STAT381 **Lia Liu**

Sections 1.1-1.4

List of topics: Describing data

1. **Population vs. sample**
2. Given a set of 1-d data: $x_1, x_2, ..., x_n$
3. **Stem-and-leaf plot**: To arrange data in increasing/decreasing order.
4. **5 number summary**: min, Q_1, Med, Q_3, max
5. **Inter-quartile range**= $Q_3 - Q_1$
6. **Boxplot**
7. **Outlier**: 1.5IQR rule
8. **Histogram**: frequency, relative frequency
   - Rule of thumb: number of classes $\approx \sqrt{n}$
   - Describe center: **mean(average), median, mode**
   - Describe dispersion:
     population variance $\sigma^2 = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \overline{x})^2$,
     sample variance $s^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \overline{x})^2$,
     population standard deviation σ,
     sample standard deviation s,
   - Median and IQR are more robust(less influenced by extreme values) than mean and standard deviation.
   - How to use TI-83/84 to compute and display data?

1.5

List of topics:

1. **Always draw scatter plot! Look for direction, form, strength and outlier.**
2. **Residual=data-model:** $e_i = y_i - \hat{y}_i$
3. **Linear regression:** $\hat{y} = a+bx$**, where** $b = r\frac{s_y}{s_x},\ a = \overline{y}-b\overline{x}$
4. **Correlation coefficient:** $r = \dfrac{\Sigma z_x z_y}{n-1} =$
   $= \dfrac{1}{n-1}\Sigma_{i=1}^{n}(\dfrac{x_i-\overline{x}}{s_x})(\dfrac{y_i-\overline{y}}{s_y})$
5. **Where is the line come from? Least square method.**
   Ex: Given paired data: $(x_i, y_i), i = 1, 2, 3, ..., n.$
   Step 1: Choose a model to fit the curve according to the scatter plot.
   Step 2: If the cloud looks linear, Choose linear model:
   $\hat{y} = a + bx$**.**
   Minimize $\Sigma_{i=1}^{n} e_i^2$**, which is** $\Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 = \Sigma_{i=1}^{n}(y_i - a - bx_i)^2$**. Note: it's a function of two variables** a **and** b**.**
   Let's call it $F(a, b)$**. Then** $F(a, b) = \Sigma_{i=1}^{n} e_i^2$
   $= \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2 = \Sigma_{i=1}^{n}(y_i - a - bx_i)^2$
   Recall from Calc III, to minimize a function of two variables, we need to take partial derivatives and make them equal to 0. Solve the system of equations. That's how a **and** b **are obtained.**

Because the line came from minimizing the sum of residual squared, and the residual is the data y - model y, if you switch x with y pair, you get totally a different line.

6. **Outliers: Influential points, large residual or high leverage.**
7. **How to tell a data point is an outlier? Find a linear regression line and correlation coefficient with and without the point and compare.**
8. **What can go wrong?**
    - **Don't fit a straight line to a nonlinear relationship;**
    - **Be aware of the effect of outliers, leverage, and influence;**
    - **Don't invert regression;**
    - **Be careful with extrapolation;**
    - **Don't infer that x causes y;**
    - **Be ware of lurking variable.**