimes you guys might want to do that from the beginning, and I'll tell you why when we work with regression. You guys can see when you put large numbers here and you go to use those in calculations it's not always very efficient. So sometimes you will substitute for the first year or the first time period the number 1 or something like that and 2 for the second. It's a lot easier to perform calculations based on smaller numbers than larger numbers.

So what we're interested is doing is testing these hypotheses. Is the IMR monotonically increasing or decreasing? And that's significantly monotonically increasing or decreasing. Everyone's eyes see different things here. We don't know what we're going to say about that.

We do know one thing: We would be surprised if we see just this outrageous trend in any of our analysis, because this trend isn't that great. It isn't that strong. So we need to keep that plot in the back of our minds when we're going through all of our analyses to make sure everything coincides with that plot in some way or another.

If we end up with something new after controlling for covariates, great. But we at least need to keep in the back of our minds we started out with that. If it's that much different, then maybe we found something. Maybe there's a reason for that and we need to pursue it.

Okay. Chi-squared test for trend. This is probably one of the simpler techniques for testing for trend. It was a test statistic developed by Cochran and Armitage in 1954 and 1955. It's a simple test, and it's appropriate for testing the relationship between two ordinal variables. A single dichotomous variable maybe considered ordinal always.

The problem comes when you have nominal variables that you want to treat as ordinal. That's never appropriate. But if you're going from zero to one that can always be considered ordinal. When I say two ordinal variables, just remember that if you have a dichotomous variable, that can always be considered as ordinal.

Now, if you have three, like a treatment effect. Treatment one, treatment two, treatment three. There's nothing you can do there, because treatment two doesn't mean any more than treatment one. That's not on a continuous -- that's not an ordinal scale. However, if I go from treatment one to treatment two, I can consider that on an ordinal scale. That can be done.

Okay. Let's take a look at how I have the data set up. Cross-sectional right now. We'll come back to the cohort IMRs I'm just going to look at the cross-sectional at this point. Consider the cross-sectional IMR data and tabulate as you see here.

Once again, I could have considered the cohort data after going through all that linking process and everything, I had the number of births and I had the number of deaths for the cohort data. Easily could have done it. But this is for cross-sectional, what you see up here.

The question is, are the associated infant mortality rates monotonically increasing or decreasing?

Are they going up? We could easily calculate the IMRs themselves. That would be the one for 1990, 796 divided by 72,112 times 1000. That's the calculation for the IMR. This is how the data are set up for testing for trend. You see here.

We're going to set it up a little differently and I'm going to show how to modify that data a little bit to test, to perform this test in a different way. Okay, begin by calculating IMR I and generate a time trend plot. In a statistical analysis, this is probably the first thing you'd want to do.

This is essentially a repeat of what I showed you earlier. The only difference is I didn't do this by race. Okay. This is for all races. Okay. And I take a look at the time trend plot and I ask myself what do I see before I conduct the test.

I say there may be a slight decreasing trend. I don't know. You may look at that in another way. I don't really know. Okay, based on this. Looks to me like it might be a slight decrease, slight decreasing trend.

Okay. So I come here. I tabulate the data as you see here for the Cochran-Armitage test. Notice how this is set up. The one thing that you guys may not see right away is the number of deaths -- this is the same. The number of deaths in 1990. This is the same as what you see back here.

It's almost the same table. I'm working with the same data, 796. But notice this number has changed. It's not 72,112, it's 71,316. The reason it's like this is because when you're using the Cochran-Armitage test, the number dead is not going to be included in -- this is dead and alive.

Alive is not the number of live births. Alive is the number of infants who didn't die. Okay. So this 71,316 is going to be this one minus that one. Okay. So that's how Cochran-Armitage is set up. And that's how the statistic's calculated, okay, is like that.

So that would be the only thing you'd have to worry about when it comes to data setup. After that you're going to have -- now, I call this variable something else. I call it status, because the whole -- almost the whole way through this we're going

to be working with the number of live births, but I give this variable a different name because I don't want you guys to get confused with that.

This isn't the number of live births it's actually the number of live births minus the number of infant deaths. I give it a name so you don't confuse yourself. After that it's pretty straightforward to perform this test in SAS.

Once again what we're testing here is this hypothesis. Now I've hypothesized -- you see this greater than or equal to. If we have a decreasing trend, this is how we write the hypothesis.

If there's a decreasing trend, then IMR 1 at the beginning of the time period should be greater than IMR 2 and IMR 3 and so on. So this may look counterintuitive. This may be counterintuitive, but it isn't.

If there's a decreasing trend like I'm hypothesizing, the first one will be greater than the other ones. Okay. Here are the results. And we're going to go ahead -- we're going to do an in-class exercise for some of this stuff. But I'm just going to show you the results for now.

After running the Cochran -- after performing the Cochran-Armitage trend test in SAS here are the results I got -- notice the sample size. The way we've done this

here is we've done this at the individual level that's why you see the over a million observations there, at the individual level.

With that in mind, at the individual level, I'm going to reject the null hypothesis and both one-sided and two-sided, and I'm going to reject the null and I'm going to conclude that there is a slight decreasing trend.

I say slight because it looks to me like it's kind of slight. But this doesn't give you a measure of the magnitude of that trend. So there's really not much you can say about it. It's a hypothesis test for testing for significant trend. And it's one place where you might want to start. And it's an exploratory technique. It doesn't give you too much information.

Okay. Now, how do you interpret whether it's increasing or decreasing, how did I come up with that? Well, a positive Z value can be shown to be coincident with a decreasing trend. Okay. So if the Z value is positive, that means you have a decreasing trend. If the Z value is negative that means you have an increasing trend.

Now, you can probably -- most often you're going to be able to come here and say what that is. But that's probably not very good practice to do that. You should let the results tell you the results, not what you want to see.

Okay. So that's what I ended up with for that. Now, keep in mind that this was done at the individual level. This is essentially telling me that the likelihood of infant mortality for individual is decreasing over time. I'm going to show you guys another way to perform this test that's going to talk about the entire state.

Now, when I do it this way, keep in mind I have considered these populations separately. I talked about that earlier, considering race as the numerator/denominator combined, as a single rate or considering the deaths and underlying population separately.

Here what I've done is I've considered the deaths and the population separately, which is good. I've accounted for these population disparities sequences in that we're not going to be able to talk about population disparities. We're not going to properly account for those.

Now let me make sure I have that up here. Maybe I don't. Okay. I do. But I'm going to illustrate it with Kendall's Tau. Okay. Kendall's Tau -- so that's coming up. Keep that in mind, I'm going to consider these as a rate later.

Kendall's Tau, we have two cases. Chi-squared test for trend is going to be done like this. We're going to do the chi-squared test for trend like this, although we could do a chi-squared test for trend using the rates. I didn't do that here. I did it

for Kendall's Tau. So what you need to be asking yourself is you want to perform some type of test for trend.

And here's what you're given. If you have it set up like this, you could set it up as you see here or set it up in the form of a rate. If you set it up in the form of a rate you lose the ability to talk about the underlying population. You didn't account for the underlying populations properly.

If you perform it like this, you have. However, if you talk about it like this, you're talking about it at the individual level. If you talk about the rates, you're talking about the state level. So keep that in mind. I'm going to illustrate a case where I do look at the rates but not for that test. Ordinal measures of association. Now, this is different. Chi-squared test for trend simply tests the hypothesis of monotonicity. That's it. It raises the questions what is the association, how strong is it? We weren't able to answer that with chi-squared test for trend.

Now we're going to take a look at a measure of association that allows us to analyze that or measure that. Kendall's Tau-b, Tau-beta, is a nonparametric measure of ordinal association. And it ranges over negative one-to-one for nonsquared tables. Think about this as like an analog to R square, to just R, not R squared, just R. Pearson's correlation coefficient. This is an analog parametric to that.

This statistic ranges over negative one-to-one for nonsquare tables and most of the tables we're going to look at today are going to be nonsquare. Think about the births status, birth, alive, dead, and many years. So that's a nonsquare table.

What I'm talking about, nonsquare, I'm talking about this. The table's set up in this fashion and that's nonsquare. You can use this for square tables. It's not a big deal. It's going to change the end points. Negative one and negative one will be able to be included. If you want to use it as square tables you can, but this is the range for nonsquare.

Now, each extreme reflects a significant or a -- I can't say significance yet. We have to perform a special test for that. But I can say a strong monotonicity. Just as if you go back to Pearson's R, Pearson's correlation coefficient, each extreme reflects a strong linear correlation.

For this, each extreme is going to reflect a strong monotonic relationship. So it's interpreted a lot like that. Positive numbers are going to indicate monotonically increasingly bivariate relationships, and negative numbers are going to indicate monotonically decreasing bivariate relationships.

Let's look at some data. Once again data are tabulated the same way. I just copied the table I showed you before and pasted it here. It's the same table.

Cross-sectional IMR data. We have two cases. Now this one I'll illustrate for both rates and individual level stuff. I talked about individual level stuff before.

Okay. Individual level stuff, once again, we're going to interpret as the likelihood of an infant death for an individual. You're talking about this from an individual's perspective. Now we're going to also talk about it at the state level from its state perspective, too.

Consider the individual level case like we did before, the data set up like this, very common way to do this, and here are the results. Now, what I have here, I have a confidence interval. But more importantly, well, not more importantly, we have to look at the significance of this confidence interval, too, and we see it's significant. You shouldn't be surprised to see significance here, because we have a large sample of individuals.

So the significance isn't that surprising. You can have many cases where the significance is where something's very significant, but a large sample will almost always give you that.

So you want to consider the magnitude of the significance. If something's significant but it only contributes negative point triple 01 to something then it might not be that important.

In this case I ended up with a Tau-beta. Now it looks like it's significant here. My confidence interval doesn't include 0. So I can say that it's significant. But take a look at the value. Negative .0026. The magnitude of that number isn't large at all. Now, I guess I can say that there's a slight decreasing likelihood of infant mortality of a single individual over time. The number is negative. You can interpret that as negative, negative being going down.

But it's not very big. So this is saying there's some there but it's not -- the magnitude of the number isn't very much. But it is significant. It's not surprising, though, because we have over a million live births.

Okay. Now I'm going to show you something different. I considered these data as you see them here death is -- status is either dead or alive, meaning I've accounted for these populations. These alives are essentially the number of live births.

Now, this is how the test is conducted. But what we're essentially doing here is accounting for the populations. That's just simply how the test statistic needs to be calculated based on the alives. But you guys can see that the underlying population here is simply this plus this.

So we've accounted for it. Next what I'm going to do is I'm going to include these as only rates. I'm going to combine those two pieces of information into one. I

lose all ability to discuss what's going on between the denominators. Meaning, if I back up again, if I had in 2005 if I had 598 -- well, let's bring this number down.

Let's say I had one infant death and I had a population of one. Well, I mean, you know we're going to consider that, if we were to do that, we would consider this as having the same population as that.

Probably not a very good idea. Right? This is an unstable rate. This is a very stable rate. So that's what happens when you consider these as rates. That's what I'm going to do here is a single number. I'm going to run the same test, the Kendall's Tau beta with doing this on the rates, as you see here. And I can do this.

You guys will be able to do this easily. It's not very much different at all in SAS. Here are the results you see here. Okay. I have -- it looks like -- now I'm not interested in what you see going down the diagonal. Okay. Now, notice the output is indicated a little bit differently than when you do it like this.

You use the -- you set up the data like this and you run a Kendall's Tau on it. This is the kind of output you get on it. You set it up like this. You run a Kendall's Tau. This is the kind of output you're going to get. Same measure, just comes out a little differently. I think I ran the same proc core. Might have not on the other one.

Anyhow they're interpreted the same way. Let's look at the measure. The measure is negative .3. It's given right here.

So this says if it was significant, this would say I have a decreasing trend over time. Now, I look right underneath it, the P value, and it's insignificant. So it's saying I don't have a significant trend based on these IMRs as you see them there.

Okay. So that's what that means. Now, this is a very basic test. These are two very basic tests. One thing is significance. The other thing is ordinal measures of association. Where did all this start. Well, one of the places it started was with that chi-squared test for trend. So you guys go back and maybe read some of Cochran and Armitage's stuff to read some of the motivation for this stuff. Those are two very classical tests.

Any questions about that? Those are just some basic tests that you guys might want to perform at the beginning of a study like this.

UNKNOWN SPEAKER: How would you release this [Inaudible] to those two tests? The individual list.

JEFFREY SHAFFER: That's a good question.

UNKNOWN SPEAKER: [Inaudible].

JEFFREY SHAFFER: Yeah. Yeah. You would say that there is a significant -- now it doesn't seem to be consistent. At the state level you would say there is no evidence of a decreasing IMR. At the individual level, after accounting for the population disparities between the years, there is a very slight difference. But the reason these are coming up like that, why do you think this significance level, why do you think that's gone? What's our sample here?

UNKNOWN SPEAKER: The population is getting smaller.

JEFFREY SHAFFER: Well, you almost have it. We have a sample size here of 16 years. Our sample here was over a million. We're talking about individuals. So we have over a million observations going into this test statistic. Was it this one? Yeah. We have over a million here. A million people. But we have only 16 years.

So think of the sample as being 16 and 1 and over a million here. That's the significance. So the reason you're seeing that is because of those. I think what's more important to see is when I come and look at the magnitude of the number for the significant one, it's very small.

I suspect if I had a large enough sample, if I had a million of these, that number would be significant, too. So that's the reason you're losing the significance. It's likely to happen by chance.

UNKNOWN SPEAKER: I have a question. We flipped back to the slide that has the infant mortality rate that you were looking at the trends, back a little bit more.

JEFFREY SHAFFER: Tell me when to stop. We go here. Here.

UNKNOWN SPEAKER: That one. Is there a minimum number of years in which you have confidence about modeling a trend? Because if you just look at that graph and you picked out '90 to '95 you would come up with something much different than if you looked at 2000 through 2005.

JEFFREY SHAFFER: Any confidence at all. The only place that's going to show up is in your standard errors. It depends how the standard errors are calculated in the technique I guess is what I would say. Now, you also have to use some common sense like you're doing there.

Does it make sense to even talk about three points or five points? I say this. We don't have that many years here. And that is a common problem with data like this, is they haven't gone back that far.

And sometimes even when they have gone back that far, the reporting issues were different and you can't put them all together like you see here. So that is a good question in that essentially our sample of years is 16.

Okay. So that isn't very much. And different techniques have different thresholds for what constitutes the minimum required sample. For the arena techniques that we'll be looking at tomorrow, the time series techniques, they give an exact way to say whether you should use the technique or not.

But I guess I would -- I guess my answer to that is it depends on the technique. And it also depends -- because there's going to be a different standard error associated with each technique you use.

We're going to look later at a technique, typically the nonparametric techniques do better with smaller samples. That's one of the main uses for nonparametric statistics.

UNKNOWN SPEAKER: I would guess that's a large sample.

JEFFREY SHAFFER: Depends on what you're modeling.

UNKNOWN SPEAKER: I'm guilty of looking at more recent data and figuring it out. Because people we're talking, they don't care what happened in '95 or '90, because it's not relevant to --

JEFFREY SHAFFER: You're looking at this. If you're looking at -- we have both large and small samples here. If you're looking at this, if you ask me what is -- how confident are you in the IMR for 1990 based on these data, I'm very confident in that.

And that would come up in a simple confidence interval for an IMR. Okay. Because I have a huge population. I'm very confident in that. I have a population of over 71,000 live births. However, so for 1990 specific, I'm very confident.

When you come here, I'm actually very confident because we're including all of the denominators here. If I consider these as individuals, I'm very confident in those results. But when you come here, as soon as you start doing this and you say, okay, I'm going to move from considering populations, considering the infant births to rates, your sample is no longer over a million.

That's gone. Your sample is now 16. You have 16 years. So one of the main arguments for not modeling rates as continuous data is you're essentially losing a lot of your sample. Is that clear? They don't have risk ratio interpretations, because you're assuming your populations are the same over all years.

That's saying that this population is exactly the same as that. Now, it may be close for these data, but for other people's data, maybe not. Okay. I'm saying they're all the same. And I have exactly 16 values. Is it done like this all the time? Sure. But keep in mind you're dealing with a sample of 16 when you do that.

We're going to look at some techniques that actually deal with small samples, too. We're going to look at some nonparametric regression techniques that are a nice -- an effective way for handling that.

One thing you need to consider when you're interested in modeling these is what type of technique to use. What type of technique's appropriate. Well, I would start off with the one that requires the least assumptions, okay, about your response surface.

So that's coming up later. But the thing I want you guys to see here is, yes, the sample all of a sudden is 16, and we have all these people, a million, over a million people, but the sample size, when you look at the rates, it's gone. 16.

All right. So that's probably why that P value is insignificant at this point. All right. So let's go ahead and look next at some basic descriptive measures for trend. And the answer to your question, I don't know if I have it here. Per annum. Yeah. I have some stuff coming up on what you talked about. But as far as his question

goes, results arising from a bad sample or a small sample, not necessarily a bad sample, a small sample, are going to show up in your standard errors.

And you're going to look at those standard errors and look at your confidence intervals. And if they are so wide where they just don't make any sense at all or cannot be used, I mean if you have a confidence interval about glucose levels that ranges from, let's say, the mean glucose level is between let's say 100 to a thousand milligrams per deciliter, I mean how useful is that?
You have to ask yourself, is it useful? And that's where the small sample is going to show up in all of the analytical techniques. Now, a lot of the analytical techniques will give you that from the beginning. They'll say, well, based on our previous studies you shouldn't use this amount. And I see that in a lot of literature and a lot of the reports. They'll say this is considered a small sample.

But it depends on the technique. All right. So let me back up to what we have here. Okay. So more simple basic descriptive stuff. Now you may not have seen this before, but it's basic in the sense that -- oh, I went the whole way down -- it's basic in the sense that we're not controlling for any covariates. We're not really -- this is purely exploratory. It doesn't have to be exploratory.

Okay. Let's talk about annual percent changes. These are very commonly reported in the literature for health outcomes. This is a simple measure used to

describe time trends. And for the simplest case, assume that the IMR trend, the trend in IMRs is constant over time.

Order the IMRs as IMR I equals 1 to K where 1 and K are respectively correspond to the first and most recent observation times. Just think about that as notation for ordering the IMRs in terms of their time, their associated times, meaning the IMR from 1990 is here. Next to it IMR for 1991 and so on.

Put them in order. Order them in ascending order of year. Okay. And here's how you calculate an average annual or not an average. We're going to talk about that next. Just an annual percent change. There's a difference between average annual percent change and just annual percent change or APC.

For just an annual percent change, you take the difference of the range of the IMR -- of the end points of the IMR, not the range, the end points, meaning the IMR for the last year is here. For our study, it would be IMR 2005 minus the IMR in the first year, which would be 1990. Divide that by IMR for the first year times K minus 1. Times the number of years minus 1. So for our study this number would be 14.

Multiple it by 100 and that's called an annual percent change in IMR. And this is one thing that you guys would probably want to report. We're going to talk about significance levels for IMRs and things like this later.

But this is what an IMR is. This is in the most basic sense this is what an IMR is or not an IMR, an annual percent change. I'm going to show you guys how to calculate an annual percent change using regression techniques as well. And this allows for a very straightforward interpretation and explanation of your results.

You can simply say that based on a calculation like this that the IMR increases at an annual rate of something. That's a very straightforward interpretation. So that has some appeal.

That's an annual percent change. Here's an example, okay. Using the cross-sectional data for the Louisiana births and deaths data that I showed you earlier, the IMR for the first year, okay, or 1990, if you like, one is 1990. That was our 16. We had 16 years. Either way you want to look at it, is 10.006. So I perform this calculation. I substitute these numbers in, and I end up with negative .59 percent per annum.

Okay. So that's saying that my IMR is decreasing at .59 percent per annum. Okay. So now this is probably -- this may have some appeal if you only are given the end points or you're given the range. If you're given more than that, use it. And I'm going to show you how to use it next. But this only considers the range. This says is my IMR going up or down over the entire range. Now, there's a lot

going on within that range. So we haven't done anything with that yet. Is this appropriate? Well, it goes back to your plots. Can you say that the end points adequately describe the pattern of what's going on within the range? Then this is appropriate if it does.

A more effective approach would be to use what's called an average annual percent change. Now, why would you ever use this if you have a more effective approach? Well, you may only be given the end points. If you're only given the end points, this is all you have. You don't know what's going on between the end points. Maybe you have an idea but you don't have the numbers.

Okay. So if you do have the numbers, more appropriately I should say within the range of the years. There's a difference between the range of the years and the range of the IMRs.

We order these IMRs in terms of year. Within the range of the years. Within the range of the study period, I guess, is what I should have said.

If a constant trend over time is implausible, which oftentimes it isn't, meaning that what's going on at the end points adequately describes what's going on in the middle, which it usually doesn't, a better calculation, a more precise estimate is what you see here. This is called an average annual percent change.

Okay. And this involves -- this is essentially an average of all the IMRs from one year to the next. And once again the subscript corresponds to the year. Year I. So you would just calculate this sum.

Basically just a plug-in. I think it will probably come into clear focus once I show you guys the example. But this is just evaluating a sum and plugging the numbers in. K is the total number of years for the time periods you're considering.

Let's look at an example of this. Given these data -- so I just give you the IMRs. 1990 to 2005. Let's take a look and see how we evaluate the sum that I showed you here.

Okay. The first thing we do is we see that we have 16 years. So the denominator is I minus 1. Let's plug a 2 in here because that's where it starts. So this would be the second IMR for the second year. IMR 2. So if I move ahead, IMR 2 is 10.46, second year.

Divided by I minus 1. Plug 2 in here that's the IMR or the previous IMR. What you guys need to do when you see a notation like this just put it in your own words instantly. If you try to remember this by brute force, it will never happen. Put it in your own words. I just think IMR current year. IMR previous year, minus 1. And then you'll have it forever.

For the previous year minus 1, evaluate that sum. Keep going, do that for all of them, and times 100. I end up here with negative .42 percent per annum, which isn't much. But it's not even a whole percentage point. So that's what I'm seeing. This is a more precise estimate. Okay. Next I'm going to talk a little bit about rate reliability.

We've been seeing these numbers and we haven't been talking about how reliable are they. Let's go ahead and take a look. I've adopted some of this stuff from some of the reports that I've seen out there, the linked -- I can't remember exactly where I got this from. But I know I've reported it in your manual where I picked this up from. It was one of the infant mortality reports out there. And I'm sure I noted that in my manual about where to get additional information about what you see up here.

Next I'm going to talk about rate reliability. One way of measuring rate reliability is to talk about this in terms of a relative standard error. Okay. Relative standard error is often known as a coefficient of variation. They're the same thing. The infant mortality reports, they refer to it as an RSE. I usually refer to it as a coefficient of variation. In the statistical books they seem to refer to it as coefficient of variation, but they're the same.

This is how it's defined. RSE IMR, the notation, the RSE of this argument. This means RSE IMR-I, you would read this as the IMR for the Ith year, if we wanted

to relative standard error for IMR 1990, put in 1990 as the subscript. Equals 100 times the standard error of the estimate divided by the estimate.

Okay. Estimate is some estimate. Something you're estimating. Could be a parameter estimate. It could be something like we just discussed here. The APC divided by the standard error of that, if we had it. The standard error of that.

Sometimes the standard error is hard to come up with. Sometimes the formulas are this. In a uniform -- we would be able to present this in a way that everyone could relate to.

Okay. Small RSEs are coincident with reliable estimates. That should make sense to you. Because you want the numerator to be small. You don't want the standard error to be large. You want it to be very small. So obviously small is better. Small numerators mean small number is better.

The NCHS, National Center of Health Statistics and BRSS, you may have used that before, it's a national -- statewide survey performed annually.

They recommend, though, they use a threshold of 30 percent. The NVSS says that 23 percent is unreliable. This can be shown in an infant mortality context to be approximately equivalent to 20 deaths per thousand live births. So that's

saying if I don't have at least 20 deaths per thousand live births, then I'm probably going to have an unreliable estimate for IMR.

An unreliable IMR, that's how I can determine if it's a reliable IMR or not. Now, a lot of you are probably dealing with numbers much larger than that. But this is saying if you go below 20 and you may if you go to a small enough level. At the state level probably not.

If you go to a local level, and you start seeing numbers around 20, this is saying that maybe you shouldn't do this, because they recommend you need a threshold for what's reliable and what isn't. And they're saying less than 20 percent or 23 percent or more is unreliable.

I don't know. One of these two. As an example, and then we're getting pretty close to lunch, I think we'll go ahead and break after this. As an example of rate reliability. So how reliable is one of the IMRs that we looked at? I say here of the 72,000 live births for Louisiana in 1990, there were 796 infant deaths. The relative standard error for the Louisiana IMR in 1990 is what you see here.

Where did I get this number? You're going to have to look for my manual. I can tell you this: That this is going to be equivalent to the RSE for an infant mortality rate. But I will say that the standard error of this estimate is based on a Poisson distribution. So we're interested in this and the estimate itself.

So it can be shown that the RSE for this is going to be 3.56. This number depends on the technique you're using. If somebody else asks you what's the RSE of this estimate based on linear regression. It won't be the same as this. This is based on a Poisson distribution and considering the IMR's numerator and denominator separately.

So this is what that's based on. Now, ending up with this, I ended up with 3.56. Now, that's very reliable. They're using a threshold of 23 percent. That's extremely reliable. Shouldn't surprise you, because that's an enormous sample. I have 796 deaths. By the way, the RSEs can always be calculated like this for IMRs. The general form is going to be one over the number of deaths, the square root over the number of deaths plus one over the number of births.

But I have the general form in the notes, but I will say, like I said, if you have a different technique, there's going to be a different form for standard error. So this is the general form for RSE, for any technique you use. And this is how you would interpret it.

Now, we conclude that the estimated IMR for Louisiana in 1990 is extremely reliable. But like I said, you shouldn't be surprised because it can be shown that 23 percent -- an RSE of 23 percent for IMRs is approximately equivalent to about 20 deaths per thousand infants.

Here I have much more than 20 deaths per thousand infants. I have 796 per 72,112. So that's pretty good.

All right. Let me see if this breaks off nicely. Yeah, I think we're going to start here when we come back from lunch and we're running a little behind today, but we'll pick it up a little bit in the afternoon and tomorrow as well. Let's go ahead. We're ready for lunch, right?

UNKNOWN SPEAKER: Starved.

JEFFREY SHAFFER: Are they ready for us?

UNKNOWN SPEAKER: They are ready.

JEFFREY SHAFFER: Thank you. We're going to go ahead and break for lunch. Let's go ahead and we're going to come back at 1:00 p.m.