# INFO 1301

Prof. Michael Paul
Prof. William Aspray

Friday, September 16, 2016

# Topics – Additional Topics in Descriptive Statistics

- Percentiles
- Box plots
- Outliers
- Euclidean distance and its use in statistics

# Percentiles

- Sometimes you want to know how a particular data point stands in comparison to the entire data set. For ordinal variables, we can create percentiles
- Example: SAT scores are given in percentiles. Thus if you are in the $90^{th}$ percentile on a given variable (e.g. you SAT general math aptitude test), that means your value is higher than 90% of the people who took the test at the same time.
- Of particular interest are the $25^{th}$ and $75^{th}$ percentiles because they are at the halfway point between the median and the lowest or highest value, respectively (not in actual score but in how many data points are higher or lower)
- Example: For the set 2,2,2,4,10 the median=?, $P_{75}$=? But the average between the median and the max = ?, $P_{25}$= ?

# Box Plot (1)

- A box plot summarizes a data set using five statistics plus some additional comments about unusual circumstances.
- Let's consider a data set including 1,1,2,3,5,8,13,21,34, and some more numbers.
- To create a box plot:

0. Plot the data along a vertical axis, suitably spaced.
1. Calculate the median and draw a dark horizontal line
2. Calculate the $25^{th}$ and $75^{th}$ percentiles ($P_{25}$ and $P_{75}$) and use them as the lower and upper edges of a box that represents the middle 50% of the data (The $25^{th}$ percentile is the median of all data points below the median. Similarly for the $75^{th}$ Percentile. Common terms are the first, second, third, and fourth quartiles.)
3. Calculate the Interquartile Range (IQR) = $P_{75}$-$P_{25}$.

# Box Plot (2)

4. Calculate maximum and minimum whisker reach.

    Max = $P_{75}$+ 1/5(IQR)

The whiskers are intended to reach out a little above the third quartile and below the first quartile – to capture more than the half of the data that is captured by the interquartile range.

5. Identify specific outliers, i.e. data points above the maximum whisker reach or below the minimum whisker reach.

[See the visualization of a box plot in Figure 1.26 (p. 35)]

# Euclidean Distance

- The distance in the x-y (2-dimensional) plane between points $(a_1,a_2)$ and $(b_1,b_2)$ is given by the Pythagorean Theorem: $[(a_1-b_1)^2 + (a_2-b_2)^2]^{.5}$
- This generalizes to n-dimensional space, where the distance between points $(a_1,a_2,\ldots,a_n)$ and $(b_1,b_2,\ldots,b_n)$ is given by $[(a_1-b_1)^2 + (a_2-b_2)^2+\ldots+(a_n-b_n)^2]^{.5}$
- This is called the Euclidean distance, and it appears in several important places in basic statistics:
  - In (Pearson) correlation [remember the number r, where $-1\leq r\leq 1$, that was given by Minitab Express to represent how closely two variables were associated?]
  - In hypothesis testing ($\chi^2$ [pronounced "chi squared"] distributions)
- More on both of these topics later in the semester

# The Hidden Life of Data Points

- Alice Marwick, “I’m More Interesting than my Friendster profile.”
- Consider a data set with the following variables: name, age (years), savings (pennies), weight (pounds), GPA, shoe size (US)
- These 6 variables tell us a lot but not everything about the person
- What the data set tells us about each data point (each person) can be represented as a vector with 6 dimensions $<a_1,a_2,\ldots,a_6>$, which are the values in the row of the database representing that person.
- Remember that order matters in a vector.
- Then the Euclidean distance can be used to find the distance between two rows (i.e. the distance between two data points in the database)
- This is then incorporated into the analysis of how any two variables are associated (correlated).

# Scale Matters!

- Note in the last database that:
  - Major changes in savings or weight are represented by large numerical changes in those variables; while
  - Major changes in shoe size or GPA are represented by small numerical changes in those variables.
  - E.g. a change in weight of 2 lbs. is not large but a change in GPA of 2 pts is enormous.
- Statisticians scale the variables to make the change in numerical value of each variable proportional to the amount of change.
- This is why $-1 \leq r \leq 1$ in Pearson correlation.
- We will come back and see some of the details later in the semester.

# In-class Exercise

- Watch the Arthur Benjamin youtube video on Fibonacci numbers
- Using the baseball statistics that you captured in Minitab Express, draw a BoxPlot [graphs/boxplot/simple]
- Compare it to the boxplots of other students. See if you can explain their differences.