Math 114 ACTIVITY 3: Numerically summarizing data – Measures of Position

Why

The mean and standard deviation are particularly useful for describing symmetrically distributed variables. Another commonly used set of summary values is the *positional* descriptions - cutoffs below which we find a quarter of the data, or forty percent, or any other selected fraction. These are particularly useful with non-symmetric distributions, since they can give a sense of the shape – they are in common use for describing the location of individual values within a distribution and appear frequently in medical and educational descriptions (especially related to development). You have already seen examples – the maximum, minimum and median of a set of values are all positional measures.

LEARNING OBJECTIVES

1. Learn the most common measures of position – and how to determine them in a set of values
2. Learn to read shape information from a set of positional measures s
3. Continue developing skills in working with teams.

CITERIA

1. Success in working as a team and in fulfilling the team roles.
2. Success in involving all members of the team in the conversation.
3. Success in completing the exercises.

RESOURCES

1. The course syllabus
2. The team role desk markers (handed out in class for use during the semester)
3. Your text - especially section 3.4 and 3.5
4. Your calculator
5. 40 minutes

PLAN

1. Select roles, if you have not already done so, and decide how you will carry out steps 2 and 3 (5 minutes)
2. Work through the exercises given here - be sure everyone understands all results & procedures(30 minutes)
3. Assess the team's work and roles performances and prepare the Reflector's and Recorder's reports including team grade (5 minutes).

TERMINOLOGY

Note: In any small data set (any set we are likely to work with by hand) we are very likely to be forced to "split the difference" between two values to estimate a positional value (as you have seen with the median). We will take the mean of the two values closest above and below the location [some people use more complicated procedures which we will not duplicate]

The *median* of a collection of numbers is the number in the middle when the numbers are sorted in order. Half the values are less than or equal to (to the left, in a histogram) the median, half the values are greater than or equal. [If there is an even number of values, we split the difference between the two middle values in the list]. It is the number in position $i = \frac{1}{2}(n+1)$ when the n values are sorted in order. (If i is a fraction, we average the two numbers before and after this position).

The *first quartile* (Q_1) of a set of values is the number with one-quarter of the values less than or equal to that number and three-quarters greater than or equal to the number. It is the number in position $i = \frac{1}{4}(n+1)$ when the n values are sorted in order. (If i is a fraction, we average the two numbers before and after this position).

The *third quartile* (Q_2) of a set of values is the number with one-quarter of the values less than or equal to that number and three-quarters greater than or equal to the number. It is the number in position $i = \frac{3}{4}(n+1)$

when the n values are sorted in order. (If i is a fraction, we average the two numbers before and after this position).

The five numbers min, Q_1, Median, Q_3, max are often referred to as the *five number summary* of the set of numbers.

The *interquartile range* (IQR) of a set of values is the number $Q_3 - Q_1$. It gives the range for the *middle half* of the data. This is a more resistant measure of spread than the range – but is less intuitive for most people.

A number in a set of values is an *outlier* if it is very far from (larger or smaller) the pattern of the other values. [Note: This *is* a bit fuzzy and the meaning is subject to context.] Outliers often create difficulties for statistical analysis and are often signs of trouble in data collection or interpretation. [We will worry about outliers – but in this introductory course we will not have time to discuss what to do with them – other than look for expert help].
The *lower fence* for a set of values is the number $Q_1 - 1.5 \times IQR$. The *upper fence* is the number $Q_3 + 1.5 \times IQR$. A value greater than the upper fence or less than the lower fence is potentially (but not certainly) an outlier.

A *boxplot*, based on the five-number summary and the fences, gives a quick way to represent the distribution (and is useful for quick comparisons between two sets of values for the same variable): Above a number line, there is a horizontal box from Q_1 to Q_3 with a vertical line at the median, and lines extend out from the quartiles to the largest and smallest values that are not past the fences. Values that are past the fences (potential outliers) are plotted as asterisks (*). (see pp. 167-168 in the text for details and for information on interpretation). (The TI-8x calculators will draw boxplots using the "statplot" commands; Minitab draws boxplots using the Graph>Boxplot command - but draws them vertically, as shown below.)

The *k-th percentile* (P_k) of a set of values is the number with $k\%$ of the values less than or equal to that number and $(100 - k)\%$ of the values greater than or equal to that number. It is the number in position $i = \frac{k}{100}(n + 1)$ when the n values are sorted in order. (If i is a fraction, we average the two numbers before and after this position).

The *z-score* of a value x in a set of numbers is the difference between x and the mean of the numbers, divided by the standard deviation (measured in standard deviation units). If the set of numbers is considered a population, $z = \frac{x-\mu}{\sigma}$; if the set is a sample, $z = \frac{x-\bar{x}}{s}$. The difference is in the symbols, not in the idea. We sometimes use z-scores to compare relative positions in different data sets (see text Example 1 p.1155-156). [This is another use of the idea of measuring position in "how many standard deviations from the mean" that shows up in the "empirical rule" shown in class Wednesday]

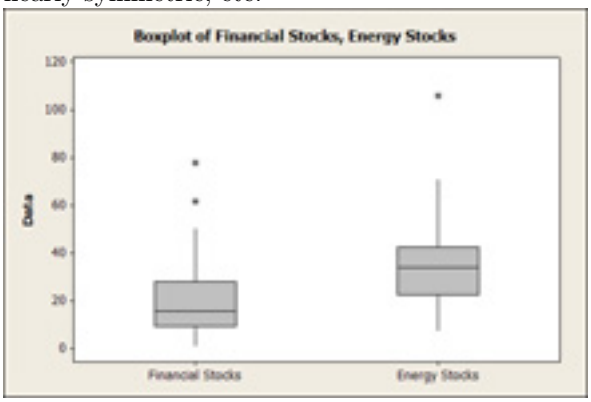
EXAMPLE

Consider the data sets in Exercise 33 on p.144: These show five-year rates of return for two groups of stocks – 32 financial stocks and 32 energy stocks - we will look at the returns on financial stocks: Sorted into increasing order, the values are
.086 1.05 7.24 7.82 8.30 8.58 10.01 12.15 13.53 14.44 15.21 15.52 16.01 16.27 16.54 18.30 125.89 16.95 126.57 29.66 31.40 47.16 50.75 61.90 77.82

The median return is in position $\frac{1}{2}(25 + 1) = 13$—so the median is 16.01. Similarly, Q_1 is in position $\frac{1}{4}(25+1) = 6.5$ so we take $\frac{8.58+10.01}{2} = 9.29$ as Q_1. Q_3 is in position $\frac{3}{4}(25+1) = 19.5$ and is $\frac{26.57+29.66}{2} = 28.12$.
The five-number summary for these values is minimum = 0,86, $Q_1 = 9.29$, Median = 16.01, $Q_3 = 28.12$, Maximum = 77.82 (This is what they mean—they are easier to find using Minitab)
We can see that this set of values is somewhat skewed right – the minimum is 15.15 below the median and Q_1 is only 6.72 below the median, but the maximum is 61.81 above the median and Q_3 is 12.11 above the median — the right (larger values) side of the distribution stretches out further.
The interquartile range (containing the middle half of the values) i s $Q_3 - Q_1 = 28.12 - 9.29 = 18.83$. The lower fence is $9.29 - 1.5 \times 18.83 = -18.28$, so there are no outliers on the low side. The upper fence is $28.12 + 1.5 \times 18.83 = 56.37$ so there are two possible outliers (61.90 and 77.82) on the high side.

The 90-th percentile (which should separate the largest 10% of values from the lower 90%) will be in position $\frac{90}{100}(25 + 1) = 23.4$ and is $\frac{50.75+61.90}{2} = 53.32$
The mean of the values is 22.36, standard deviation 18.94. The z-score of the value 16.27 is $z = \frac{16.27-22.36}{18.94} = -.32$, meaning that this number is .32 of a standard deviation below the mean. The z-score of the value 77.82 is $z = \frac{77.82-22.36}{18.94} = 2.93$, so this value (the maximum) is 2.93 standard deviations above the mean. [The sign matters and tells us "above" or "below"– we have to take the value minus the mean]
We can calculate similar values for the other set of values.(From Minitab: Mean = 36.25, StDev = 20.08, Min = 7.22, $Q_1 = 22.68$, Median = 34.18, $Q_3 = 42.67$, Max = 106.05)
A pair of side-by-side boxplots (from Minitab - which draws them vertically) shows that the median for the

enrgy stocks was above the third quartile for the financial stocks, distribution for energy stocks was more nearly symmetric, etc.



EXERCISE

1. For the data in exercise 11 on p. 170 [Note this set is available as a Minitab data set - take advantage]:
   (a) Give the five-number summary(minimum, Q_1, median, Q_3, maximum)
   (b) Draw a boxplot (copy it onto your report or print it to hand in with the report)- does the distribution appear to be symmetric? Skewed left?skewed right?
   (c) Find the upper and lower fences - are any values identified as likely outliers? (if so, which?)
   (d) Give the 80-th percentile of the weights for these 25 tablets.
   NOTE: You can use Minitab top sort the data into order using the CALC>SORT command (store the sorted data in a new column—say c5—in the same worksheet)
   (e) Give the 40-th percentile of the wights for these 25 tablets.
   (f) Give the mean and standard deviation of the values.
   (g) What proportion (fraction) of the weights are between one standard deviation below the mean and one standard deviation above the mean? (This is *not* a bell-shaped distribution)
   (h) Give the z-scores for the values .601 and .612.
2. Use side-by-side boxplots to complete exercise 16 on p. 170 (note there's a Minitab data set available)
3. Use z-scores to do exercise 9 on p.161

READING ASSIGNMENT (in preparation for next class)
In Sullivan, read section 4.1 on scatter diagrams and correlation for Monday

SKILL EXERCISES:(hand in - individually - at next class meeting Note Minitab or calculator can be used to get quartiles, median, min, max) Sullivan p.161 #10, 15, 22 [remember Minitab] p169 # 10, 15