The big picture focus of Exam 1 is: Given some data, you should know how to...

- Describe the basic characteristics of that data (number of cases/observations, types of variables, distributions of each variable, etc.)
- Describe relationships between variables, for any combination of variable (ie: categorical and categorical, quantitative and categorical, and quantitative and quantitative).
- Understand any multivariate relationships using stratification or regression to find conditional measures of association.

Specific topics you should be familiar with:

- **Data basics**
  - What is meant by the terms: "cases" (or "observations") and "variables" and how to find/describe them for "tidy data" (ie: data stored in a spread sheet format).
  - What is meant by "explanatory" and "response" variables and how to identify them when given a research question.
- **Data visualization** – The goal here is to *qualitatively* describe your data.
  - Visualizations of 1 variable
    - Categorical variable = bar chart
      - Describe the overall distribution (ie: be able to name the more common and less common categories)
    - Quantitative variable = histogram (or boxplot)
      - Describe the center, shape, spread, and unusual points.
  - Visualizations of 2 variables
    - 1 categorical variable and 1 quantitative variable = side-by-side boxplots or histogram
      - Be able to determine whether the variables are associated using differences in means/medians.
    - 2 categorical variables = stacked, clustered or conditional bar charts
      - Know how to identify and describe associations from these graphs.
      - Know the advantages of conditional vs. stacked/clustered.
    - 2 quantitative variables = scatter plot
      - Be able to describe the form, direction, and strength of an association.
- **Descriptive statistics** – The goal here is to *quantitatively* describe your data.
  - Descriptive statistics for 1 variable
    - 1 categorical variable = one-way tables (frequencies and/or proportions)
    - 1 quantitative variable
      - Descriptions of center = mean or median
      - Descriptions of spread = standard deviation, IQR, range
  - Descriptive statistics for 2 variables
    - 1 categorical and 1 quantitative = differences across groups, such as differences in means or medians
    - 2 categorical variables = contingency tables (two-way frequency tables)
      - Difference in conditional proportions

- • Risk Difference, risk ratio (relative risk), and odds ratio
  - ▪ 2 quantitative variables
    - • Correlation coefficient, either Pearson's or Spearman's rank
    - • Slope of the simple linear regression line
- **Multivariate Relationships –** The goal here is to properly account for things that might influence or obscure the relationship between an explanatory and response variable.
  - o **Confounding variables**
    - ▪ Know the definition of confounding (ie: third variable associated with both the explanatory and response variable)
    - ▪ Know when a third variable is/isn't a confounding variable based upon this definition.
  - o Controlling for confounding variables
    - ▪ Know how to conduct a stratified analysis to control for a categorical confounding variable.
      - • A stratified analysis is just repeating the same analysis you'd normally do for your explanatory and response variable multiple times for different subgroups of data defined by the confounding variable.
    - ▪ Know how to interpret the conditional effects of variables in a regression model that includes multiple explanatory variables.
      - • Understand how categorical variables are re-coded and used in regression models.
      - • Understand how/why coefficient estimates change when other variables are included/excluded from a model, and the related concept of omitted variable bias.