Data Analysis Basics – Part I

Judith A. Savageau

University of Massachusetts Medical School

Follow this and additional works at: https://escholarship.umassmed.edu/healthpolicy_pp

Part of the Health Services Administration Commons, Health Services Research Commons, and the Statistics and Probability Commons

Repository Citation

https://escholarship.umassmed.edu/healthpolicy_pp/206

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in Center for Health Policy and Research (CHPR) Publications by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Hello! I’m Judy Savageau from the Center for Health Policy and Research at UMass Medical School. A recent post from Pei-Pei Lei, my colleague in our Office of Survey Research, introduced some options for statistical programming in R. I wondered whether a basic introduction to statistics might be in order for those contemplating ‘where do I begin’, ‘what statistics do I need to compute’, and ‘how do I choose the appropriate statistical test’. While most AEA365 blogs don’t cover every topic in detail, perhaps a basic 2-part introduction will help here. Analyses are very different with qualitative versus quantitative data; thus, I’ve concentrated on the quantitative side of statistical computations.

**Hot Tip:**

Analyses fall into 3 general categories: descriptive, bivariate, and multivariate; they’re typically computed in that order as we:

- explore our data ([descriptive analyses](#)) with frequencies, percentile distributions, means, medians, and other measures of ‘central tendency’;
- begin to look at associations between an independent variable (e.g., age, gender, level of education) and an outcome variable (e.g., knowledge, attitudes, skills; [bivariate analyses](#)); and
- try to identify a set of factors that might be most ‘predictive’ of the outcome of interest ([multivariate analyses](#)).

**Hot Tip:**

The decision about what statistical test to use to describe data and its various relationships depends on the ‘nature’ of the data. Is it:

- **Categorical data:**
  - nominal: e.g., gender, race, ethnicity, smoking status, participation in a program: yes/no;
  - ordinal: e.g., a Likert-type scale score of 1=Strongly disagree to 5=Strongly agree or 5 levels of education: ‘Less than high school’, ‘High school graduate/GED’, ‘Some college/Associate degree’, ‘College graduate – 4-year program’, and ‘Post-graduate (Masters or PhD degree)’;
  - interval: ordinal data in fixed/equal-sized categories; e.g., age groups in 10-year intervals or salary in $25,000 intervals; or is it:
- **Continuous data:**
  - For example: age, years of education, days of school missed due to asthma exacerbations), etc.

Of course, data are often collected in one mode and then ‘collapsed’ for particular analyses (e.g., age recoded into meaningful age groups, Likert-type scales recoded as ‘agree’/’neutral’/‘disagree’).

**Hot Tip:**

Decisions must take into consideration whether the data are ‘normally distributed’ (i.e., is there ‘skewness’ in the data such that the values for age are mostly in persons under 45 though you have a small number of people who are in their 60’s, 70’s, and 80’s?). Most statistical tests have a number of underlying assumptions that one must meet – all starting with data being normally distributed. Thus, one typically begins looking descriptively at their data: frequencies and
percentile distributions, means, medians, and standard deviations. Sometimes, graphing the data shows the ‘devil in the detail’ with regard to how data are distributed. There are some statistics one can compute to measure the degree of skewness in the data and whether distributions are significantly different from ‘normal’. And, if the data are not normally distributed, there are several non-parametric statistics that can be computed to take this into account.

Tomorrow’s post will focus on bivariate and multivariable statistics. Stay tuned!