Statistics Frequently Asked Questions

1. What statistical software should I use?

We would recommend you use Stata because this software is available from Gold Coast Health and we can show you how to use it.

Stata is very powerful but relatively easy to use. Because it is very popular, there are many online sites that provide teaching resources for Stata, there's even a Stata YouTube channel. Googling, "How to….in Stata" is a good start. Stata has a help tab at the top of the screen that will also direct you to useful resources.

There are other options such as R, SPSS, and SAS. Many software providers offer free trial versions for academic purposes, allowing you to explore their features. R is free and very powerful but is challenging to learn.

Having said all this, you will first need to understand your research question, know a bit of statistics, and, preferably, have discussed analysis options with a statistician before diving into Stata or one of the other statistical packages.

2. What format is best to collect research data in for later statistical analysis?

Most commonly, people will collect their data for research projects in an Excel spreadsheet; either directly or from a data capturing software such as RedCap. There are two formats that can be used, Wide Format and Long Format, both are fine to use. In wide format each row shows the data for a single participant. If there are repeated measures of a variable, they need to be named appropriately, eg BMI_3, BMI_6 for BMI at 3 months and 6 months.

In long format, there are multiple rows for each participant relating to each time point that data was collected. So, in the above example, each participant would have two rows, there would be a single BMI variable, and a Time variable that would contain values 3 or 6.

Here is a link to a publication titled "Data Organization in Spreadsheets" (Broman and Woo, 2018) that gives some useful tips on how to prepare your spreadsheet for data analysis.

There are likely to be some specific things to be aware of depending on which statistical software you plan to use but the basic principles are:

> ***Create a new Data tab:*** Copy the original data into a new tab that can be edited and modified to best suit analysis. Another, clean, data tab may be used to create data in a format that is best for export to the statistical software (ie simple variable names, no calculations). Save the file under a new name so that you always have the original.
>
> ***Be consistent, systematic, and simple:*** Consistent entries for the same thing; eg one of Male, male, or M. (when collecting data, Excel's drop down choices are useful for this). Simple and consistent column names (which will become variable names); eg BMI_3, BMI_6 for BMI at 3 months and 6 months; not BMI (kg/m2) at 6 months. When naming variables, go from the general to the specific, eg Temp_max, rather than Max_Temp. Make sure column/variable names are unique.

Do not put anything in a cell if data is missing: Excel (and Stata, after data has been imported) recognize an empty cell as missing data. If you want to do any preliminary analysis in Excel, it is only possible if missing data is an empty cell. Broman and Woo (2018) recommend using a missing value indicator such as NA. This, however, assumes your data will be saved as a text file and exported to the R statistical package. Be careful when importing data into Excel from a data capturing software such as SurveyMonkey. They may give a numerical value for missing data; SurveyMonkey uses “11”.

Do not mix numerals and text in a numerical variable: Any text in an otherwise numerical variable will mean the variable is saved as text and cannot be analyzed. Do not use a “?” to designate uncertainty. If laboratory results use “<” or “>” for out-of-range results, decide upon a value that is reasonable to assume for such results. If a value in variable X needs a comment, create a new “VariableX_Comment” column.

Put just one thing in a cell: For example, for the TNM cancer staging system, rather than writing, T1N0MX, create separate columns for Primary Tumor (T), Lymph Nodes (N), and Metastasis (M) with 1, 0 and nothing in each of the cells respectively.

Create a data dictionary: The data dictionary is created in a separate Excel tab. Each column name (variable) is defined and described, eg Acronym and units defined; BMI_3: Body Mass Index (kg/m^2) at 3 months. A trick to creating the Data dictionary, and to simplify variable names, is to copy the top row of the data collection sheet, the column/variable names, into the data dictionary and use the transpose function of paste to create a column of these names in the first column. In the second column, you can then write the description of the variable. If the original data collection sheet used long column names, such as whole survey questions, these can be copied into the second column of the data dictionary and simplified versions created in the first column. Once completed, the simple names can be copied and transposed back into the top row of the data collection tab.

The Data Dictionary can also be used as a place to do some minor analyses or help tidy up your data (see later).

Continuous variables and categorization: Generally, if a variable can be collected as a continuous variable do so. It can always be categorized later if needed. Sometimes a variable in the original data consists of an inordinate number of categorical options, eg diagnoses (some of which may actually be spelling variations or due to inclusion of stray spaces). You may want to reduce this to a more manageable number of more relevant categories. A trick to doing this is to copy the relevant column into the Data Dictionary and use the “remove duplicates” function in Excel to reduce the column to unique values. Then, in the adjacent column to the right, you can provide a numerical designation to each unique value based on your clinical grouping strategy. These two columns form a “lookup” table. Go back to the data tab and create a new column after the column you copied. Then use the vlookup command to populate the new column with the numerical values from your look up table. Eg =VLOOKUP(AF2,'Data Dictionary'!F$1:G$19,2,FALSE). Where AF2 is the cell whose contents you are “looking up” and 'Data Dictionary'!F$1:G$19 is the location of the lookup table (you highlight it). The $ signs are inserted to fix where Excel expects to find the look up table.

Dates: Dates can be tricky but Excel is reasonably robust in that it will recognise most things that look like a date as a date. Try to be systematic and consistent, always using the same date format, eg 21/06/2024. Computer programs store dates as a number, for Excel it is the number of days from the 1st of January 1900. So, to find age or length of stay, for example, it is a simple matter of subtracting the past date from the most recent date. If date data also has a time component, as is often the case when extracting dates and times from

ieMR, the time unit could be hours or minutes. I have even seen milliseconds. The time difference, therefore, may have to be converted to something more convenient, eg, years for age by dividing the number of days by 365.25.

Data for specific instruments or indices: When collecting data from a particular survey tool it is important to know how it is to be recorded. Some instruments, for example, expect results from a Likert scale to be recorded as 1 to 5, others 0 to 4. Some questions may be weighted differently or the Likert scale for some questions interpreted differently. Some instruments will have groups of questions that form domains that should be summed (with appropriate weightings) together or the whole questionnaire summed. It is a good idea to know beforehand how the instrument expects the data to be handled so that you can design the Excel sheet accordingly.

Organize the data as a single rectangle: Participants as rows and variables as columns, and with a single header row containing the variable names. Often people will collect data into two or more sheets for treatment and placebo, for example. This data needs to be together so that it can be analysed. It can be tricky to combine two spreadsheets so it is better to set up the data collection into a single spreadsheet with a Treatment variable to designate Placebo (0) or Treatment (1).

Do not use font colour or highlighting as data designators: If it is important to designate a data entry as something and that designation is likely to be something that an analysis can be based upon, then create a variable for that designation with values such as 0, 1, 2.

Do not justify your data: Excel has the useful feature that it automatically justifies text to the left and numbers (including dates) to the right. This means if text is accidentally entered into a numerical variable, even if it is a number that has been stored as text, it will look obvious and is easy to correct. Choose good names for things, make backups, use data validation to avoid data entry errors, and save the data in plain text files.