1. Overview

The usual two numbers summarizing a distribution are the "center"[the "typical" value] and the "spread" [how close or far the data are to each other, i.e. variability].

2. "Center"

A. The mean or average is denoted: $\bar{y}$ *(pronounced "y-bar") for a variable called "y"*

The mean is computed as follows: given a list of *n* numbers: $y_1, y_2, \ldots, y_n$,

Apply the formula:

$$\bar{y} = \frac{sum\_of\_all\_the\_values}{total\_number\_of\_values} = \frac{y_1 + y_2 + y_3 ... + y_n}{n} = \frac{\sum_1^n y_n}{n}$$

B. The Median is often denoted M

The median is the "middle value" of a list that has been sorted **in order** from lowest to highest. Half of the data are larger than (or equal to) the median, and half of the data are smaller than (or equal to) the median. The median may or may not be an actual value in your list.

The median is computed as follows:

(1) Given a list of *n* numbers $y_1, y_2, \ldots, y_n$,

(2) Sort all the numbers from lowest value to highest value

(3) Select the middle number from the list. This is the median

(4) If a list *n* is even-numbered, the average of the two middle numbers is the median.

C. Remarks

(1) The mean treats values as if they were little weights, the median treats values as if they all weighed the same amount.

(2) Symmetric histograms have means and medians which are equal or nearly equal, skewed histograms have means and medians which can be very different

(3) You can distort the value of the mean with a single outlier whereas the median is relatively insensitive to outliers.

(4) To calculate a mean, it is not necessary to know HOW MANY numbers are in a list, only the RELATIVE FREQUENCY of the values. Example: If we have 5 players with salaries 1,2, 3, 4, 4, the mean of the list is 2.8. If we had 10 players with salaries: "1, 1, 2, 2, 3, 3, 4, 4, 4, 4" the mean is still 2.8. As long as the values in the list maintain their relative frequencies (in this example: 20% y1's, 20% y2's, 20% y3's 40% y4's) the mean will be unchanged.

3. "Spread"

A. Minimum
B. Maximum

C. Range
D. Percentiles & Quartiles
E. IQR (inter quartile range)

A. The Sample Standard Deviation (SD)

The usual measure of spread is the STANDARD DEVIATION, written as SD or as a lowercase "s" when calculated for samples.

Formula:

The sample SD, s, is defined as follows – given a list of *n* numbers: y_1, y_2, ... , y_n,

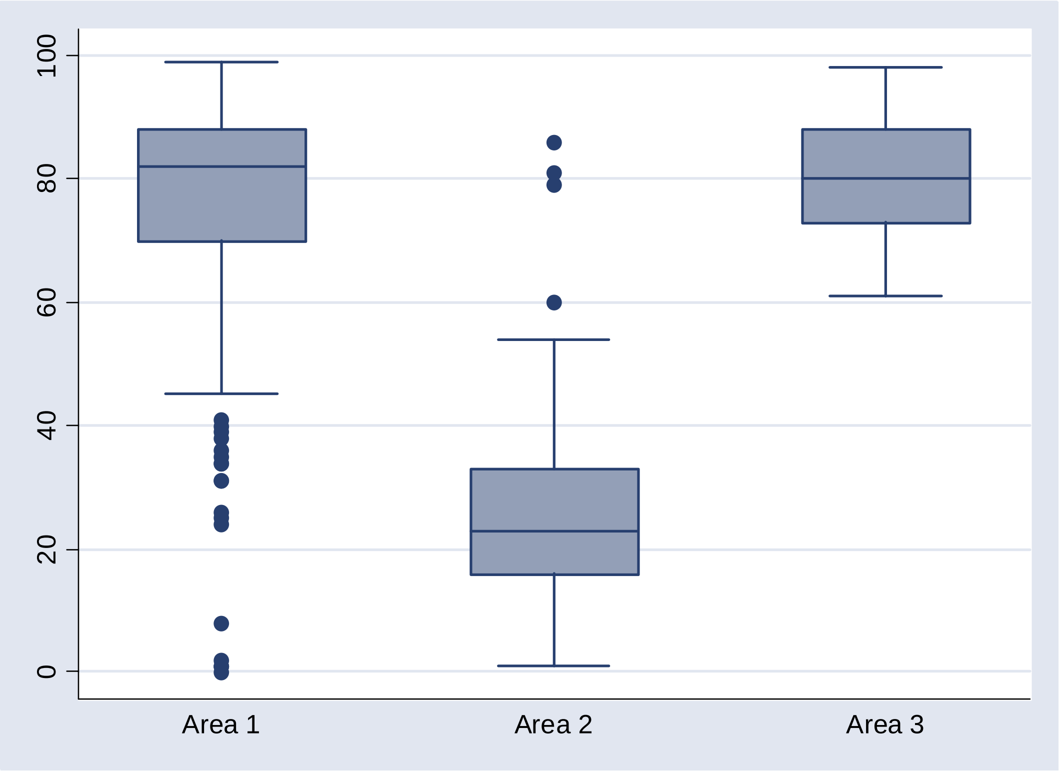
$$s = \sqrt{\frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_n - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum_1^n (y_i - \bar{y})^2}{n-1}}$$

where y-bar is the mean of the n numbers.

The square of the sample standard deviation or s^2 is called the sample variance

B. Five-Number Summary & Boxplots

A five number summary for a variable is its minimum, first quartile, median, third quartile, and maximum. A Boxplot is a graphical summary that use the five number summary to provide a lot of information in a very simple drawing.



C. Remarks on the SD

Standard deviations usually make more sense when you are comparing them for example, these are comparisons of the age of death in Los Angeles for 3 different areas:

Let’s take a numerical look at the areas to see how very different they are.

```
AREA 1
                             age
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              0
 5%           35              1
10%           51              1       Obs                 268
25%           70              1       Sum of Wgt.         268

50%           82                      Mean           75.99627
                        Largest       Std. Dev.      19.01557
75%           88             98
90%           92             98       Variance       361.5917
95%           95             99       Skewness      -1.874109
99%           98             99       Kurtosis       6.926898
```

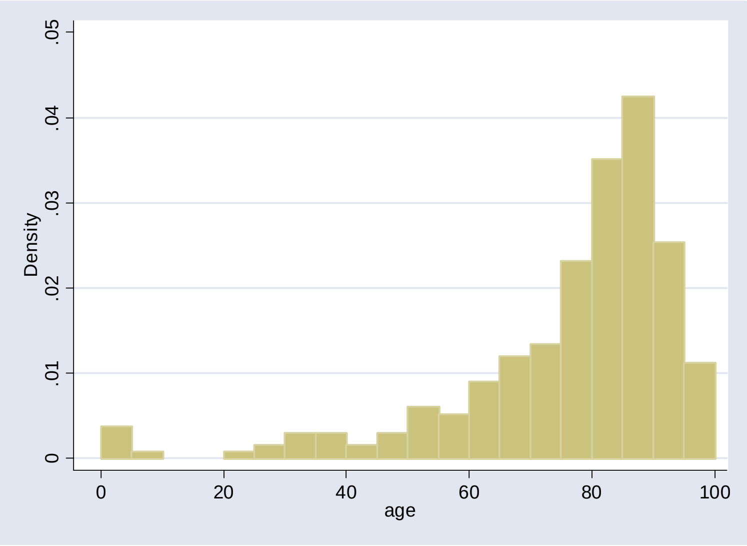
```
AREA 2
                             age
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1              1
 5%            1              1
10%            2              1       Obs                  79
25%           16              1       Sum of Wgt.          79

50%           23                      Mean           25.65823
                        Largest       Std. Dev.      18.45013
75%           33             60
90%           48             79       Variance       340.4073
95%           60             81       Skewness       1.003433
99%           86             86       Kurtosis       4.464647
```

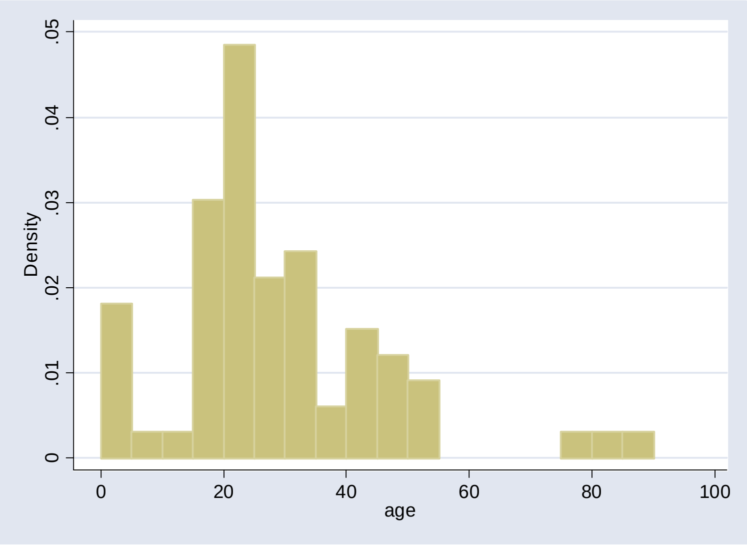
```
Area 3
                             age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           61             61
 5%           67             65
10%           67             66       Obs                  79
25%           73             67       Sum of Wgt.          79

50%           80                      Mean           79.89873
                        Largest       Std. Dev.      8.995148
75%           88             95
90%           91             96       Variance       80.91269
95%           95             97       Skewness       .0550155
99%           98             98       Kurtosis       2.007271
```

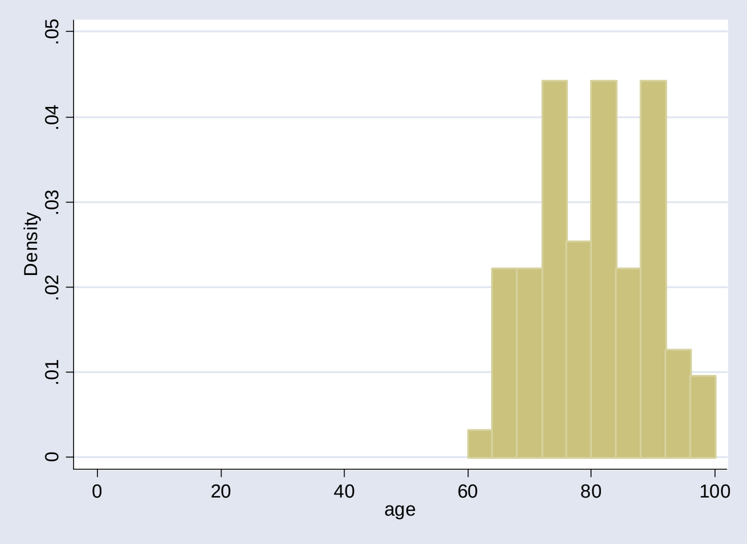
Area 1 again



Area 2 again



And Area 3 again


Density
age