# Lecture 2 –EDA Fundamentals Continued

## One Quantitative Variable

A) Histogram –

a. Example - Midterm exam scores

Notes regarding the histogram:

1) Bins can be labelled at the endpoints or the midpoints. If labelled at the midpoint, how do we find the endpoints?

2) The number of bins makes a difference!

B) Stem and Leaf Plot – best for small datasets and retains & sorts the actual data

a. Example – Actress Ages

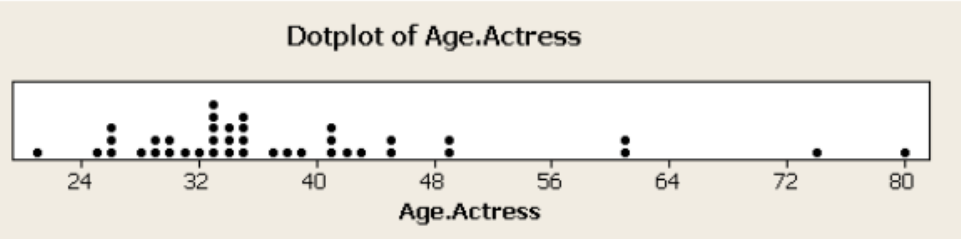
**Stem-and-Leaf Display: Age.Actress**

```
Stem-and-leaf of Age.Actress  N  = 40
Leaf Unit = 1.0

 8    2  15666899
(19)  3  0012333334445555789
 13   4  111235599
 4    5
 4    6  11
 2    7  4
 1    8  0



 1   2  1
 8   2  5666899
 20  3  001233333444
 20  3  5555789
 13  4  11123
 8   4  5599
 4   5
 4   5
 4   6  11
 2   6
 2   7  4
 1   7
 1   8  0
```

Datapoints in this data set:

C) Dotplot

a. Example – Actress Ages




Notes:

a.

b.

D) Boxplot – visually displays the 5-number summary; flipping can help display distribution

## Numerical Measure – Descriptive Statistics "Describing the Distribution"

1) Shape

    a. Symmetry: skewed right (positive skew), skewed left (negative skew), bell-shaped (symmetric curve)

    b. Peakedness – number of modes, uniform (no modes), skewed data can also be bi-modal

2) Measures of Central Tendency – what can be considered to be a "typical" value

    a. Mean –

        i. Formula

        ii. Example – Test Scores (56,72,72,95,100)

b. Median –

i. Formula

ii. Example 1 – Given a list of numbers

1. Case 1 – n is odd

2. Case 2 – n is even

iii. Example 2 – Given a stem and leaf plot

1. Case 1 – n is odd

2. Case 2 – n is even

c. Mode –

i. Case 1: 1 clear mode

ii. Case 2: bimodal

iii. Case 3: no mode

iv. Case 4: multiple numbers with frequency greater than 1

Notes on Central Tendency:

i. When to use median versus mean

ii. Example - On an easy exam, which is higher mean or median?

iii. The mean should also be used if the distribution only has few distinct values (Anything with less than 5 distinct values)

**However, the center alone does not tell you everything!!!!**

| Group 1 | Group 2 |
|---|---|
| 64, 65, 66, 67, 68 | 45, 46, 66, 86, 87 |

3) Measures of Spread/Variability (about the center)

a. Range

b. Standard deviation –

Notes on Standard Deviation:

i. Use only with the mean

ii. $s \geq 0$ always

iii. $s = 0$ means that every value is the same in the dataset

iv. Empirical Rule - for Bell-shaped Distributions ONLY

1. Example – Using the Empirical Rule

v. Chebyshev's Inequality – when you don't know about the distribution, or know it is skewed

1. Formula

2. Example - income

c. IQR –

i. Example 1 – Given a list of numbers

1. Case 1 – n is odd

2. Case 2 – n is even

ii. Example 2 – Given a stem and leaf plot
    1. Case 1 – n is odd

    2. Case 2 – n is even

***Note : Use IQR whenever using the median***

4) Outliers – observations that fall outside the overall pattern; tied into spread

    a. Can use your eye, or take a more quantitative approach

    b. Outliers using IQR – 1.5 IQR Rule of Thumb

        i. Formulas

        ii. Example

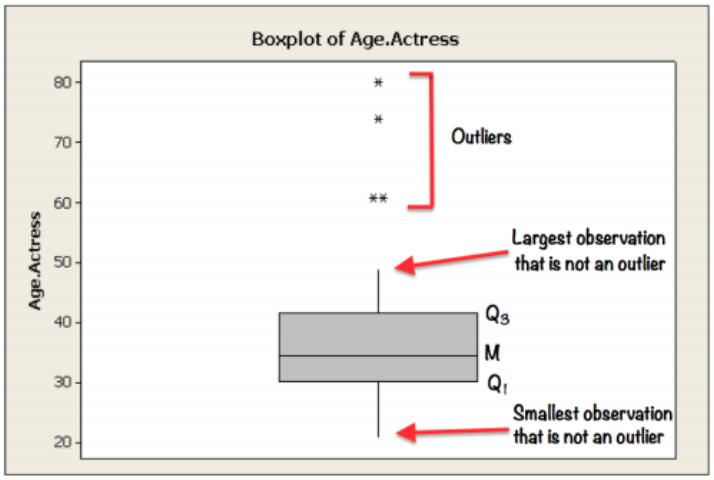
    c. Outliers generally imply skew

    d. For Bell shaped data →

Descriptive Statistics Summary

**Note**: there are different variables used for sample info vs population info

Re-examine the boxplot:




Potential Questions:

1) 50% of observations have greater than what value?
2) What percentage of the data falls below some value?
3) Within which interval would you expect to find the largest number of observations?

**To summarize:**

**1 Quantitative Variable:**

*Graphical*:
*Numerical Summary:*

## NOW WE MOVE ONTO 2 VARIABLES

We have discussed types of variables (quantitative versus categorical).

Now, we will discuss roles of variables. Specifically, now with 2 variables, we are interested in determining which is the dependent variable (Response Variable) and which is the independent variable (Explanatory Variable).

Dependent =
Independent =

Examples:

1) Does smoking cause lung cancer?

   Dependent =
   Independent =

2) Does height affect salary?

   Dependent =
   Independent =

3) Can SAT score predict freshman GPA?

   Dependent =
   Independent =

**Note → Always identify the ROLE and TYPE of each variable**

**A) C → Q**

1) Here, we are comparing the distributions of the quantitative response variable across the different categories of the explanatory variable

*Graphical*: Side-by-side boxplots
*Numerical Summary:* descriptive statistics of the response for each level of the explanatory

2) Example - does type of meat dictate calories consumed?

Dependent =
Independent =

3) Need Descriptive Statistics for Beef, then for Turkey, and then for Chicken (3 sets). Then you can create the side-by-side boxplots.

**B) C → C**

1) Explanatory variable and response variable both categorical

   *Graphical*: Contingency table & stacked bar graph
   *Numerical Summary:* conditional percentages of the response for each level of the explanatory variable separately

2) Need to fill in contingency table (aka two-way table of observed counts)

3) Then, supplement the contingency table with conditional percentage of the response for each category of the explanatory separately (condition over the **EXPLANATORY** variable) → can use stacked bar graph (doesn't matter if row or column, only matters where explanatory variable is)

4) Example - does hair color impact eye color?

   Dependent =
   Independent =