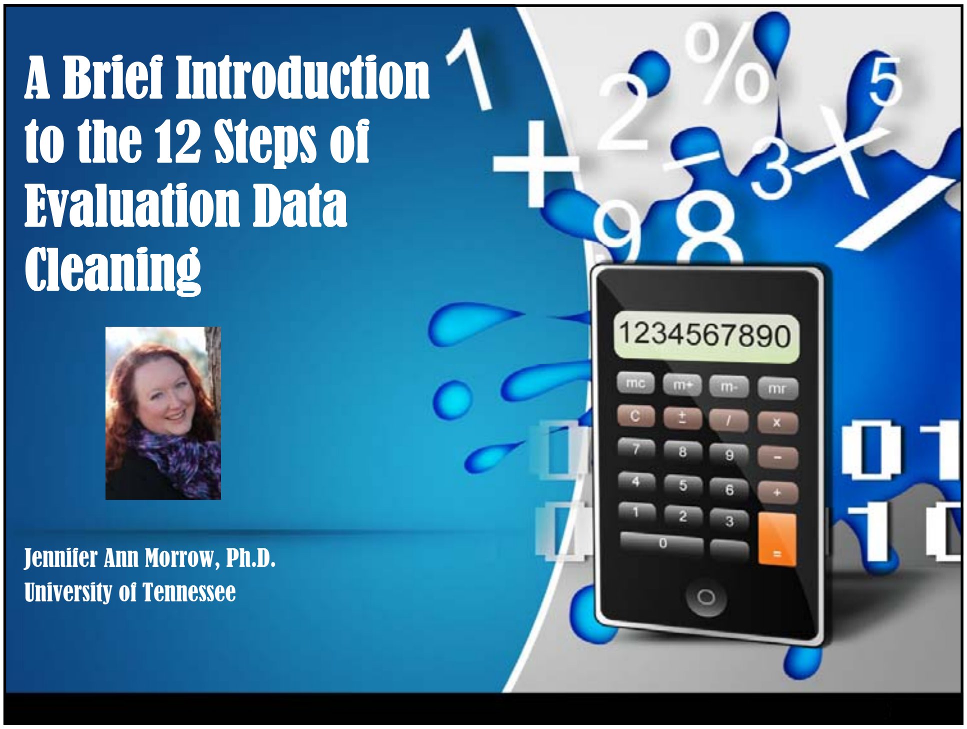
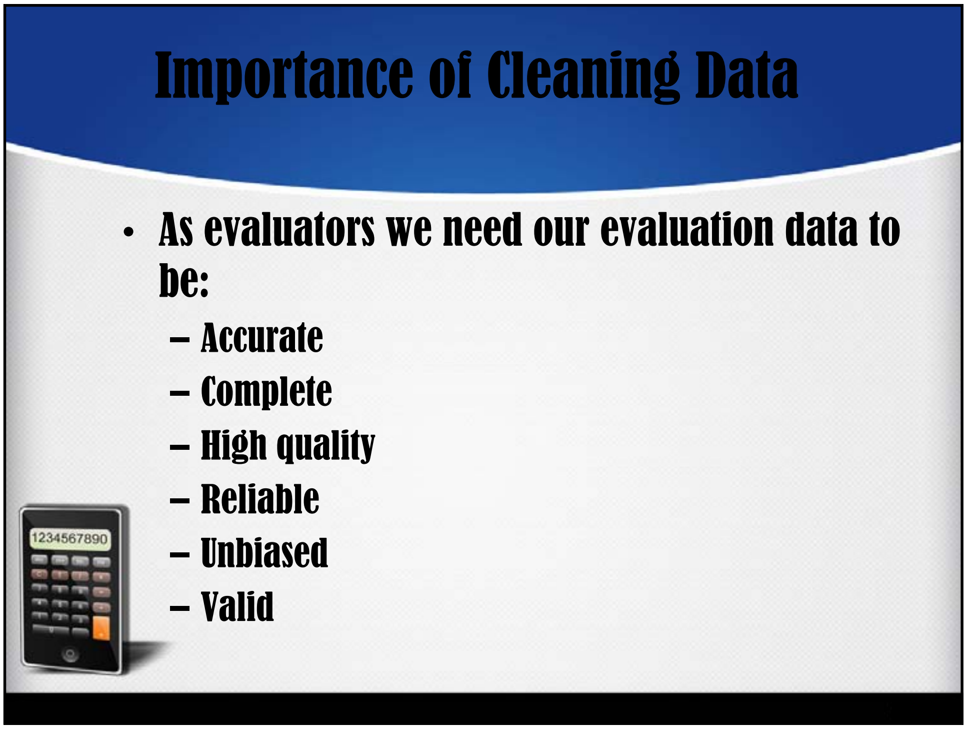


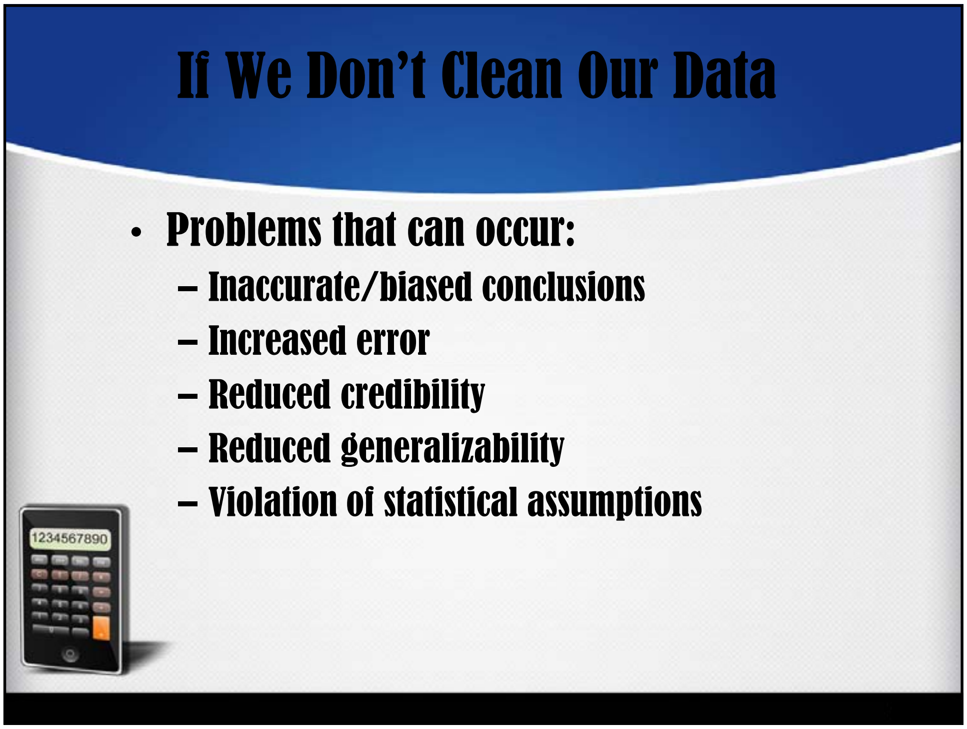
# A Brief Introduction to the 12 Steps of Evaluation Data Cleaning

Jennifer Ann Morrow, Ph.D.
University of Tennessee



## Importance of Cleaning Data

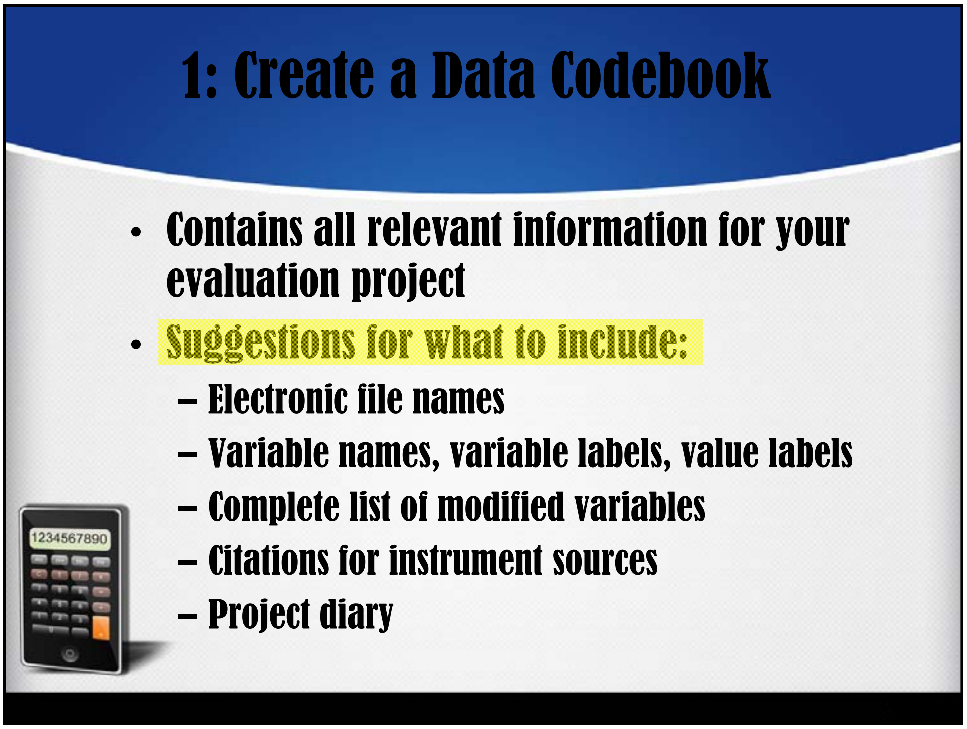
- As evaluators we need our evaluation data to be:
  - Accurate
  - Complete
  - High quality
  - Reliable
  - Unbiased
  - Valid



## If We Don't Clean Our Data

- Problems that can occur:
  - Inaccurate/biased conclusions
  - Increased error
  - Reduced credibility
  - Reduced generalizability
  - Violation of statistical assumptions

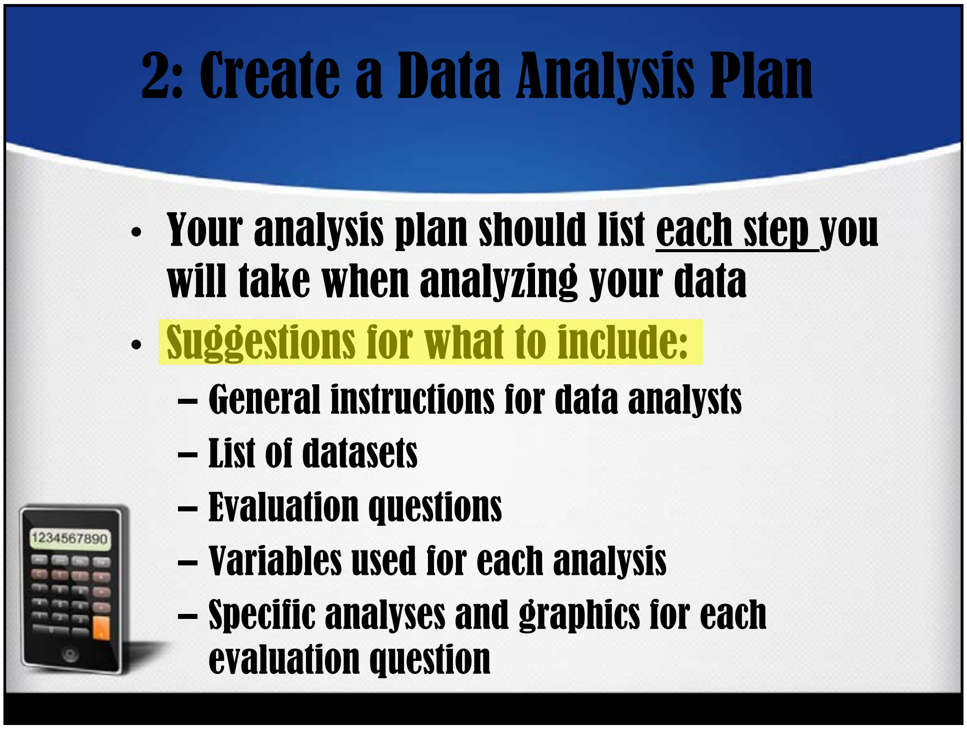
1234567890



## 1: Create a Data Codebook

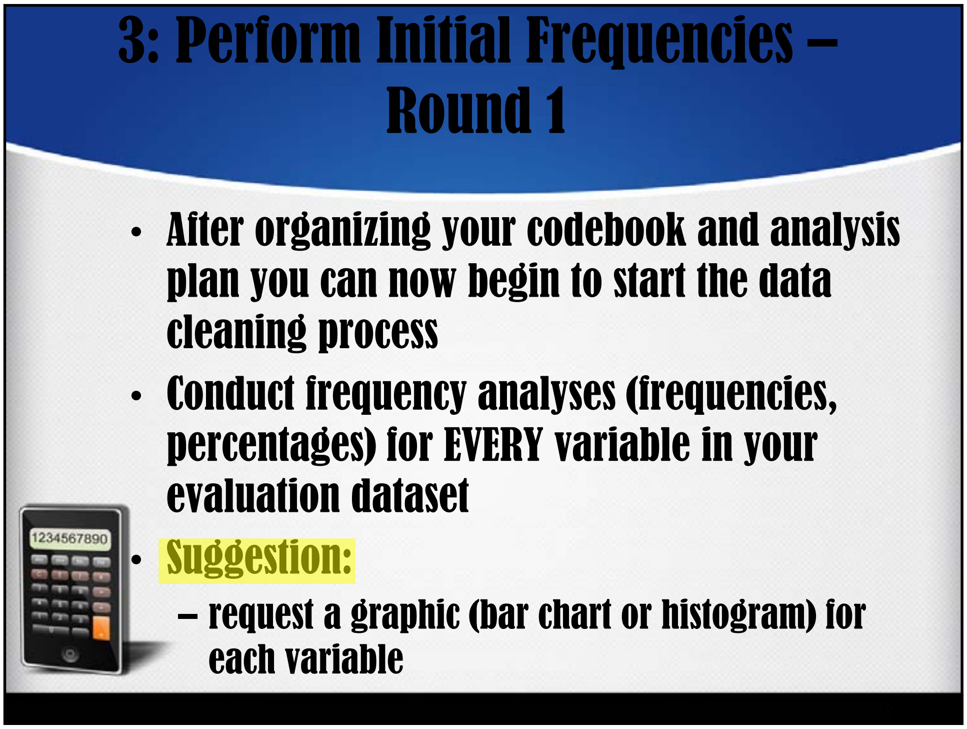
- Contains all relevant information for your evaluation project
- Suggestions for what to include:
  - Electronic file names
  - Variable names, variable labels, value labels
  - Complete list of modified variables
  - Citations for instrument sources
  - Project diary

1234567890



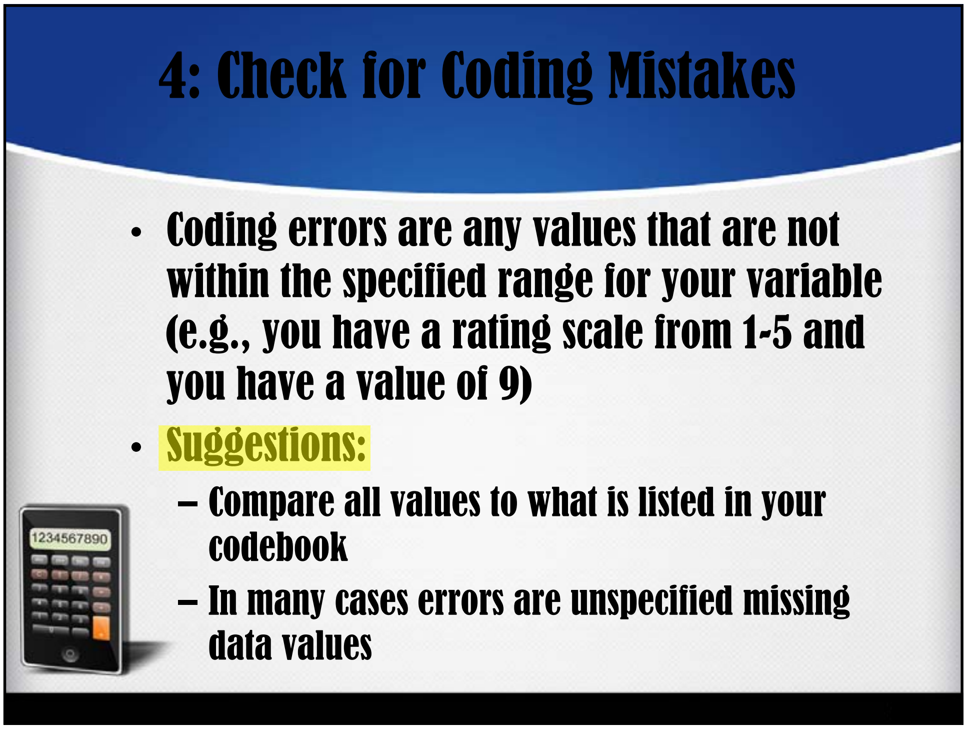
## 2: Create a Data Analysis Plan

- Your analysis plan should list each step you will take when analyzing your data
- Suggestions for what to include:
  - General instructions for data analysts
  - List of datasets
  - Evaluation questions
  - Variables used for each analysis
  - Specific analyses and graphics for each evaluation question



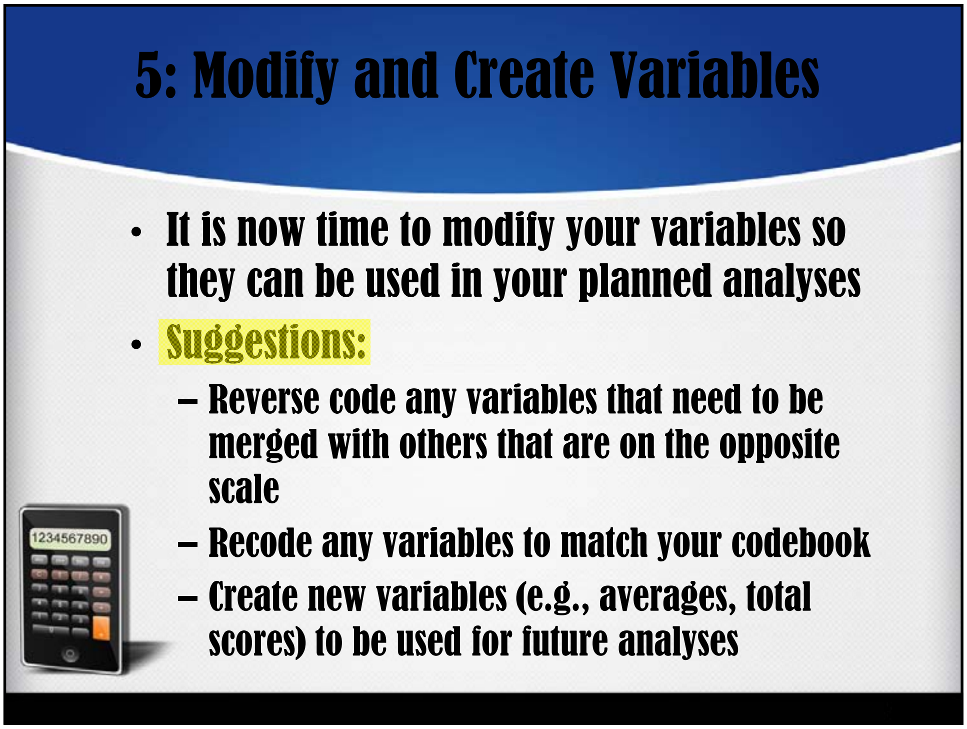
## 3: Perform Initial Frequencies – Round 1

- After organizing your codebook and analysis plan you can now begin to start the data cleaning process
- Conduct frequency analyses (frequencies, percentages) for EVERY variable in your evaluation dataset
- Suggestion:
  - request a graphic (bar chart or histogram) for each variable



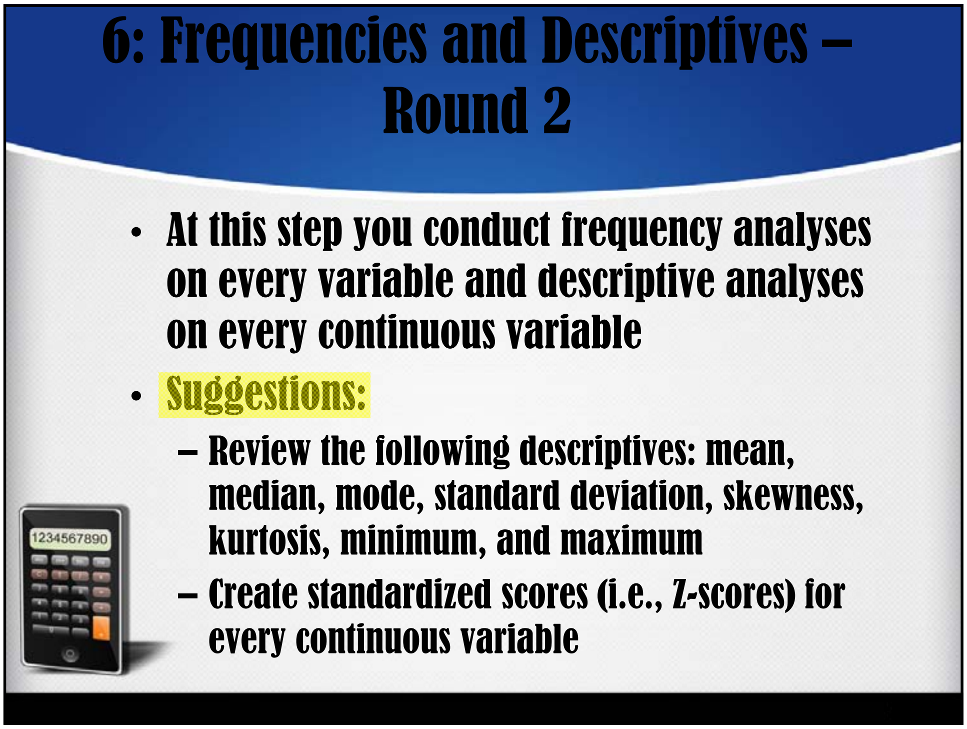
## 4: Check for Coding Mistakes

- Coding errors are any values that are not within the specified range for your variable (e.g., you have a rating scale from 1-5 and you have a value of 9)
- Suggestions:
  - Compare all values to what is listed in your codebook
  - In many cases errors are unspecified missing data values



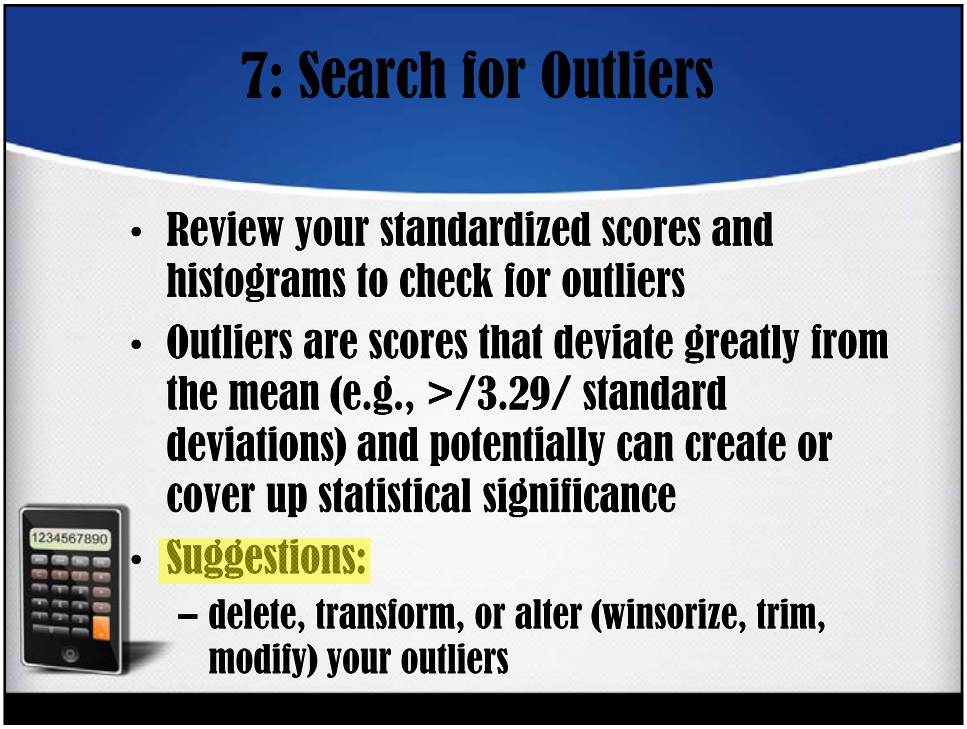
## 5: Modify and Create Variables

- It is now time to modify your variables so they can be used in your planned analyses
- Suggestions:
  - Reverse code any variables that need to be merged with others that are on the opposite scale
  - Recode any variables to match your codebook
  - Create new variables (e.g., averages, total scores) to be used for future analyses



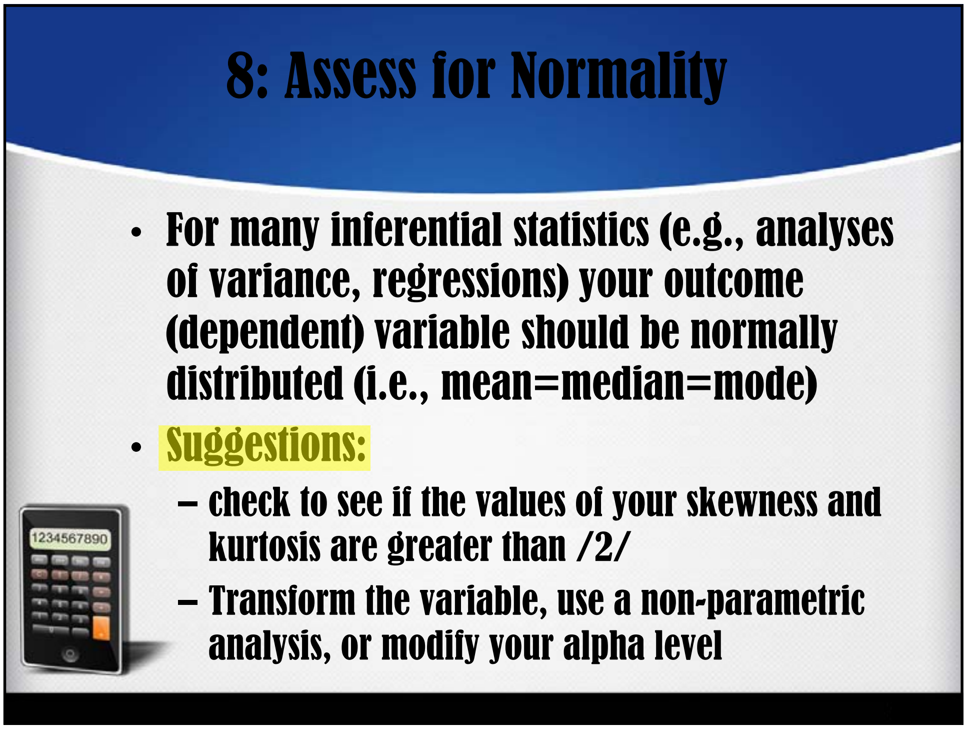
## 6: Frequencies and Descriptives – Round 2

- At this step you conduct frequency analyses on every variable and descriptive analyses on every continuous variable
- Suggestions:
  - Review the following descriptives: mean, median, mode, standard deviation, skewness, kurtosis, minimum, and maximum
  - Create standardized scores (i.e., *Z*-scores) for every continuous variable



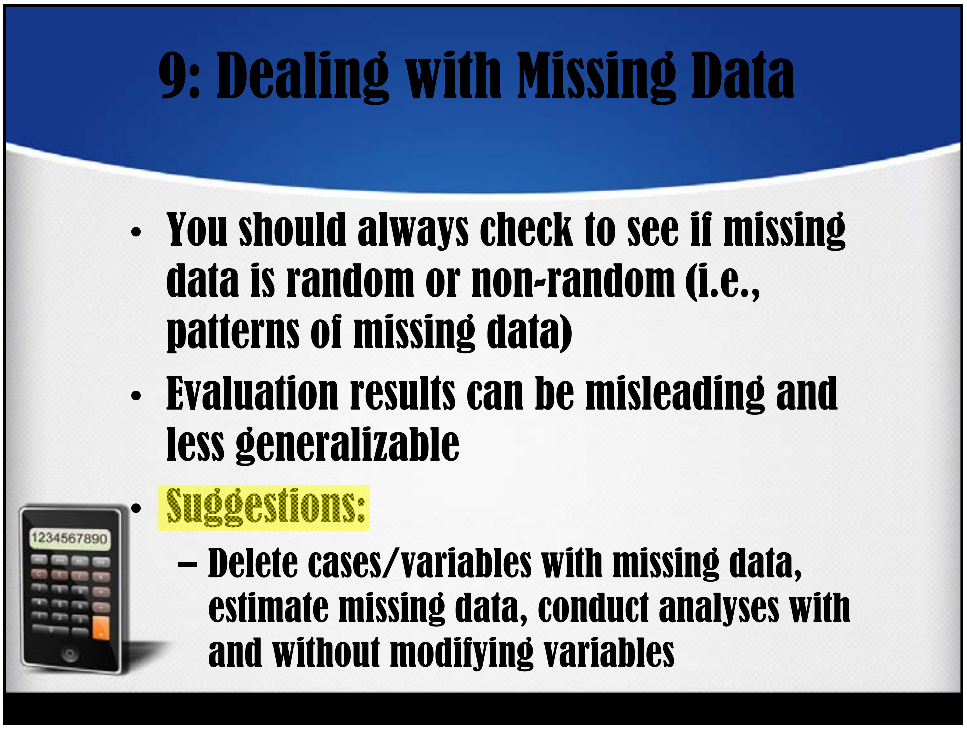
## 7: Search for Outliers

- Review your standardized scores and histograms to check for outliers
- Outliers are scores that deviate greatly from the mean (e.g., >/3.29/ standard deviations) and potentially can create or cover up statistical significance
- Suggestions:
  - delete, transform, or alter (winsorize, trim, modify) your outliers



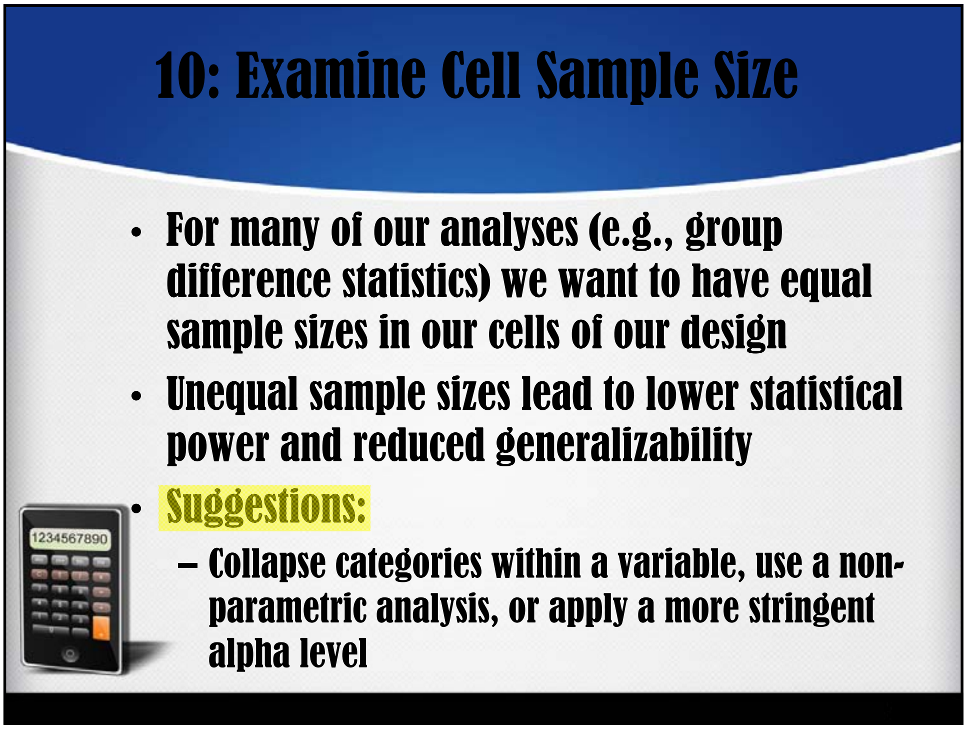
# 8: Assess for Normality

- For many inferential statistics (e.g., analyses of variance, regressions) your outcome (dependent) variable should be normally distributed (i.e., mean=median=mode)
- Suggestions:
  - check to see if the values of your skewness and kurtosis are greater than /2/
  - Transform the variable, use a non-parametric analysis, or modify your alpha level



# 9: Dealing with Missing Data

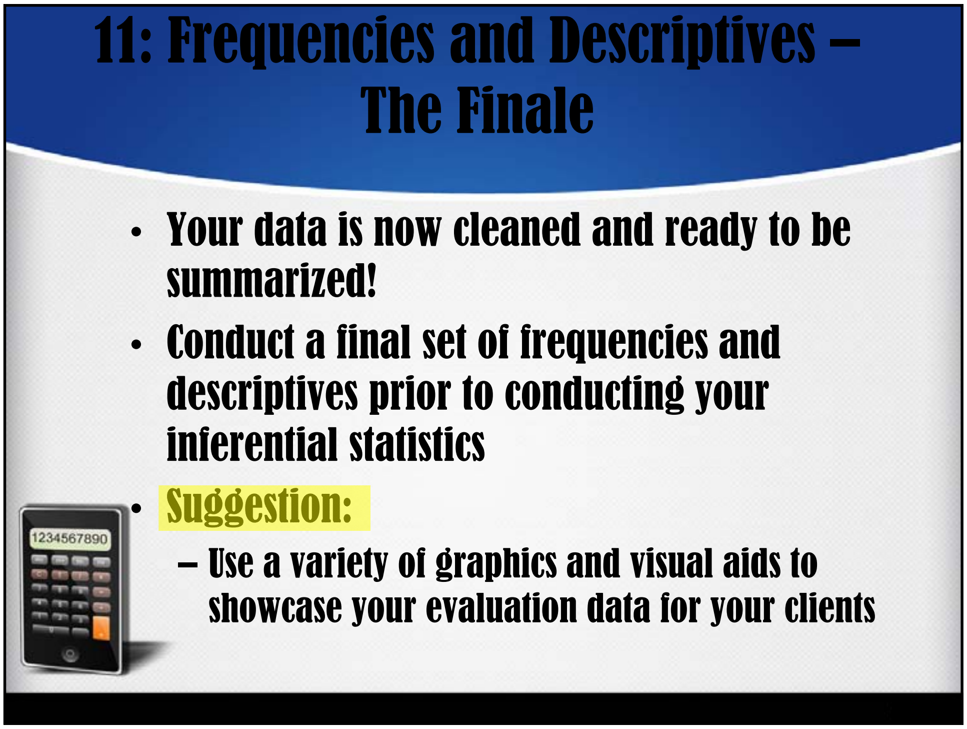
- You should always check to see if missing data is random or non-random (i.e., patterns of missing data)
- Evaluation results can be misleading and less generalizable
- Suggestions:
  - Delete cases/variables with missing data, estimate missing data, conduct analyses with and without modifying variables



## 10: Examine Cell Sample Size

- For many of our analyses (e.g., group difference statistics) we want to have equal sample sizes in our cells of our design
- Unequal sample sizes lead to lower statistical power and reduced generalizability
- Suggestions:
  - Collapse categories within a variable, use a non-parametric analysis, or apply a more stringent alpha level

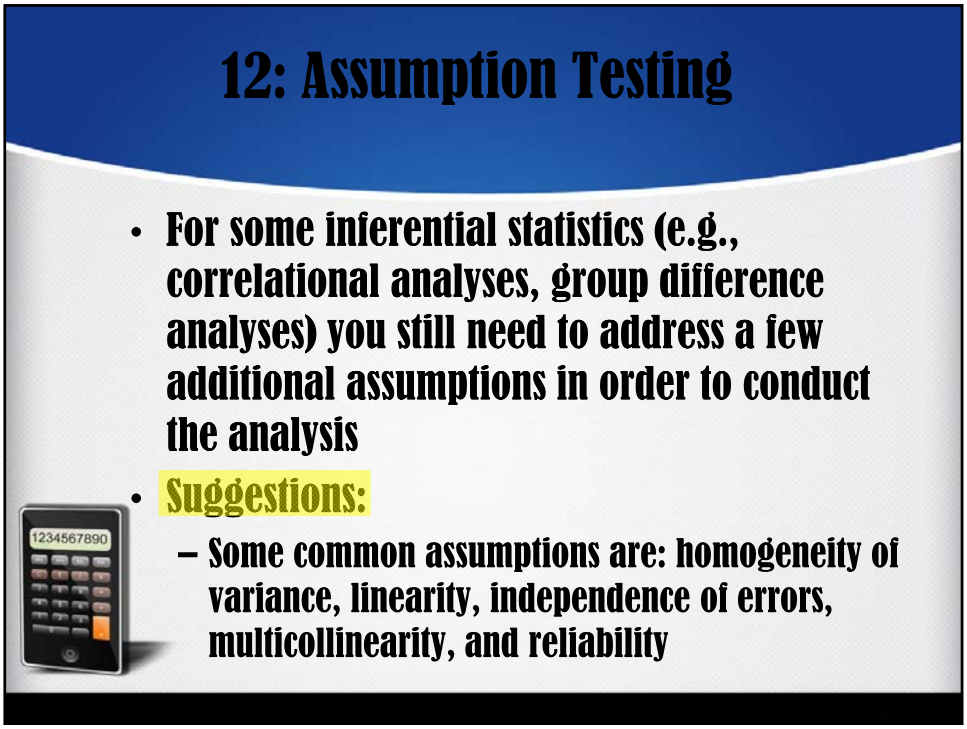
1234567890



## 11: Frequencies and Descriptives – The Finale

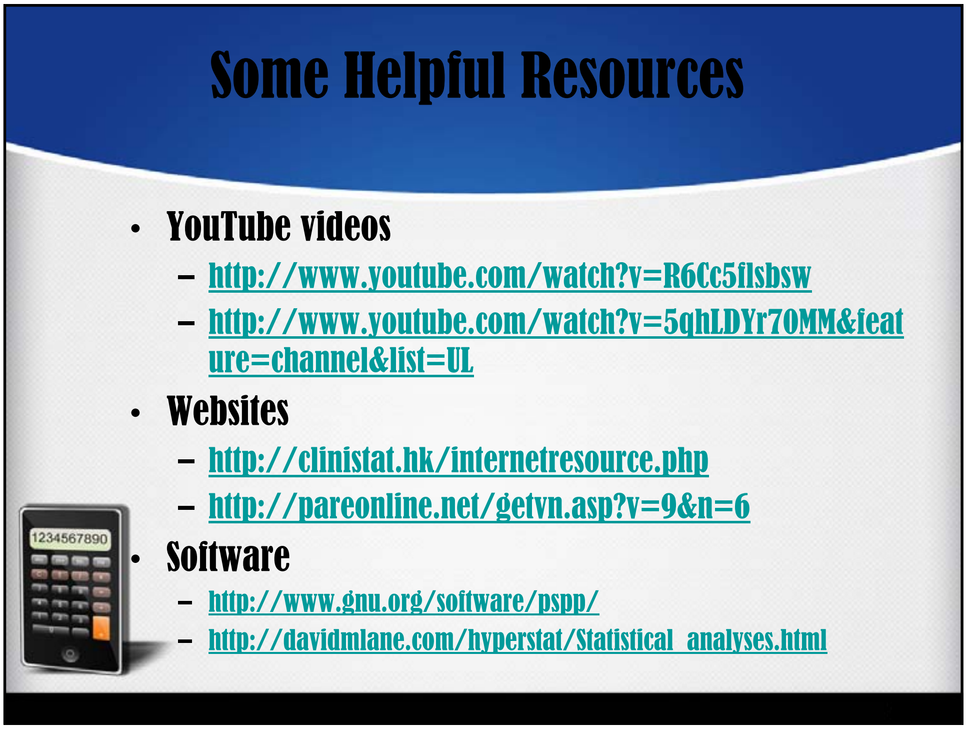
- Your data is now cleaned and ready to be summarized!
- Conduct a final set of frequencies and descriptives prior to conducting your inferential statistics
- Suggestion:
  - Use a variety of graphics and visual aids to showcase your evaluation data for your clients

1234567890



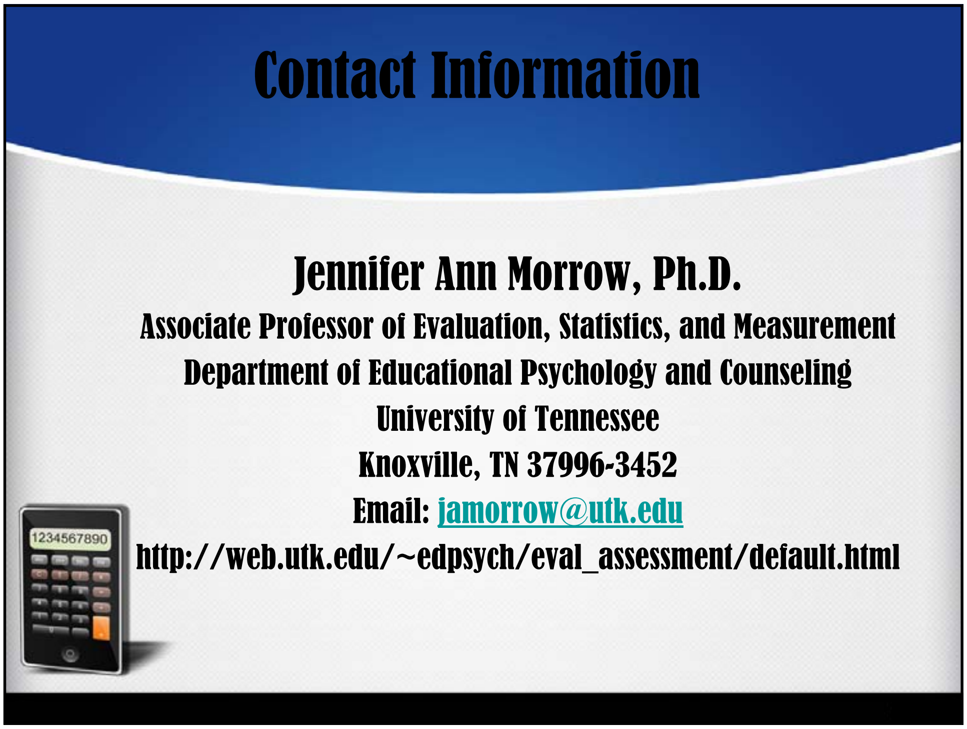
# 12: Assumption Testing

- For some inferential statistics (e.g., correlational analyses, group difference analyses) you still need to address a few additional assumptions in order to conduct the analysis
- Suggestions:
  - Some common assumptions are: homogeneity of variance, linearity, independence of errors, multicollinearity, and reliability



# Some Helpful Resources

- YouTube videos
  - http://www.youtube.com/watch?v=R6Cc5flsbsw
  - http://www.youtube.com/watch?v=5qhLDYr70MM&feature=channel&list=UL
- Websites
  - http://clinistat.hk/internetresource.php
  - http://pareonline.net/getvn.asp?v=9&n=6
- Software
  - http://www.gnu.org/software/pspp/
  - http://davidmlane.com/hyperstat/Statistical_analyses.html



## Contact Information

**Jennifer Ann Morrow, Ph.D.**

Associate Professor of Evaluation, Statistics, and Measurement
Department of Educational Psychology and Counseling
University of Tennessee
Knoxville, TN 37996-3452
Email: jamorrow@utk.edu
http://web.utk.edu/~edpsych/eval_assessment/default.html