STAT 2120: Notes on Topic 1

Introduction to Examining Distributions:

- A variable records characteristics of cases (*i.e.*, objects of interest) in its values.
- Classify a variable by its possible values:
  - Categorical: records group labels; numeric labels mean nothing, except possible order.
  - Quantitative: records meaningful numbers; may be discrete or continuous
- A time series is a record of values across time.
- A variable's distribution describes the counts or relative proportions of its values.
- Exploratory data analysis seeks to describe distributions and relationships in data.

Displaying a distribution with graphs:

- Bar graphs and pie charts describe the distribution of a categorical variable.
  - Bar graphs emphasize counts; pie charts, proportions.
  - A Pareto chart is a bar graph with categories ordered by decreasing frequency.
- Histograms are essentially bar graphs of a quantitative variable.
  - Bar-widths are not absolute; use equal bar-widths and "eyeball" for best picture.
  - Look for overall pattern, shape center, spread, deviations in shape, and "outlier" deviations.
  - A symmetric distribution is such that its histogram mirrors itself about its center.
  - A right- or left-skewed distribution shows a long tail to the right or left in its histogram.
- Stemplots are back-of-the-envelope histograms drawn with the digits of quantitative values
  - "Stem" digits define bars; "leaf" digits display counts and sub-counts.
  - Customize by rounding digits and splitting stems.
- Time plots graph time series values by time.
  - Emphasize patterns of change over time, such as trends and seasonal variations.

Describing distributions with numbers:

- Denote by $x_1, \ldots, x_n$ the values of n observations.
- p^{th} percentile is a number such that p percent of values fall on or below.
- Describe a distribution with numerical summaries of shape, center, and spread.
- A summary is resistant if it is insensitive to changes in skewness or extreme values.
- Measure of center: mean, $\bar{x}$
  - $\bar{x} = \frac{1}{n}\sum x_i$, the arithmetic average.
  - $\bar{x}$ is not resistant.
- Measure of center: median, M
  - M is the 50^{th} percentile.
  - Calculate as the middle value or average of two middle values.
  - M is resistant.
- Measure of spread: extreme values
  - Smallest and largest values
  - Extreme values are not resistant.
- Measure of spread: quartiles, Q_1 and Q_2
  - Q_1 is the 25^{th} percentile; Q_2 is the 75^{th} percentile
  - Calculate Q_1 and Q_2 as medians of values falling to the left or right of (but not on) M.
  - Q_1 and Q_2 are resistant.
- Measure of spread: standard deviation, s
  - $s = \sqrt{s^2}$, where $s^2 = \frac{1}{n-1}\sum(x_i - \bar{x})^2$, a rescaled average of squared-deviations from $\bar{x}$.
  - s^2 is the variance; $n-1$ is "degrees of freedom;" square-root to match units with x_i.
  - Calculate by computer.
  - s is not resistant.
- Measure of shape: mean-median comparisons:
  - $\bar{x} \approx M$ if the distribution is symmetric; $M < \bar{x}$ if right-skewed; $\bar{x} < M$ if left-skewed
- Useful descriptions of a distribution:
  - Summarize center and spread, *e.g.*, as $\bar{x}$ and s (for symmetric, outlier-free distributions).
  - Display $\bar{x}$ and s graphically as "error bars."
  - Five-number summary: smallest extreme, Q_1, M, Q_2, largest extreme.
  - Display the five-number summary graphically as a box plot.

Normal distributions:

- A density curve is an idealization for describing patterns seen in histograms.
  - Denote by x a variable representing an idealized observation.
  - "Area under the curve" in a range represents the proportion of observations in that range.
  - Total "area under the curve" is one.
  - The median is the point that divides "area under the curve" equally to the left and right.
  - Denote by μ and σ idealizations of $\bar{x}$ and s formulated on a density curve.
  - μ is the balance point.
- Normal distributions are described by the class of density curves called "Normal curves."
  - Symmetric, single-peaked, and bell-shaped.
  - Indexed by μ and σ, denoted $N(\mu, \sigma)$.
  - $\mu \pm \sigma$ mark a Normal curve's inflection points.

- 68-95-99.7 rule: For observations having a Normal distribution: 68% fall within $\mu \pm \sigma$; 95% fall within $\mu \pm 2\sigma$; and 99.7% fall within $\mu \pm 3\sigma$.
- The standard normal distribution is $N(0,1)$
- Suppose x has a distribution with mean μ and standard deviation σ. The z-score, or standardized value, of x is $z = (x - \mu)/\sigma$.
- Measures "location from μ in units of σ."
- If x is $N(\mu, \sigma)$ then z is $N(0,1)$.
- To calculate an "area under the curve" for $N(\mu, \sigma)$ translate to a z-score and use $N(0,1)$.
- Calculations involving $N(\mu, \sigma)$ might be forward (What proportion p has $x \leq c$?) or backward (For what c is the proportion of $x \leq c$ equal to p?)
- A Normal quantile plot is a graph of percentiles of $x_1, \ldots, x_n$ plotted against those of $N(0,1)$.
- Plots on a straight line indicate a Normal distribution.
- Calculate by computer.

Introduction to Examining Relationships :

- Approach: plot data, calculate summaries; look for patterns and deviations; consider idealizations
- An explanatory (or independent) variable explains variability in the response (or dependent) variable.
- Scatterplots: graph two quantitative variables measured on the same set of individuals.
- Look for overall pattern; general deviations, "outlier" deviations.
- Scatterplots are sometimes "smoothed" using algorithms that fit curves to the data.
- A transformation (*e.g.*, the log transformation) is sometimes applied to skewed data.
- A scatterplot be extended by adding categorical variables, color- or symbol-coded.
- The overall pattern of a relationship:
- The form of a relationship may involve linear patterns, clusters, or lack of any pattern.
- The direction may be positive or negative.
- A stronger relationship is observed as points falling more closely to a clear from.

Correlation:

- Measure of direction and strength: correlation, r
- $r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$, the rescaled average of the product of standardized deviations from $\bar{x}$ and $\bar{y}$.
- Calculate by computer.
- Interprets only linear relationships.
- Response and explanatory variables are interchangeable
- Unitless, and independent of variables' units.
- $r > 0$ indicates a positive relationship, $r < 0$ a negative relationship.
- $-1 \leq r \leq 1$, always.
- Stronger relationships are indicated by values farther from 0; $r = \pm 1$ is a perfect relationship with points on a straight line.
- s is not resistant.
- Report r together with $\bar{x}$, s_x, $\bar{y}$, s_y to summarize two-variable data.

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.