

STATISTICS AND SAMPLE SIZES IN GRANT/ETHICS APPLICATIONS:

WHAT WOULD A STATISTICAL REVIEWER LIKE TO SEE?

WARNING

This presentation is about what a **STATISTICAL** reviewer would like to see in a grant/ethics application – you should not let this over-ride your experience with what reviewers in your field want.

WHAT WE WILL LEARN TODAY

- This presentation will give some statistical background and some “cheat sheets” for presenting your project proposals:
 - The statistician’s view, and some basic dos and don’ts;
 - An overview of the statistical analysis process, and the elements of what you need to explain (in a logical order);
 - An overview of how sample size calculations work, and the elements of what you need to explain (in a logical order).

The background is a dark blue-grey color. In the four corners, there are decorative white line-art patterns that resemble circuit traces or data paths, with small circles at the end of the lines. The text is centered in the middle of the page.

THE STATISTICIAN'S VIEW OF GRANT/ETHICS APPLICATIONS

THE STATISTICIAN'S VIEW

- Grant and ethics applications for research projects involving data generally include the following statistical information:
 - Description of your **methodology** and experimental design;
 - Description of the **data** you will use in your analysis;
 - Description of your proposed **statistical analysis/modelling**; and
 - Your **sample size** calculations/justification.

THE STATISTICIAN'S VIEW

- Naturally, your statistical reviewer will be brilliant. But sadly, he/she cannot read your mind — the reviewer will read only what you write in your application, in the order you write it.
- Your statistical reviewer is probably also a suspicious (bordering on cynical) person who has seen plenty of bad analysis before. He/she will be alert for certain “red flags”.

THE STATISTICIAN'S VIEW

- Your statistical reviewer will be familiar with a wide class of statistical models, and their underlying theory, but this does not remove the burden of explaining things simply and clearly.
- Your statistical reviewer is unlikely to be impressed by statistical complexity or jargon — they will be much more impressed with *clarity, soundness, and simple exposition.*

THE STATISTICIAN'S VIEW

- Statisticians fancy ourselves as general scientific experts who get to “play in everyone’s backyard”. So we want to be able to read your proposal, and be confident we could start work on it.
- We will consider whether we can do this project from your description of it. If we can’t, we will get cranky, and blame your bad description (and we are probably right to do so).

THE STATISTICIAN'S VIEW

- If your description is done well, your statistical reviewer will feel like they can “see” the data and model. This should be **easy** for the reviewer — they should not have to search around.
- The reviewer should be able to write out your (empty) data frames, and your proposed model, including specifying the input and output variables.

CHEAT SHEET – METHODOLOGY DOs AND DON'Ts

DO	DON'T
Do describe your project in clear and simple terms and in a logical order .	Don't try to impress (or bamboozle) the reviewer with excessive statistical complexity and jargon.
Do commit to research questions before seeing the data, and account for multiple comparisons.	Don't do anything that looks like “data dredging” — i.e., cherry-picking the results from a large group.
Do have “falsifiable” hypotheses, and be open to any possible conclusion from your data.	Don't say anything that “pre-empts your conclusion”, or suggests that you want a certain conclusion to be true (e.g., getting a “significant” effect).
Do be clear and precise about whether or not you are interested in cause-and-effect, and use precise and consistent language for this.	Don't equivocate between statistical association and causality, or treat predictive and causal terms as if they were interchangeable.
Do consider confounders, mediators, etc., and use appropriate experimental/statistical protocols to deal with these.	Don't write in a way that might make the reviewer think that you don't understand that “correlation is not cause”.

CHEAT SHEET – DATA AND MODELLING DOs AND DON'Ts

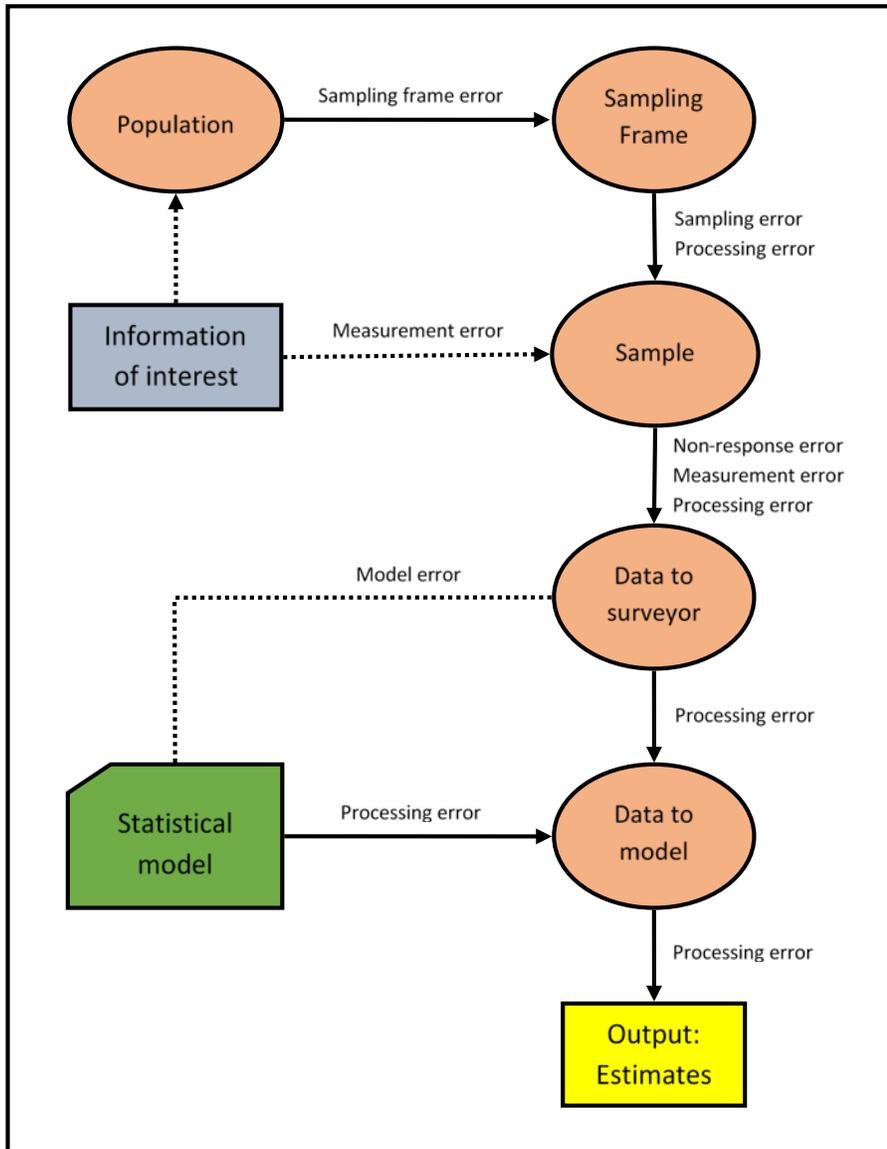
DO	DON'T
<p>Do ensure that your data are obtained from an appropriate sampling method that allows you to make sound statistical inferences.</p>	<p>Don't gloss over the sampling method as if it were unimportant, or omit it and hope no-one notices.</p>
<p>Do ensure that your sample size is appropriate to the constraints and trade-offs of the project, and justify your sample size.</p>	<p>Don't leave available data unused unless there is good reason to, or make arbitrary or unjustified choices on trade-offs.</p>
<p>Do give clear heuristic explanations of statistical methods and models, backed up with references.</p>	<p>Don't get into the weeds with all the technical details of the statistical methods and models.</p>
<p>Do formulate a proposed model that you think will be appropriate to your data.</p>	<p>Don't lock yourself into specific model choices that might turn out to be inappropriate.</p>
<p>Do consider common statistical issues, and frankly acknowledge these. State clearly how you will deal with these issues.</p>	<p>Don't “gloss over” tricky statistical issues, or try to bamboozle the reviewer with jargon to avoid dealing with them properly.</p>

The background is a dark blue gradient. In the four corners, there are decorative white line-art elements that resemble circuit traces or data paths, with small circles at the end of the lines. The main text is centered in a large, white, sans-serif font.

OVERVIEW OF STATISTICAL ANALYSIS

OVERVIEW OF STATISTICAL ANALYSIS

- In most applied research projects that use data, your goal is to answer one or more **research questions** about **characteristics** of some **population** of interest.
- The statistical part of your research deals with the long road from characteristics of a population, to a sample, to a statistical model, to give estimates of the thing you want to know about.



- Here is a flow diagram showing information on a population, flowing down to outputs of analysis.
- Each transition here has an associated **source of error**. You should be aware of these sources of error.

OVERVIEW OF STATISTICAL ANALYSIS

- Some research is **predictive inference**, where you just want to predict one thing from another. This type of research looks at observed statistical associations.
- Some research is **causal inference**, where you want to find out the *causal effect* of one thing on another. This type of research usually also looks at experimental controls.

OVERVIEW OF STATISTICAL ANALYSIS

- We often cannot sample directly from the population of interest, so we form a **sampling frame** of units that can be sampled. You will choose a sampling frame that is similar to the population.
- The observed **sample** is obtained from the sampling frame by some **sampling method**. Measuring sampling error generally requires randomised sampling done by proper methods.

OVERVIEW OF STATISTICAL ANALYSIS

- Data is usually costly or inconvenient to collect, so it may be necessary to limit the **sample size**. Data might be collected in one go, or sequentially (i.e., you can get more later).
- Data from the sample to the researcher is often subject to non-response or other problems that lead to **missing data**. This can create complications when data are not “missing at random”.

OVERVIEW OF STATISTICAL ANALYSIS

- Research questions should be specified clearly before seeing the data, and multiple comparisons should be taken into account. You might want to **pre-register** your research questions publicly.
- Causal analysis is also complicated by **confounding** variables and **mediator** variables that might or might not be observed. This necessitates particular statistical/experimental protocols.

OVERVIEW OF STATISTICAL ANALYSIS

- The appropriate class of statistical models is probably known in advance, but you want your **final model** to fit the data well and have good “diagnostic” properties.
- You should be aware of all the **sources of error** in all the transitions from the source of data to your conclusion.

CHEAT SHEET - DESCRIBING YOUR METHODOLOGY

Description	What you should explain
Methodology and research goal (2-6 sentences)	What is your population of interest? What characteristic(s) are you interested in, and how are these measured ? Give an clear and logical overview of how you will use data to answer your research question. Distinguish clearly between the variable you are interested in, and any “proxy” or operational measurement being used as a stand-in for this.
Experimental protocols (if needed) (4-8 sentences)	Are you doing predictive inference or causal inference — i.e., do you need to know the <i>causal effects</i> of variables or not? If you are interested in making causal inferences, how will you go from statistical associations to cause-and-effect? What (if any) protocols have you imposed — e.g., randomisation/blinding. How are you dealing with confounding variables and mediator variables?
Pre-registered research questions (1 sentence)	Have you publicly pre-registered your research questions? If so, where are these registered? (This is something you should consider – pre-registration of research questions in a public repository prevents <i>post hoc</i> analysis and adds credibility.)

CHEAT SHEET - DESCRIBING YOUR DATA

Description	What you should explain
Source of your data (1-2 sentences)	What is your sampling frame , and how did you get your sample from this sampling frame — i.e., what is your sampling method ? If necessary, reiterate the target population in your research and its difference to the sampling frame.
Joining data from multiple sources (1-3 sentences)	If you are joining data from multiple sources, specify the sampling frame and sampling method for each source. Describe, in general terms, how you join these data sets (i.e., how you match people/items in the different sets). Don't go into technical detail — just give the reviewer a clear idea of the final data.
Variables and types (1-3 sentences)	List the variables in the final data and their types (e.g., numerical, categorical, ordinal, etc.). If there are too many variables to list these easily, at least give an overview of contents of categories. Specify which variables are to be predicted, and which are being used as predictors.
Overview and structure (1-3 sentences)	What do your “data frame(s)” look like? How many variables? Don't be ambiguous – are numbers from long-form or wide-form of data? How many data points will you have (sample-size calculation – discussed soon)?

CHEAT SHEET - DESCRIBING YOUR STATISTICAL ANALYSIS/MODEL

Description	What you should explain
Describe the proposed model (1-2 sentences)	You should generally specify a reasonably broad class of model (e.g., linear regression model, negative binomial model, etc.) Describe the model type, and state the output variable and input variables. Don't use model terminology that is not widely known; if you refer to a model form that the reviewer might not know, give a rough explanation to go with it. Have you given enough information to allow the reviewer to write out your model form?
Always specify some wiggle-room (1-3 sentences)	You propose to use a particular class of model, but your specific model will be chosen <i>after</i> seeing the data, based on statistical considerations and diagnostic testing. It might involve transformation of variables, changing model form, etc. Reviewer wants to know that you won't cheat to get a pre-conceived conclusion, so make sure you say that your choices will be based on <i>statistical considerations</i> .
Outputs, tests and comparisons (1-3 sentences)	What outputs will you get from your model, and what tests and comparisons will you look at? How will you deal with multiple comparisons (if they arise)? Do your outputs answer your research questions? Have you established this clearly?

The image features a dark blue background with white, stylized circuit board traces in the corners. These traces consist of lines and small circles, resembling electronic components or connections. The central text is in a bold, white, sans-serif font.

SAMPLE SIZE CALCULATIONS

SAMPLE SIZE CALCULATIONS

- Sample size calculations are hard, and poorly understood. First we need to understand *what needs to be specified* before you can calculate the desired **sample size**.
- The “best sample size” is a full census of the whole population. The only reason to sample less than this is because there is some **constraint, cost, or trade-off**.

SAMPLE SIZE CALCULATIONS

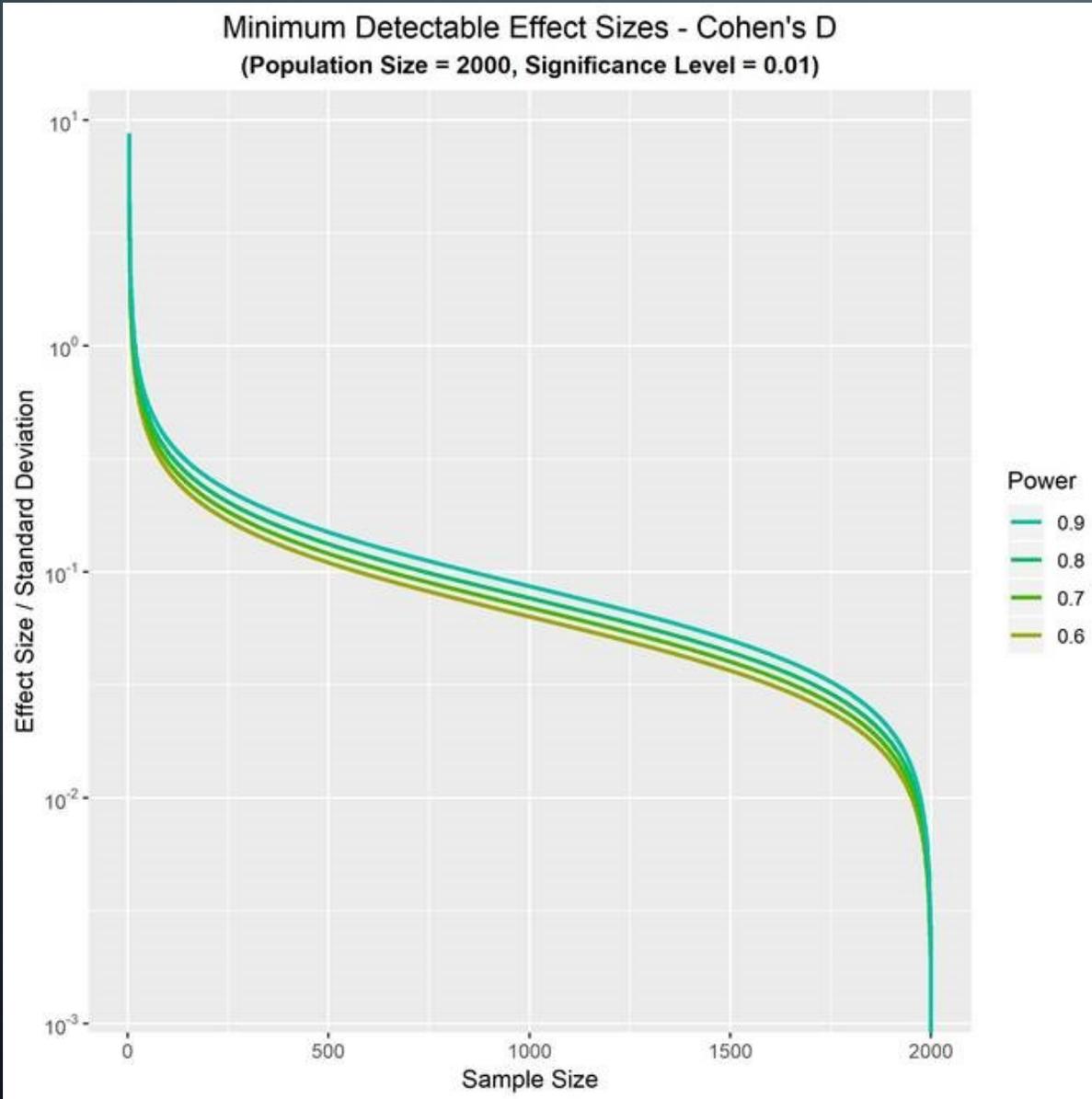
- There is no such thing as the “right sample size” in the absence of a specified **level of desired accuracy** for a specified statistical inference/prediction.
- Sample size calculations are done *before you see the data*. Thus, the level of accuracy you want is usually framed in terms that are *relative to the unknown standard deviation* (e.g., Cohen’s D).

SAMPLE SIZE CALCULATIONS

- Be clear about the content of **one data point** — this might not correspond to a single participant (e.g., if each participant gives multiple data points). Convert your calculation to an intuitive measurement (e.g., number of participants needed).
- Sample-size calculations require **time and consideration** — they are not something you do the day before the application is due.

SAMPLE SIZE CALCULATIONS

- Your sample size calculations should show the trade-off between accuracy and number of data points. This can be presented in a **plot** showing the sample size and the resulting level of accuracy.
- In your application, you will want to remove all the technical analysis to **supplementary materials**. This should be written up so that someone can come back to it in a year and understand it.



- **Determining sample size:** Set fixed values (e.g., significance level + power). Show sample size against detectable effect. You may need to convert one or both axes to a log-scale to see things properly.
- This allows you to visualise the “marginal benefit” of each additional data point. This assists in choosing the desired level of accuracy

CHEAT SHEET - DESCRIBING YOUR SAMPLE SIZE

Description	What you should explain
Describe the basis for the calculation (1-2 sentences)	Sample size is determined by reference to a statistical test or inference (e.g., test/confidence interval). What is used as the basis of your calculation? Is data collection done in one go, or is it done sequentially?
Specify desired level of accuracy (1-2 sentences)	There is no such thing as the “right sample size” without specifying the desired level of accuracy for a statistical inference/prediction. This accuracy will be relative to the standard deviation of your —yet to be collected— data. For a hypothesis test , specify significance level, power, and minimum detectable effect size. For a confidence interval , specify confidence level and interval length.
Describe expected missing data (1-2 sentences)	Anticipate non-response/missing data and account for this in your calculation. If your data collection is done sequentially then this might not be an issue, since you can keep collecting data until you get the amount you want.
Justify your specifications (1-4 sentences)	Show trade-off between accuracy and number of data points, and use this to choose your sample size (calculations in supplementary materials). Make sure you specify your sample size in an appropriate unit of measurement.

QUESTIONS?

