STAT 2120: Notes on Topic 2

Least-squares regression:

- A regression line describes a one-way linear relationship between variables.
  - An explanatory variable, x, "explains" variability in a response variable, y.
  - Often one wants to predict y from a given x. Such a prediction is denoted $\hat{y}$.
- The least-squares regression line makes the sum of squared-prediction errors as small as possible.
  - A prediction error is the vertical distance between a given point and a regression line.
  - The formula for the least-squares regression line is $\hat{y} = b_0 + b_1 x$, with "slope" $b_1 = r\frac{s_y}{s_x}$ and "intercept" $b_0 = \bar{y} - b_1\bar{x}$. Predictions are made by plugging in values of x.
  - Slope, b_1, is the amount of change in $\hat{y}$ when x increases by one unit. Intercept, b_0, is the prediction at $x = 0$.
  - Calculate b_0 and b_1 by computer.
- Properties of the least-squares regression line:
  - Interchanging x and y modifies the formulation.
  - The line $\hat{y} = b_0 + b_1 x$ always passes through the point $(\bar{x}, \bar{y})$.
  - The slope formula $b_1 = r\frac{s_y}{s_x}$ interprets the relationship in units of s_x and s_y through r.
  - Similarly, r^2 measures the proportion of variability in y that is explained by x.
- The residuals describe the leftover variation in y after fitting the least-squares regression line.
  - Each residual is defined by $y - \hat{y}$.
  - The average of the residuals is zero.
  - Analysis of residuals helps to assess the suitability of a linear relationship.
  - A residual plot is a scatterplot of residuals against the values of x.
  - The ideal residual plot should exhibit no systematic pattern; patterns indicating a departure from the linear relationship are: curvature, trends in spread, outliers in the residuals.
  - An outlier in y corresponds with an outlier in the residuals. Such is observed as an observation that outside of the overall pattern of the relationship.
- Influential observations are those whose individual deletion would have a strong impact on the regression line.
  - An influential observation is often an outlier in x, but may not be an outlier in y.

Cautions about correlation and regression:

- Basic cautions:
  - Correlation is for two-way relationships, regression for one-way relationships.
  - Only relevant for linear relationships.
  - Neither is resistant.
- Extrapolation is when predictions are made outside the range of data.
  - Often untrustworthy since the linear relationship may not hold for x-values far outside those observed.
- Correlation calculated on "averaged" data is higher than that calculated on individuals.
- The relationship between two variables may be influenced by a third, "lurking" variable that is not observed.
  - Lurking variables may influence relationships between any type of variables, quantitative or categorical.
- Association is not causation.
  - An observed association may reflect the influence of a causal lurking variable. Such is called a "nonsense correlation."
  - An experiment that controls lurking variables is best for establishing causation.
  - It is possible to establish causation without performing an experiment that controls for lurking variables, but the evidence that arises is weaker.

Relationships in categorical data:

- Relationships in categorical data are explored by compiling variables in two-way tables.
  - A two-way table involves a row variable and a column variable.
  - A two-way table may record counts or percentages. Percentages are most useful because they are easy to compare in the form of distributions.
- Relationships are described through specialized distributions appearing in the table.
  - Bar graphs provide a useful means of presenting the relevant distributions.
  - The distributions of the row and column variables appear in the margins of the table, and are called marginal distributions. Given as counts they are called row and column totals.
  - A conditional distribution is calculated from the counts of one variable limited to a given category of the other variable.

- An association may be described by examining the conditional distributions of one variable across the categories of the other variable. Typically, the former would be the response variable and the latter the explanatory variable.
- Lurking variables may give rise to Simpson's paradox: patterns seen in individual categories are reversed in the patterns of the combined data.
  - A lurking variable may arise when a three-way table is "aggregated" into a two-way table.

Introduction to producing data:

- Designing the production of data allows data analysis to answer specific questions.
  - Data are produced on a small scale, and the intent is to generalize to a wider scale. Answering questions this way (with "confidence") is statistical inference.
  - Anecdotal evidence arises haphazardly and may not represent any relevant group of cases.
  - "Available data" arise for other purposes, but may help to answer present questions.
  - In a sample survey, a sample of cases is drawn from a population. (A census is a sample that consists of the entire population.)
- Confounding arises between explanatory variables when their relationships with the response are indistinguishable.
- An intervention is where one imposes a change in the conditions of data-production.
- In an observational study, individuals are observed, but no attempt is made to control the conditions of data-production.
  - Observational studies are often plagued by confounding between an observed variable and an unobserved lurking variable.
- In an experiment, the conditions of data-production are controlled by applying treatments to individuals.
  - One objective in designing an experiment is to avoid confounding between explanatory variables.
- Controlling data-production may involve questions of ethics.

Designing a sample:

- The key elements of a sampling study:
  - A population is a collection of individuals about which we want information and the conclusions of statistical inference are to be relevant.
  - A sampling frame lists the individuals in the population.
  - A sample is the subset of a population on which data are measured and put to analysis.
  - The response rate is the proportion of measured individuals in a preliminary sample.
  - The design of a sample refers to the method used to select it from the population.
- A sampling design is biased if it systematically favors certain portions of the population over others.
- Examples of biased sampling designs:
  - A voluntary sample arises when individuals are self-selected for the sample by responding to an incentive.
  - A convenience sample arises when selection for the sample is determined by the convenience of the selection-maker.
- Simple random sampling (SRS) selects a sample randomly, in such a way that every fixed-size subset has an equal probability of being selected.
  - Bias is avoided in SRS by its use of chance.
  - SRS may be carried out by drawing labels from a hat, or by simulating that procedure with computer software or a table of random digits.
- A probability sample is a sample selected by chance (through the use of probability sampling), based on known selection probabilities of each sample.
  - SRS is an example of unbiased probability sampling.
  - In general, bias may be accommodated in probability sampling using knowledge of the selection probabilities.
  - Stratified random sampling is an example of probability sampling in which simple random samples are drawn in distinct strata and aggregated.
  - Multistage sampling is an example of probability sampling that is carried out in stages. At each stage, each in a current list of sampling units is narrowed to a list of more refined sampling units, and a SRS of those units is selected. The final list of sampling units is of individuals.
- Bias in sampling may originate from sources other than the sampling design, including:
  - Under-coverage in the list of individuals in the population.
  - Non-response of individuals selected for the sample.
  - Inaccurate responses of the respondent, which leads to response bias. This may be unintentionally encouraged by the interviewer.
  - Poor wording and design of a questionnaire.

Designing experiments:

- In an experiment, a response variable is observed under controlled conditions that reflect carefully chosen values of explanatory variables.
- Terminology associated with experiments is:
  - Individuals are referred to as subjects, or experimental units when they are non-human.
  - Explanatory variables are referred to as factors.
  - Each specific value of an explanatory variable is referred to as the level of a factor. It reflects the application of a treatment used to modify the experimental conditions in a specific way.
- Experiments provide: focus on interesting treatments by holding uninteresting factors steady; simultaneous study of multiple factors.
- The basic principles of experimental design are:
  - Control the effects of lurking variables by using comparisons.
  - Allocate subjects among treatments by using randomization.
  - Reduce random variability by using replication.
- In randomized experiments, each subject is assigned a treatment by chance.
  - A completely randomized design allocates subjects to treatments with equal probability.
  - Randomization might be implemented using techniques similar to those used in SRS to select a sample.
- In comparative experiments, the influence of lurking variables is canceled by comparing treatments against each other
  - A control group is a sham treatment (also called a placebo) used as a baseline comparison.
- Randomized comparative experiments may establish that treatment differences cause patterns seen in the response variable.
  - The underlying context is that random assignment homogenizes subjects applied the same treatment, while comparison cancels the influence of lurking variables. Patterns in the response must therefore be due to the random variability of treatment assignments or the causal effect of the treatments.
  - Patterns in the response are said to be statistically significant if they are of such magnitude that they would rarely be observed by chance.
- The efficacy of randomized, comparative experimentation may be undermined by such issues as:
  - Unconscious bias of the experimenter or subject, which might be avoided by double-blinding knowledge of the treatment assignments from both the experimenter and the subject.
  - Lack of realism in the subjects, treatments, or the experimental setting.
- A block design randomly allocates subjects across treatments multiple times in separate predefined blocks of subjects.
  - Blocks are formulated so that the relevant characteristics of subjects within the same block are roughly the same.
  - A matched-pairs design is a block design with two treatments and blocks of two subjects each.
  - Block designs employ the basic principles of experimental design.