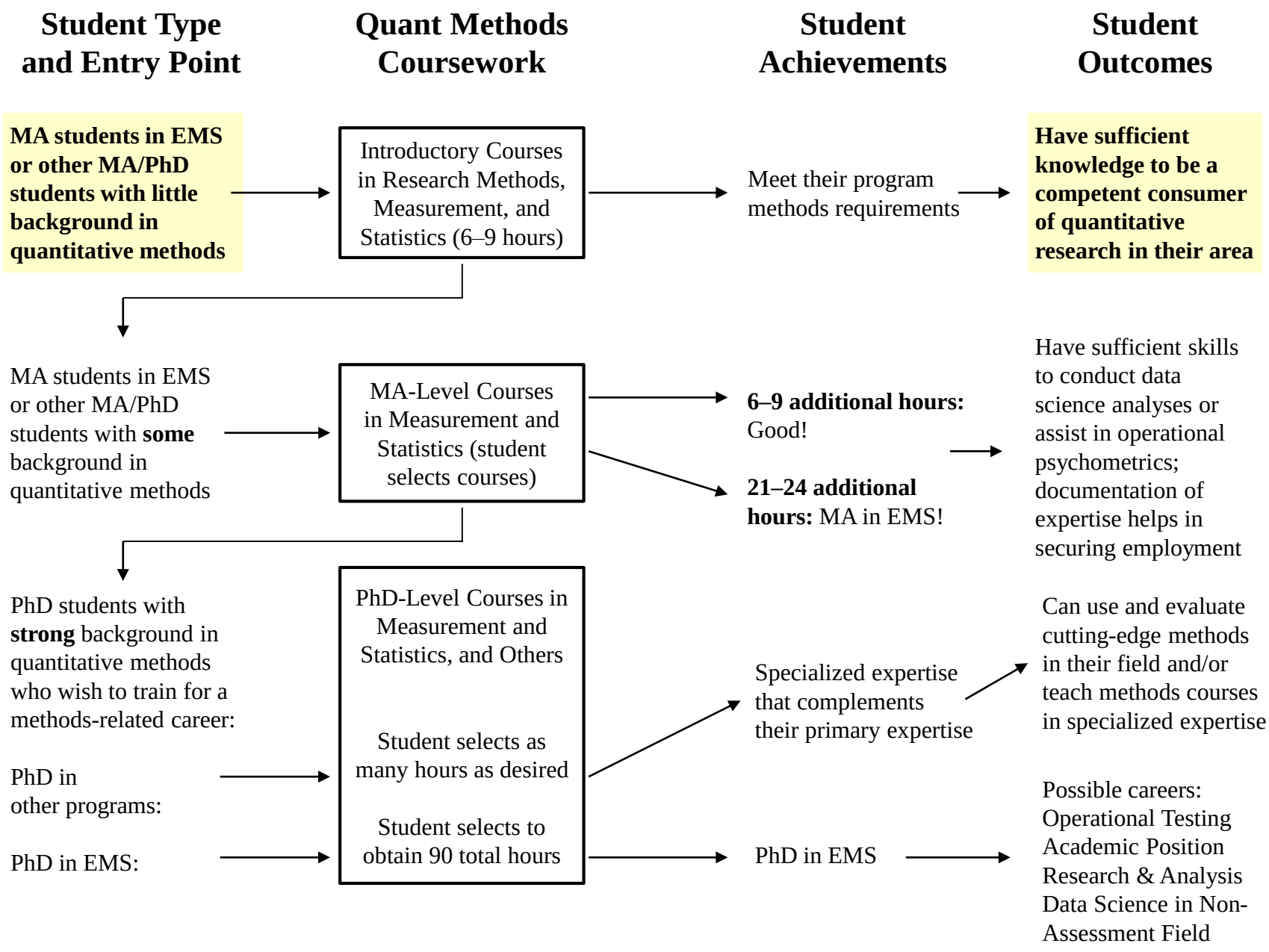
Introduction to PSQF 6242 and to Quantitative Methods

- Topics:
  - What to expect this semester (and beyond)
  - Independent versus dependent sampling
  - Types of variables in quantitative analysis
  - Tips for storing variables in databases

Reasons Why You Are Here

1. This class fulfills a requirement (true, but uninteresting).
   - It's ok if this is the *only* reason you are here, but I hope to convince you otherwise!

2. This class will teach you about research using **quantitative methods**—yes!
   - One method by which to answer questions—in real life or in research settings—is by collecting quantitative data
   - The process of summarizing that data by searching for patterns that answer questions occurs through statistics
   - **Quantitative methods = Quantitative data + application of statistical models to answer questions**
   - Let's examine the levels of expertise you can acquire…


Student Type and Entry Point
Quant Methods Coursework
Student Achievements
Student Outcomes
MA students in EMS or other MA/PhD students with little background in quantitative methods
Introductory Courses in Research Methods, Measurement, and Statistics (6–9 hours)
Meet their program methods requirements
Have sufficient knowledge to be a competent consumer of quantitative research in their area
MA students in EMS or other MA/PhD students with some background in quantitative methods
MA-Level Courses in Measurement and Statistics (student selects courses)
6–9 additional hours: Good!
21–24 additional hours: MA in EMS!
Have sufficient skills to conduct data science analyses or assist in operational psychometrics; documentation of expertise helps in securing employment
PhD students with strong background in quantitative methods who wish to train for a methods-related career:
PhD-Level Courses in Measurement and Statistics, and Others
Student selects as many hours as desired
Student selects to obtain 90 total hours
PhD in other programs:
PhD in EMS:
Specialized expertise that complements their primary expertise
Can use and evaluate cutting-edge methods in their field and/or teach methods courses in specialized expertise
PhD in EMS
Possible careers:
Operational Testing
Academic Position
Research & Analysis
Data Science in Non-Assessment Field

What Will "Statistics" Mean to Us?

- Statistics will be **applied math** used for a relevant purpose!
- Competent consumers and users of quantitative methods must learn the **logic behind the uses of statistical models**

- **This will NOT require anxiety-provoking behaviors like:**
  - Calculating things by hand—computers are always better, and more advanced statistical models cannot be implemented by hand anyway
  - Deriving formulas or results—it's ok to trust the people who specialize in these areas to have gotten it right and use their work (for now, at least)
  - Memorizing formulas—it's ok to trust the computer programmers who have implemented various estimation techniques (for now, at least)

- It WILL require learning and implementing **new language and decision guidelines** for matching data, questions, and models

How You will Acquire the Language and Logic of Statistical Modeling

- I will NOT:
  - Use infrequent high-stakes testing to assess your level of learning
  - Ask you to complete practice problems by hand
  - Present statistics as a series of unrelated ideas and formulae

- I WILL:
  - Use **formative assessments (in ICON)** to help you review concepts (6 planned; 12 points for **completing them at all**)
  - Use **homework (in my custom online system)** to give you hands-on software practice (6 planned; 88 points for **accurately completing** both computational and interpretation questions)
  - Present statistics by linking data, questions, and models explicitly

Our Responsibilities

- My job (over Zoom exclusively this semester):
  - Provide custom lecture materials and examples that are accurate, comprehensive, and with the necessary scaffolding for your future use
  - **Answer questions** via email, in individual meetings, or in group-based office hours—you are ALL invited to attend to work on homework during office hours and get immediate assistance if you want it

- Your job:
  - **Ask questions**—preferably in class, but any time is better than none
  - **Review** the class material **frequently**, focusing on mastering the vocabulary (words and symbols), logic, and procedural skills
  - **Practice** using the software to implement the techniques you are learning **on data you care about**—this will help you so much more!
  - Read the texts *if you feel they are helpful* (they are mainly for reference)
  - Don't wait until the last minute to start homework, and don't be afraid to **ask for help** if you get stuck on one thing for more than 15 minutes

Class-Sponsored Statistical Software

- To help address the needs of different degree programs, I will show examples using **both SAS and STATA**
  - Why not SPSS? Because it doesn't have everything we need and it doesn't leave as much room to grow into advanced models
  - Caveat: I am a heavy-duty SAS user who picked up enough STATA to teach multilevel modeling workshops using it
  - So if you have STATA tips, please share them with me!

- Things to consider when choosing which one to focus on:
  - More programs = more entries in "technical skills" part of your CV
  - Both SAS and STATA are available through the UIowa Virtual Desktop, **but STATA may not be available from off campus**
  - What program will be used in your quant classes to follow? What do the other members of your research lab use?
  - Btw, I can also help you in SPSS, M*plus*, and a little bit of R

SAS vs. STATA: My Opinion

Activity	Winner	Commentary
Working with raw files or multiple datasets	SAS, hands down	As of STATA 15, only one dataset can be open at once—problematic for messy data management
Within-dataset manipulations	Tie, but STATA for some tasks	STATA wins for group-centering, stacking, and unstacking (used for multilevel models)
Data analysis	Tie, but SAS for some tasks	I've had estimation problems in STATA for certain advanced model variants (within multilevel models)
Post-estimation (i.e., predicted outcomes or simple slopes)	STATA, hands down	STATA has simple yet powerful options for doing these tasks in bulk that SAS doesn't have
Automating data tasks (i.e., loops)	Tie	Both programs have ways to do this, but I only know how in SAS…

This Semester's Topics

- Sampling, data, and quantitative methods terminology
- Univariate statistics and inferences thereof
- Bivariate measures of association and significance testing

- Intro to General Linear Models (GLMs) as a one-stop shop *for predicting one conditionally normal outcome per person*
  - Quantitative predictors? Is "(linear) regression"
    - One predictor variable? "Simple (linear) regression"
  - Categorical predictors? Is "Analysis of variance (ANOVA)"
    - One predictor variable? Is "One-way ANOVA"
    - Two predictor variables? Is "Two-way ANOVA"
  - Both kinds? Is "Analysis of Covariance (ANCOVA)"
  - We will learn how to test moderation of all kinds, too!*

Moving Forward… What Else is There?

- If your schedule permits (likely in Spring 2022), please plan to take a **new class** I have planned as a follow-up to this one:
  - Currently listed as PSQF 7375 Applied Generalized Linear Models
  - Modeling non-normal outcomes; path and mediation analysis
    - These models will REALLY help you in common research settings
    - This will really help you learn structural equation modeling (SEM)

- Want to keep going? What other quantitative methods classes you will be able to take will depend on the instructor—the material in PSQF 6242 is an acceptable pre-requisite for:
  - Clustered or Longitudinal Multilevel Models; SEM (with me)
  - Design of Experiments (with Dr. Aloe)
  - Computer Packages for Statistical Analysis (with Dr. LeBeau)
  - Not sure what classes would be good for you? Please ask me!

Where to Begin? Sampling Vocabulary

- Who are we trying to know about, more generally? →
  To what **population** do we want to make inferences?

- Accordingly, from whom should we collect data? →
  By what process should we select our **sample**?
  - **Variables** are characteristics that differ across persons in a sample

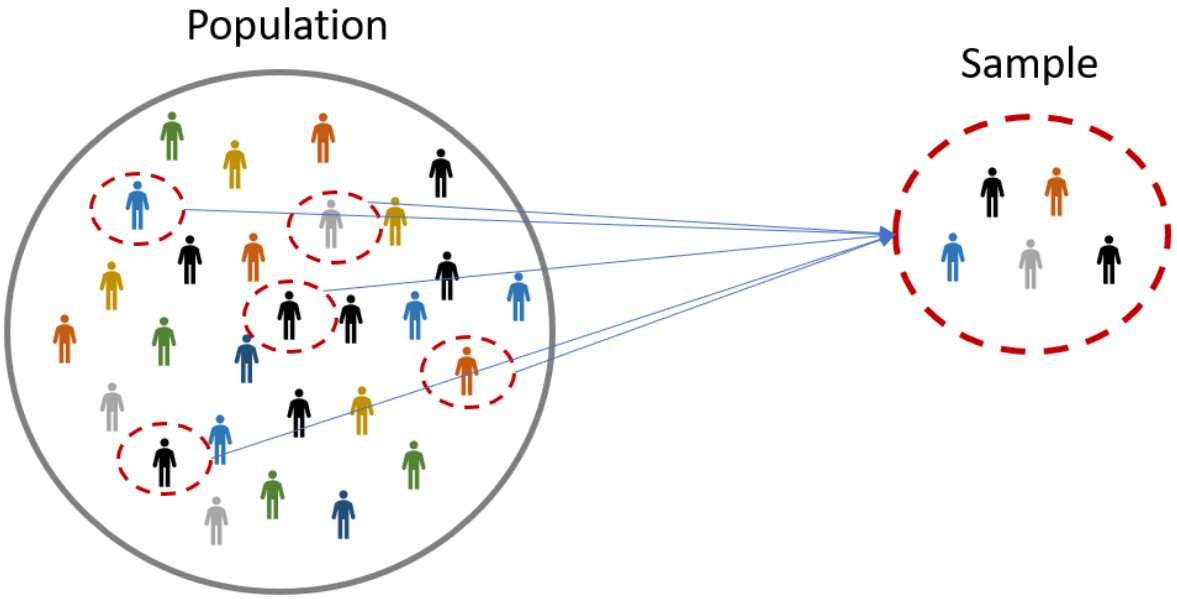


Image borrowed from: https://www.sigmamagic.com/blogs/online-sample-size-calculators/

Where to Begin? Sampling Vocabulary

- Example: Let's say a researcher wants to examine graduate student life, and uses a survey to collect self-report info on program membership, stress levels, and well-being

- So what **types of sampling** should we use? For instance:
  - Collect data for multiple students from the same program only? Then **program would be a *constant*, not a *variable***
  - To examine **differences between programs**, we'd need to sample multiple programs from the same college, at a minimum
  - But would it help our **generalizability** to include multiple colleges from the same university, or even from multiple universities?
  - Should we survey each student once? Or would **several times** be better?
  - Should we also try to collect **corresponding data** from other people who know each student well (e.g., their partners, friends, family)?
- These questions lead to **independent** versus **dependent** sampling

Independent and Dependent Samples

- Example of a (likely) **independent sample**: One occasion of measurement each from students in the same program
  - If program is a **constant**, not a **variable**, it can't be part of any research questions (but then program differences are controlled)

- Examples of **dependent** (= naturally related) **samples** (in which your analyses must account for common sampling):
  - Sample lots of programs (e.g., >20) from same university
    - e.g., Stress rates of persons from the same program may be more related (dependent) than those of persons from different programs
    - This is known as "**clustered**" or "**nested**" data
  - Sample each person more than once
    - e.g., Stress rates at occasions from the same person may be more related (dependent) than those of occasions from different persons
    - This is known as "**repeated measures**" or "**longitudinal**" data
    - Collect both self-report and another-report ratings → "**dyadic**" data

PSQF Courses that Cover Analysis of Independent and Dependent Samples

- **This semester**, we will only be able to cover analysis of quantitative data from **independent** samples (GLMs)
  - Using "**univariate**" statistical models (of one observation per variable per person) predicting a **numeric variable**

- My other courses are extensions for **dependent** samples:
  - Generalized Linear Models: predict **other kinds of variables**, as well as introduce "**multivariate**" statistical models for predicting **multiple outcomes** (and testing mediation)
  - Longitudinal Multilevel Models: multivariate statistical models for **repeated measures** data (of occasions nested in persons)
  - Clustered Multilevel Models: multivariate statistical models for **clustered data** (of persons nested in groups)

Back to This Course… What's Next?

- Rest of Lecture 0: Present labels for the **kinds of variables** collected in quantitative research studies
  - Big picture: categorical or quantitative? This taxonomy will guide how to describe them or use them in subsequent analyses
  - A little bit about managing databases for quantitative studies

- Lecture 1: **Univariate** analysis (i.e., for one variable at a time)
  - How to summarize a variable's salient features ("**statistics**")
  - How to compare those statistics from one sample to what they are thought to be in the population ("**parameters**")

- Lecture 2: **Bivariate** analysis (i.e., for two variables at a time)
  - How to assess the relationship between two variables (the proper methods for which depend on which kind of variables they are)

Types of Variables in Quantitative Analysis*

* Note: this is related to traditional levels of measurement, but I am approaching it from more of a "how-to-model them" perspective

- First, **categorical variables: *where the numbers are labels***
  - Binary (dichotomous) = 2 choices (typically coded as 0 or 1)
    - e.g., dead or alive; pregnant or not
  - Nominal = 3+ unordered choices
    - e.g., favorite type of pet, degree program
  - Ordinal = 3+ choices with some natural (undeniable) order, but the distances between the values used don't mean anything
    - e.g., 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree
    - Equally ordinal (and thus also acceptable) values: 1, 20, 300, 4000
  - Synonyms for a "**categorical**" variable: discrete variable, qualitative variable, grouping variable, factor variable, CLASS variable (stands for "classification variable" in SAS)

Types of Variables, Continued

- Next, **quantitative variables** where the **numbers are really numbers** (**interval** measurement → equal distances between values), but that have one or more natural boundaries

  - Binomial = number of occurrences out of known possible
    - e.g., # correct on a test, which is bounded by 0 and total possible
    - Correcting for different totals possible by computing proportion correct (or rate of occurrence) is still binomial (just bounded by 0 and 1 instead)

  - Count = number of occurrences out of unknown possible
    - # of cigarettes smoked each day
      (minimum = 0, but maximum could be any positive number)

  - Count variables have special cases involving zero values:
    - No zeros possible? → zero-*truncated* count
    - More zeros than expected? → zero-*inflated* count

Types of Study Variables, Continued

- Lastly, **quantitative** variables that are "**continuous**" (still with interval measurement in which the numbers are numbers)
  - But **continuous means unbounded** → can theoretically go on forever in either direction AND take non-integer values
  - Although in this semester's general linear models (GLMs) our predictors can be any type of variable, **all our outcomes must be *plausibly* continuous with interval measurement**
    - This is because GLMs use a *conditional* normal distribution (stay tuned)
    - Otherwise you need "general*ized* linear models" (in another class) by which you can choose different distributions for different variable types

- Don't worry, the key word is "**plausible**": Truly continuous and interval variables are rare, but there are many variations we often pretend are "continuous and interval enough"
  - These I like to call "continu-ish" variables…

Examples of Continu-ish Variables

- **Ordinal-treated-as-interval**: Values are still ordinal but there are enough distinct values that people justify treating them as interval
  - e.g., one item on 1-4 ordinal scale? Most likely treated as ordinal
  - e.g., sum of 10 items? Likely treated as interval and continu-ish (even though there are no non-integer values, and range is 10-40)
  - e.g., average of 10 items (better if some items are missing)? Likely treated as interval and continu-ish (non-integer values, but range is 1-4)
  - Binomial and count variables are often treated as continu-ish, too

- **Interval, but still likely continu-ish** (may be bounded in practice)
  - e.g., response time, heart rate → really is continuous with non-integer values (limited only by measurement precision) but is bounded at 0
  - e.g., latent trait estimates from measurement models (IRT, CFA,SEM) → non-integer values, but may have observed ceiling or floor effects

One Last Type of Variable: Ratio

- A **ratio scale** has a true zero point
  - Examples: length, height, volume, money
- Ratio scales allow references like "twice as long" or "half as much volume" to actually be meaningful
- Ratio scales do not apply to most quantitative variables in the social sciences (which tend to be interval at best)
  - e.g., a score of 50 vs 100 on an IQ test doesn't mean "half as intelligent" in the same way as a ratio scale
- For all intents and purposes, variables with ratio scaling can be treated as just another quantitative variable

Working with Datasets and Variables in Statistical Software (e.g., SAS, STATA)

- Quantitative data can be stored in a variety of formats
- We will use data stored in excel (with **.xlsx** extension) because it is viewable outside of specific statistical software, but it can easily be imported into SAS or STATA (and others)

- <u>3 steps to import .xlsx data into either stats program:</u>
  1. Save dataset to a folder and get the address to that folder
  2. Copy the folder address into the program syntax
  3. Run (execute) the syntax to import the .xlsx data into the program's native format (i.e., SAS or STATA) for use in analysis

- Historically this has been the hardest step, so I have made new videos using Example 1 to walk you through the process…

Working with Datasets and Variables in Statistical Software (e.g., SAS, STATA)

At least 3 useful pieces of information will be stored for each variable (see demo in videos describing use of SAS or STATA for example 1):

1. **Variable name** = column name (required)
   - No spaces or special characters, must start with letter
   - To be referred to when requesting info or results about that column

2. **Variable label** = column description (optional)
   - Longer text label that can document the variable in more detail
   - e.g., how it was created, # categories, which version or metric

3. **Value label** = verbal labels that go with the numbers (optional)
   - Used for categorical variables only (in which numbers are labels)
   - Makes results easier to read (i.e., don't have to remember values)

How to Store Variables in Databases

- When entering data, there are things you can do up front to save yourself a lot of hassle later:
  - Btw, it's fine—preferable—to use spreadsheets (e.g., excel) to enter the data, no matter how you plan to analyze it
    - But keep in mind that "meaningful" formatting will not transfer

- Put **variable names** in the first row of the spreadsheet
  - Do not use spaces or special characters other than _
  - Use only as many characters as necessary to keep it unique
    - Use variable labels to add extra detail for clarification
  - Start with a letter, not a number (is rule in stats programs)
  - Use a common stem for a series of related variables
    - e.g., stress1, stress2, stress3…. wellbeing1, wellbeing2, wellbeing3…
    - This is helpful when you need to refer to them as a series

How to Store Variables in Databases

- Enter numbers for categorical variables, not text
  - Text variable = string variable = case- and space-sensitive
    - e.g., "control group" is not the same as "Control Group "
  - Add **value labels** to indicate what the numbers represent
    - It can be helpful to use the number in the value label so that the order of the labels is the same alphabetically and numerically
      - e.g., group: 0 = "0. Control Group" 1 = "1. Alternative Group"
  - Do not mix numeric and text entries in the same variable
    - Numbers will be read as text → becomes a string variable instead
  - IMHO: Do not use missing data codes (e.g., -99 = missing)
    - You must define them as such for -99 to NOT be read as data
    - Just leave them missing values blank—if you need to keep track of reasons for missing values, use a new categorical variable to do so

How to Store Variables in Databases

- Tips for handling entry of dependent data more easily
  - Create a unique ID variable for each level of sampling
  - Create separate databases for each level sampling—you can easily merge them together so that the values of the higher-level variables are replicated automatically across the rows of the lower-level database (as is needed)

- For example: people collected from different countries?
  - Person-level database: one row per person; include person ID, country ID, and all person-level variables
  - *Separate* country-level database: one row per country; include country ID and all country-level variables (when merged, will replicate across people)

- For example: multiple occasions from same person?
  - Occasion-level database: one row per occasion; include occasion ID, person ID, and all variables measured per occasion
  - *Separate* person-level database: one row per person; include person ID and all person-level variables (when merged, will replicate across occasions)

Wrapping Up

- End goal of this semester: Learn how to use **general linear models** [GLMs; with variants known as regression, analysis of (co)variance] to analyze quantitative research data
  - Requires learning **new language** (words, symbols, and software) and **decision rules** by which to link types of variables, questions, and models
  - Starts by **summarizing** variables and **associations** between them
  - Continues with GLMs: statistical models for predicting quantitative variables in independent samples (*which need extensions to be covered elsewhere for predicting other kinds of variables or for use in dependent samples*)

- Variables can be described as falling into two main types:
  - **Categorical**: numbers are labels (binary, ordinal, or nominal measurement)
  - **Quantitative**: numbers are numbers (interval or ratio measurement)
    - May be bounded (binomial, count) or continuous (more likely to be "continu-ish")