**STAT200 – Elementary Statistics**

**Instructor: Larry Musolino**

##### Chapter 2 Summary

To Create a Frequency Distribution Table (for integer data):

1. Note: The number of classes is the same as the number of rows in the Table.
2. Calculate the class width using the formula: 
3. Note: For integer data, round Class Width up to the next higher whole number.
4. Create a table with the number of rows equal to the number of classes. Create columns for Lower Class Limit, Upper Class Limit, Frequency, Relative Frequency, etc.
5. Set Lower Class Limit for the 1st row equal to the minimum value from the dataset
6. Then calculate the Lower Class limit for the 2nd row as the lower class limit from 1st row **plus** the class width. Add the class width to calculate lower class limit for other rows.
7. The Upper Class limit for each row is one less than the lower class limit from the subsequent row. You can also add the class width to each upper class limit to determine the upper class limit for the subsequent row.
8. Frequency for each Row: count of how many datavalues fall between the lower class limit and upper class limit for that row
9. Relative Frequency for each Row: Take Frequency for that class and divide by the total number of datavalues
10. Midpoint for each Row: take lower class limit plus upper class limit, divided by 2
11. Class Boundaries for each Row: Lower Class Boundary is Lower Class Limit – 0.5
12. Class Boundaries for each Row: Upper Class Boundary is Upper Class Limit + 0.5
13. Cumulative Frequncy for each Row: Take frequency for that row, plus the frequencies for all previous rows

Important to graph/plot dataset of interest, examples of graphs include:

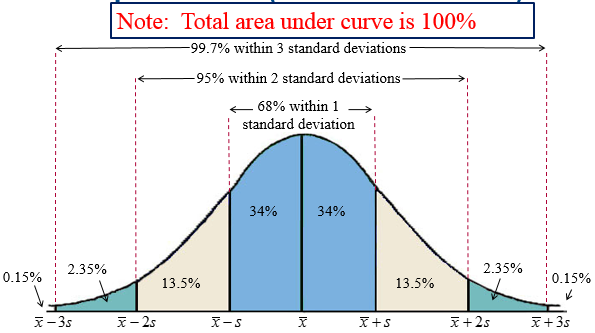
* Frequency histogram,
* Note: for frequency distribution table:
* Dot plot,
* Stem and Leaf plot,
* Boxplot
* Scatter plot for (x, y) data

Benefits of Graphing:

* Identify outliers.
* Identify shape of distribution, e.g. symmetrical (bell-shaped), right-skewed, left-skewed

|  |  |
| --- | --- |
| **Sample mean (sample average)** |  |
| **Sample median** | * Order data from smallest value to largest value, median is then the middle data entry in this ordered set. * If there are an even number of datavalues, take the two middle data values, add together and divide by 2 to determine the median. |
| **Range** | (Largest Data Value) – (Smallest Data Value) |
| **Sample Variance** |  |
| **Sample Standard Deviation (notice this is the square root of the sample variance)** |  |
| **Sample mean approximation from grouped data, i.e. frequency histogram** | where  is the midpoint of each class and  is the frequency of each class.  Note that |
| **Sample standard deviation approximation from grouped data, i.e. frequency histogram** | where  is the midpoint of each class and  is the frequency of each class. Note that |
| **Weighted mean where each datavalue has a corresponding weight** | where ***x*** are the datavalues (for example scores)  and ***w*** are the weights (in decimal format) |
| **Identifying Outliers using IQR** | To identify outliers in a dataset:   1. Find the Quartiles of the dataset: Q1, Q2, Q3 2. Find the IQR, where IQR = Q3 – Q1 3. Calculate a lower limit for outliers: Lower Limit = Q1 – 1.5\*IQR 4. Calculate a upper limit for outliers: Upper Limit = Q3 + 1.5\*IQR 5. Any datavalues below the lower limit, or above the upper limit are outliers |

**Empirical Rule - For Bell Shaped Distribution**



Empirical Rule - For Bell Shaped (Symmetric) Distribution:

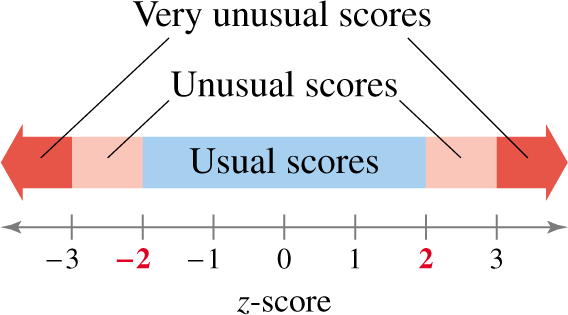
* ~68% of distribution is within one standard deviation of mean.
* ~95% of distribution is within two standard deviations of mean.
* ~99.7% of distribution is within three standard deviations of mean.

***z*** scores: measures how far a datavalue is from the mean, relative to the number of standard deviations.

For example, if a datavalue is two standard deviations above the mean then the corresponding z-score would be z = 2.0

****

|  |  |
| --- | --- |
| **Unusual data value** | Data value which is more than 2 standard deviations away from the mean. |
| **Very unusual data value** | Data value which is more than 3 standard deviations away from the mean. |



We can use z-score to determine the number of standard deviations a given data value, x, falls from the mean.

|  |  |  |
| --- | --- | --- |
| **Interpreting z-scores** | if z-score is negative: | the x-value is less than the mean |
| if z-score is zero: | the x-value is equal to the mean |
| if z-score is positive: | the x-value is more than the mean |
| if z-score is greater than 2 or less than -2 | the x-value is **unusual** since only ~5% of data values are outside of ±2 standard deviations. |
| if z-score is greater than 3 or less than -3 | the x-value **is very unusual** since only ~0.3% of data values are outside of ±3 standard deviations. |

**StatCrunch commands used in Chapter 2:**

|  |  |
| --- | --- |
| STAT 🡪 Data 🡪 Sort | To sort a column of data in ascending or descending order. |
| STAT 🡪 Summary Stats 🡪 Column | Generates summary statistics for a column of data such as mean, standard deviation, median, quartiles, etc.  Note: In StatCrunch, standard deviation refers to *sample* standard deviation, while  Unadjusted standard deviation refers to *population* standard deviation |
| STAT 🡪 Calculators 🡪 Custom | Calculates weighted mean.  Also calculates mean and standard deviation for grouped data |
| GRAPH 🡪 Histogram | Generates histogram for a column of data |
| GRAPH 🡪 Stem and Leaf | Generates stem and leaf plot for column of data |
| GRAPH 🡪 Dot Plot | Generates dot plot for column of data |
| GRAPH 🡪 Pie Chart 🡪 With Summary | Generates pie chart |
| GRAPH 🡪 Scatter Plot | Generates scatter plot for table of (x, y) data |
| GRAPH 🡪 Box Plot | Generates box plot for column of data |