YMS.4e CONTENT REVIEW
(Chapters 1-3)

I. Exploring Data

A. Analyzing Categorical Data

1. Categorical Data – nominal scale, names (e.g. male/female or eye color or breeds of dogs)

2. Displaying distributions
   Bar graphs (bars do not touch)
   Pie charts (percentages must sum to 100%)

3. From a two-way table of counts, find marginal and categorical distributions (in percents), describe relationship between two categorical variables by comparing percents and be able to recognize/explain Simpson's paradox

B. Displaying Quantitative (one variable univariate) Data with Graphs

1. Quantitative Data – rational scale, numbers where an average can be calculated (e.g. weights of hamsters or amounts of chemicals in beverages)

2. Displaying distributions
   Dot plots – can resemble probability curves
   Stem (& leaf) plots – remember to put in the key (e.g. 8|2 means 82 mg. of salt)
      Split stems if too many data points
      Back–to–back for comparison of two samples
   Histogram – put // for breaks in axis, use no fewer than 5 classes (bars)

3. Describe distributions using SOCS (Shape, Outliers, Center and Spread)

C. Describing Quantitative Data with Numbers (**1-variable stats**)

1. Measures of central tendency (center)
      mean ($\bar{x}, \mu$)
      median (middle)
      mode (most)

2. Measures of dispersion (spread)
      range (max – min)
      quartile (25% = Q_1, 75% = Q_3)
      interquartile range ($Q_3 - Q_1$)
      variance $s^2 = \dfrac{\sum (x_i - \bar{x})^2}{n-1} or \sigma^2 = \dfrac{\sum (x_i - \mu)^2}{n}$
      standard deviation (s, σ) = square root of variance

3. Mean, range, variance, and standard deviation are non-resistant measures (strongly influenced by outliers). Use mean and standard deviation with approximately normal distributions; use median and IQR for skewed distributions (where the mean chases the tail).

4. 5-number summary (min, Q_1, M, Q_3, max) shown using *boxplots* (modified shows outliers)

II. Modeling Distributions of Data

A. Describing Location in a Distribution

1. *Ogive* – cumulative relative frequency plot

2. Adding a constant *a* to a data set increases mean by *a* but has no effect on standard deviation; multiplying a constant *b* to a data set multiplies the mean and standard deviation by *b*

3. Density curves are models which smooths out irregularities of actual data (use μ, σ) where area under curve always equals 1

B. Normal Distributions

1. If the mean = 0 and the standard deviation = 1, this is a standard, normal curve

2. Use with z scores (standard scores), z = $\frac{x-\mu}{\sigma}$, where +z are scores are above the mean and –z are scores below the mean.

3. To compare two observations from different circumstances, find the z score of each, then compare

4. Use z scores to find the *p* value, the probability (or proportion or percent) of the data that lies under a portion of the bell curve, *p* values represent area under the curve. Use **normalcdf** (to find the *p* value), or **invnorm** (to find z score) on the calculator

5. ALWAYS DRAW THE CURVE and shade to show your area.

6. 68% – 95% – 99.7% rule for area under the curve

7. To assess Normality for a given data set, graph the data, apply the 68-95-99.7 rule, compare the mean/median and construct a Normal Probability Plot

III. Describing Relationships

A. Scatterplots and Correlation

1. To graph two variable (bivariate) data – DATA MUST BE QUANTITATIVE. Graph the explanatory variable (independent) on the x axis, the response variable (dependent) on the y axis

2. Scatterplots look for relationships between the variables.

3. Look for clusters of points and gaps. Two clusters indicate that the data should be analyzed to find reasons for the clusters.

4. If the points are scattered, draw an ellipse around the plot. The more elongated, the stronger the linear relationship. Sketch the major axis of the ellipse. This is a good model of the linear regression line.

5. Linear correlation coefficient (r) – measures the strength of the linear relationship $-1 \le r \le 1$

   r = 0 indicates no relationship (the ellipse is a perfect circle)
   –r indicates an inverse relationship
   r is a non–resistant measure (outliers strongly affect r)

   $$r = \frac{1}{n-1}\sum\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$ (**2–variable stats**)

6. *Correlation does not imply causation.* Only a well-designed, controlled experiment may establish causation

B. Least-Squares Regression

1. Least squares regression line (LSRL) – used for prediction; minimizes the vertical distances from each data point to the line drawn. (**Linreg a+bx**)

   y varies with respect to *x*, so choose the explanatory and response axes carefully (*y* is dependent on *x*)

   $\hat{y} =$ predicted y value

   $\hat{y} = a + bx$ is the equation of the LSRL where $b = r\left(\frac{s_y}{s_x}\right)$, $a = \bar{y} - b\bar{x}$ and the point $(\bar{x}, \bar{y})$ is always on the line.

   Do not *extrapolate* (predict a *y* value when the *x* value is far from the other *x* values).

2. Coefficient of Determination (r^2) – gives the proportion (%) of variation in the values of y that can be explained by the regression line. The better the line fits, the higher the value of r^2.

   To judge "fit of the line" look at r and r^2. If $r = 0.7$, then $r^2 = .49$, so about half the variation in y is accounted for by the least squares regression line.

3. Residual = observed y value – predicted y value ($y - \hat{y}$); residuals sum to 0

4. Residual plot – scatterplot of (x, residuals)
   no pattern $\rightarrow$ good linear relationship,
   curved pattern $\rightarrow$ no linear relationship,

5. Outliers are y values far from the regression line (have large residuals)

6. Influential points are x values far from the regression line (may have small residuals) which significantly change the LSRL slope