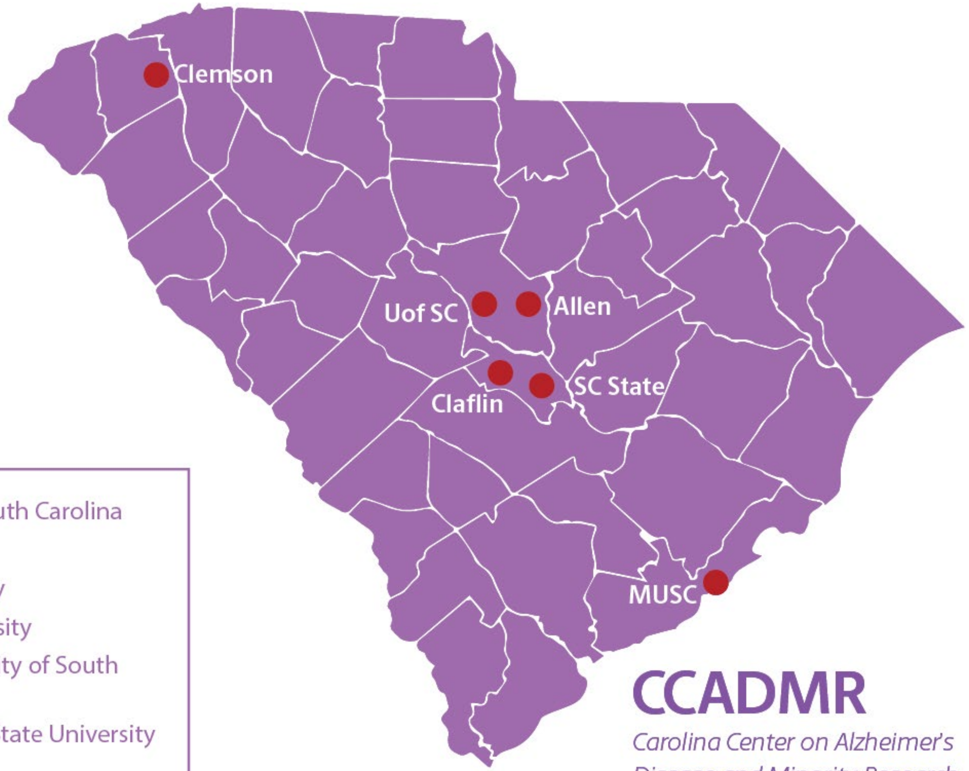
Session
will begin soon

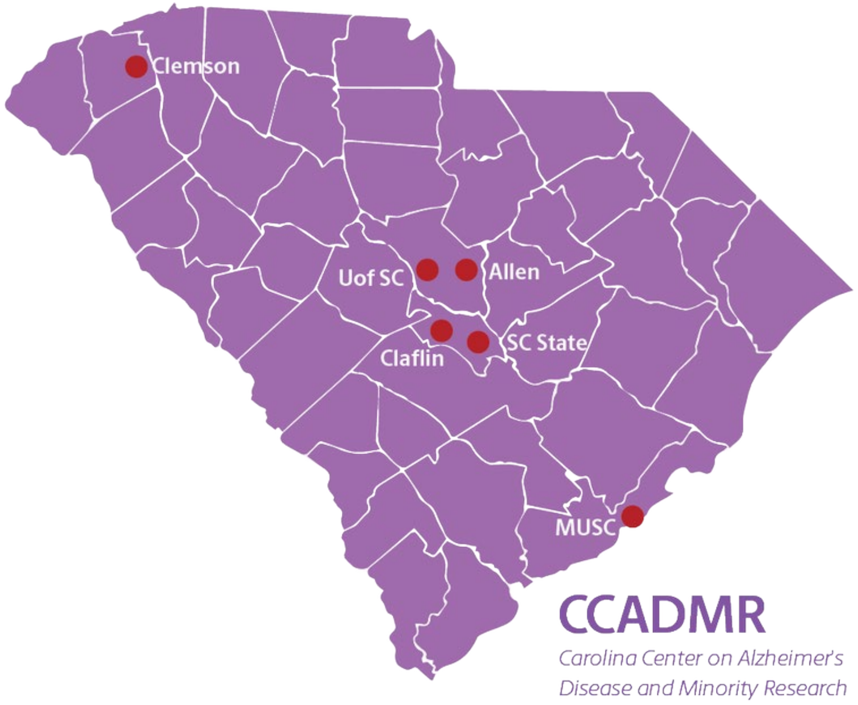
Health Disparities in Minority Aging Research Seminar Series



University Partners

- University of South Carolina
- Allen University
- Claflin University
- Clemson University
- Medical University of South Carolina
- South Carolina State University

CCADMR
Carolina Center on Alzheimer's Disease and Minority Research


Clemson
Uof SC
Allen
Claflin
SC State
MUSC
CCADMR
Carolina Center on Alzheimer's Disease and Minority Research

How to Write a Data Analysis Section

James Hardin, PhD

January 24, 2020

TODAY'S SPEAKER

CCADMR

Carolina Center on Alzheimer's Disease and Minority Research

James Hardin, Ph.D., Professor and Associate Dean of Faculty Affairs and Curriculum, **University of South Carolina**

James W. Hardin

University of South Carolina

How to …..
write a statistics section

There are actually 2 ways to write a statistics section:

The hard way

The other hard way

What do we mean when we say the statistics section?

In our discussion, we will focus on manuscripts and grant proposals

Grant proposals: abstract (specific aims), methods (power and sample size)

Manuscripts: introduction (hypotheses), methods, results, discussion

Think about the adage for giving a speech.

Some straightforward rules:

1. Your study has a purpose. That is, you have a question of interest that you are going to evaluate. The related description goes into your abstract or introduction.

2. The statistical methods that you will apply to the data are described in the methods section.

3. The conclusions that you reach are described in the results section.

4. Because there are multiple sections, you CANNOT work on these things separately!

Let’s imagine having to communicate with another person about producing the statistics jargon that we need. To highlight the manner in which we will approach our conversations, let’s consider a specific study.

Throughout our discussion, Let’s imagine that we have a study of the respiratory distress of children and that one of the covariates of interest in our data is an indicator of whether the child’s mother smokes. We have multiple measures per child.

In our abstract/introduction, we will explain

What is our hypothesis?

In our methods section, we need to explain:

What **model** we will estimate?

What **result** we will interpret?

Hypothesis: The likelihood of respiratory distress will be higher for children whose mothers smoke.

What kinds of models can we consider?

We will investigate the association of mother's smoking on the likelihood of a child having respiratory disease using a generalized estimating equation (GEE) model

$$y_{it} = g^{-1}(\beta_0 + smoke_{it}\beta_1 + Z_{it}\gamma)$$

This is known as a population-averaged model for which the correlation of within person observations are hypothesized to follow a specified structure.

Our logistic regression GEE model is specified by

$$y_{it} = g^{-1}(\beta_0 + smoke_{it}\beta_1 + Z_{it}\gamma)$$

where g^{-1} is the inverse link function relating the expected value of the outcome to the linear predictor (we will use the logit link function and interpret exponentiated coefficients as odds ratios), y_{it} is an indicator of whether the ith child at the tth visit suffered respiratory distress, $smoke_{it}$ is an indicator of whether the mother of the ith child is a smoker at time t, Z_{it} is a vector of covariates associated with the ith child at time t (sex of the child, whether the child has allergies, average daily temperature at time of measurement, race of the child, socioeconomic status of the child, and whether the child has comorbidities – see discussion section), and γ is a vector of coefficients associated with Z. Our primary interest is in evaluating $H_0: \beta_1 = 0$. While our model assumes exchangeable within-child correlation, the inference for this particular test will be based on empirical standard errors and so will be robust to misspecification of the within-child correlation of repeated observations.

Specifying your model:

Be concise: You do not need to specify every variable that is going to be included in the model:

$$y_{it} = g^{-1}(\beta_0 + smoke_{it}\beta_1 + male_i\beta_2 + allerg_i\beta_3 + temp_{it}\beta_4$$

$$+race_i\beta_5 + ses_i\beta_6 + illness_{it}\beta_7)$$

$$y_{it} = g^{-1}(\beta_0 + smoke_{it}\beta_1 + Z_{it}\gamma)$$

Specifying your model:

Explain everything that is in the equation – that is, explain everything that will be included in the model.

Explain how your model will help you answer your hypothesis of interest.

Explain whether your model makes any assumptions and whether those assumptions could affect your inference.

In a population averaged model, the interpretation of the odds ratio for smoking reflects the increased odds of respiratory distress of *children of smoking mothers* versus *children of nonsmoking mothers*.

What kinds of models can we consider?

We will investigate the association of mother's smoking on the likelihood of a child having respiratory disease using a mixed-effects regression model

$$y_{it} = g^{-1}(\beta_0 + smoke_{it}\beta_1 + Z_{it}\gamma + \delta_i)$$

This is known as a subject-specific model for which we include variance components at each structural level of the data.

Our logistic regression mixed-effect regression model is specified by

$$y_{it} = g^{-1}(\beta_0 + smoke_{it}\beta_1 + Z_{it}\gamma + \delta_i)$$

where g^{-1} is the inverse link function relating the expected value of the outcome to the linear predictor (we will use the logit link function and interpret exponentiated coefficients as odds ratios), y_{it} is an indicator of whether the ith child at the tth visit suffered respiratory distress, $smoke_{it}$ is an indicator of whether the mother of the ith child is a smoker at time t, Z_{it} is a vector of covariates associated with the ith child at time t (sex of the child, whether the child has allergies, average daily temperature at time of measurement, race of the child, socioeconomic status of the child, and whether the child has comorbidities – see discussion section), γ is a vector of coefficients associated with Z, and δ_i is the child-level variance component (random effect). Our primary interest is in evaluating $H_0: \beta_1 = 0$. Our model assumes normally distributed child-level random effects such the loglikelihood will be evaluated using numeric integration. Sensitivity analyses will investigate whether the child-level random effect is significant, and whether the distributional assumption of that effect is reasonable.

In a subject-specific model, the interpretation of the odds ratio for smoking reflects the increased odds of respiratory distress of a child if their non-smoking mother took up smoking.

Let’s assume that the estimated odds ratio for either model is 2.0 for the smoking indicator variable and that the associated p-value supports rejecting the null hypothesis.

Which model do we want?

Is there a wrong model choice?

What are the assumptions?

What happens if my data violate those assumptions?

How do I interpret a significant model coefficient?

RESULTS

Population-averaged GEE model:

We found a significant association between the smoking status of mothers and the likelihood of respiratory illness of their children. Specifically, compared to children of non-smoking mothers, children of smoking mothers have 2.0 times the odds of respiratory illness holding all other covariates constant.

Subject-specific mixed-effects model:

We found a significant association between the smoking status of mothers and the likelihood of respiratory illness of their children. Specifically, if a non-smoking mother were to take up smoking, then her child would then suffer 2.0 times their former odds of respiratory distress holding all other covariates constant.

Changes? Is there an alternative interpretation of the results?

Subject-specific mixed-effects model:

We found a significant association between the smoking status of mothers and the likelihood of respiratory illness of their children. Specifically, if a non-smoking mother were to take up smoking, then her child would then suffer 2.0 times their former odds of respiratory distress holding all other covariates constant.

We found a significant association between the smoking status of mothers and the likelihood of respiratory illness of their children. Specifically, if a smoking mother were to quit smoking, then her child would then cut their odds of respiratory distress in half holding all other covariates constant.

Implications

You should change your introduction to emphasize your investigation of the health benefits of quitting smoking rather than the dangers of smoking.

You should change your model specification to include an indicator of non-smoking instead of an indicator of smoking.

Other Topics

If you are working on the statistics section of a grant proposal, you may also need to address power and sample size.

While this is another topic that probably needs its own hour for presentation, let's look at some highlights for that section.

Power and Sample Size

For the purposes of this sample size calculation, the standard deviation (SD) of the difference between the XXXXX® and Vehicle sides was assumed to be 4.0 for the POSAS and 1.0 for the SCAR. Using these parameters, the sample size requirement of 184 was calculated for 80% power to detect a standardized difference between 0.292 (secondary) and 0.321 (primary) allowing for 20% attrition (n=146). The level of significance for the primary endpoint was set at 2.5% (Bonferroni correction since there are 2 subscales; 1 for patient and 1 for observer), and the level of significance was set at 5% for the secondary outcomes. The study is powered to detect a difference that is between a small (0.200) and a medium size (0.500) standardized effect . Number of subjects reflects 0% attrition (n=184), 10% attrition (n=164), and 20% attrition (n=146). Each subject will contribute two observations – one for each treatment, and the table illustrates the detectable standardized effect size at 80% power.

Power and Sample Size

Numeric Results for Two-Sample T-Test
Null Hypothesis: Mean0 = Mean1   Alternative Hypothesis: Mean0 ≠ Mean1
Unknown standard deviation.

				Detectable
Power	**N**	**Alpha**	**Beta**	**Effect Size**
0.80000	146	0.02500	0.20000	0.361
0.80000	164	0.02500	0.20000	0.341
0.80000	184	0.02500	0.20000	0.321
0.80000	146	0.05000	0.20000	0.328
0.80000	164	0.05000	0.20000	0.309
0.80000	184	0.05000	0.20000	0.292

Preliminary data from other studies found standardized effect sizes that are close to medium size (0.500). Using data from both cohorts similar standardized differences at standard levels of power and adjusted levels of significance can be detected.
Cohen, J. A Power Primer. 1992. *Psychological Bulletin* 112 (1), 155-159.

Highlights

Often, it is more meaningful (to a wider range of backgrounds of reviewers) to discuss standardized differences instead of absolute differences.

It is never a good idea to only report one number. Give a range of detectable effects for a power, or a range of power values for a detectable effect.

It is also often a good idea to point out the power for different sample sizes that reflect different amounts of missingness/unable to enroll/dropouts.

Summary

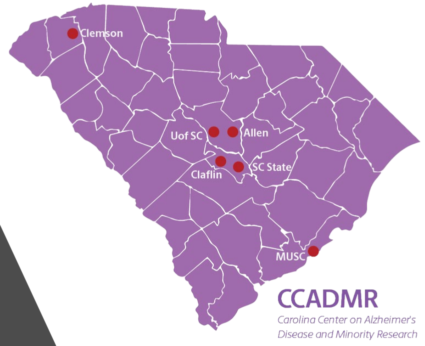
The sooner you provide context to the consultant, then sooner their verbiage can incorporate that verbiage.

It almost never works out well to completely separate responsibilities for different parts of the manuscript/grant proposal.

Engaging your consultant to prepare mock results will help all parts of your document.

Imitation being the sincerest form of flattery, basing a draft on an existing successful model of what you want to prepare will go a long way.

TODAY'S SPEAKER


Clemson
Uof SC
Allen
Claflin
SC State
MUSC
CCADMR
Carolina Center on Alzheimer's Disease and Minority Research

James Hardin, Ph.D., Professor and Associate Dean of Faculty Affairs and Curriculum, **University of South Carolina**

- James Hardin is a professor of biostatistics. He has been with the University of South Carolina since 2003. His research interests include correlated data, limited dependent variables, discrete choice analysis, generalized estimating equations, and complex survey data.

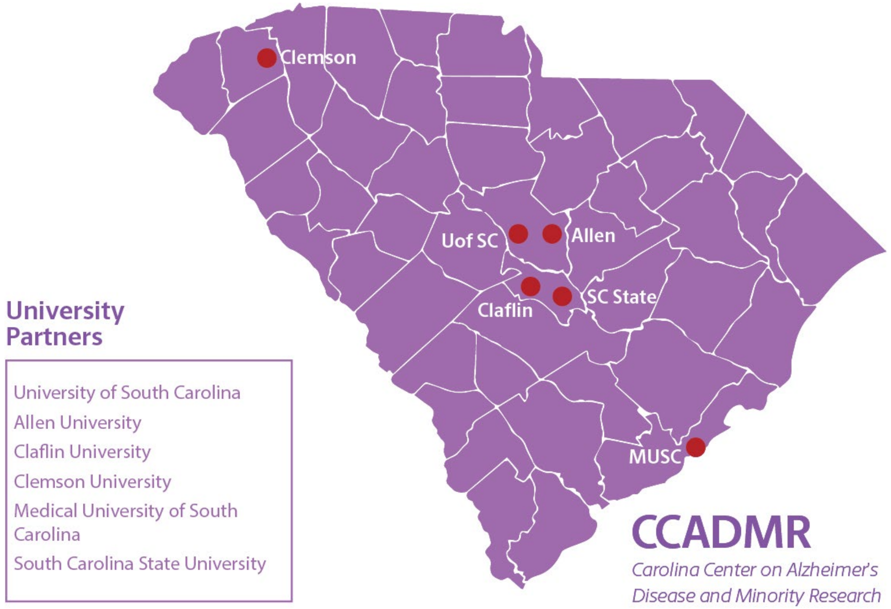
Thank you for participating!

Please give us your feedback about the session by answering a brief survey.

For the ***in-person attendees***, we will have the survey available on the tablets at the end of the session, or you can use the QR code on screen to access the survey.

For ***virtual attendees***, we will be emailing a survey link to all participants, you can access it through the QR code to the right or through the survey link.

The QR code appears here or it can be accessed via the Survey Link.



University Partners

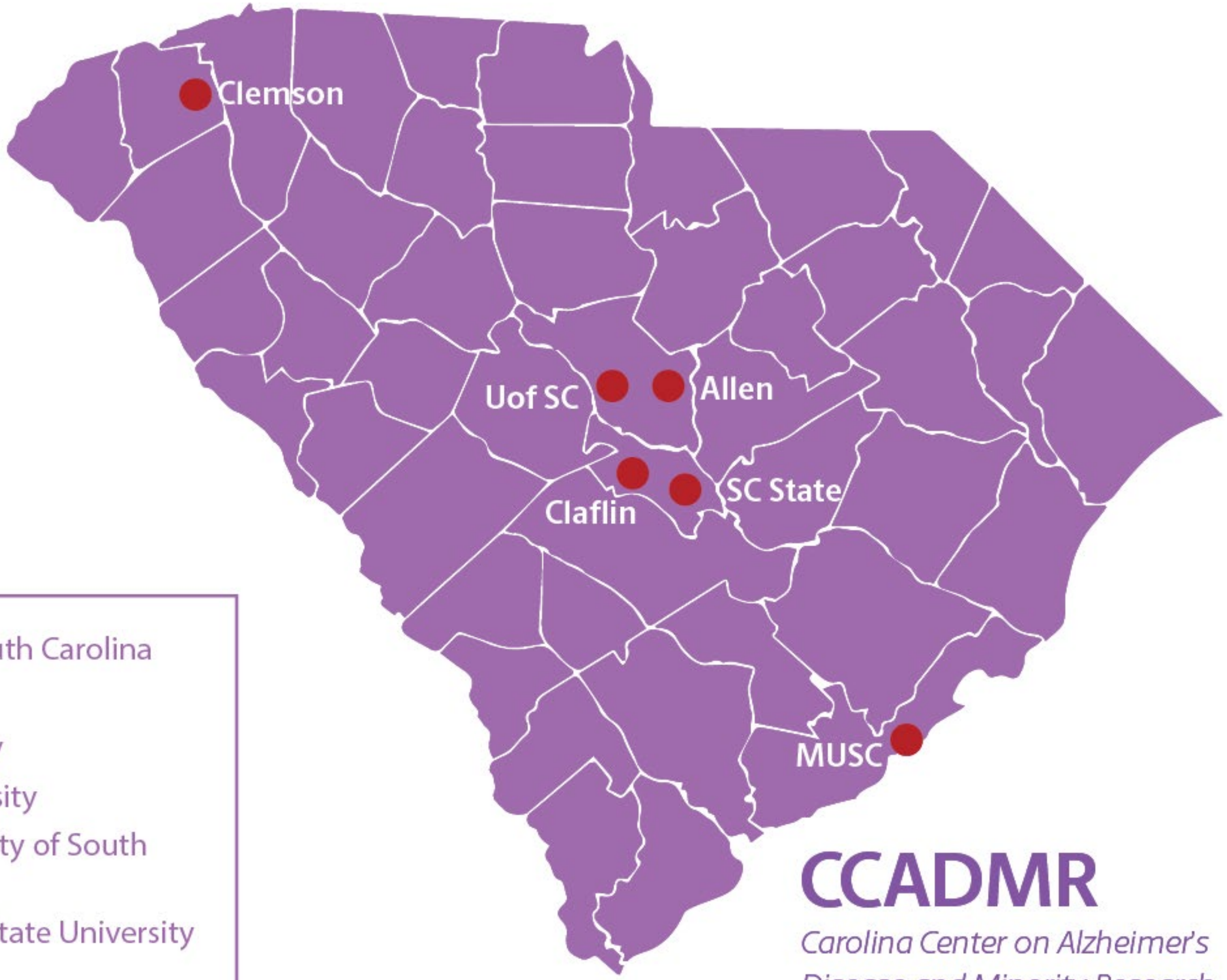
- University of South Carolina
- Allen University
- Claflin University
- Clemson University
- Medical University of South Carolina
- South Carolina State University

CCADMR

Carolina Center on Alzheimer's Disease and Minority Research

Thank you!

If you have any questions, please contact
Dr. Lucy Ingram,
Lannang@sc.edu.



University Partners

- University of South Carolina
- Allen University
- Claflin University
- Clemson University
- Medical University of South Carolina
- South Carolina State University

CCADMR

Carolina Center on Alzheimer's Disease and Minority Research