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. "Spread"

A. Minimum
B. Maximum
C. Range
D. Percentiles & Quartiles
E. IQR (inter quartile range)

THOSE ARE NICE ROBUST MEASURES (e.g. relatively resistant to extreme observations) GOOD FOR GETTING AN IDEA OF WHAT THE DISTRIBUTION LOOKS LIKE (e.g. patterns) but not as commonly used as…

3. 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.

A. Formulas

The sample SD is defined as follows: given a list of numbers x1, x2, ... , xn,

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n\text{-}1}}$$

often rewritten as

$$s = \sqrt{\frac{\sum_{1}^{n} (x_i - \bar{x})^2}{n-1}}$$

where $\bar{x}$ is the mean of the n numbers.

(we use n-1 here because we'd really like the sample standard deviation to be a good estimate of the population standard deviation. For reasons beyond the scope of this course, it turns out that the divisor n-1 gives a better estimate of the population standard deviation and population variance than the divisor n does)

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

The population has a standard deviation with notation σ and it has a variance σ^2

Example:

An interview with 5 UCLA students reveals the time in hours spent surfing the web in a given week: 7, 17, 3, 14, 7

C. 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.

D. Drawing a boxplot

- Construct either a vertical or horizontal axis
- Construct a Box whose lower edge (if scaled vertically) or left edge (if scaled horizontally) represents the value of the 25th percentile
- Construct a Box whose upper edge (if scaled vertically) or right edge (if scaled horizontally) represents the value of the 75th percentile
- Draw a line segment (horizontal if scaled vertically, vertical if scaled horizontally) at the value of the median.
- Extend vertical (if scaled vertically) or horizontal (if scaled horizontally) line segments to the minimum if the minimum does not exceed the value of (the 25th percentile – (1.5*IQR)). Your book calls this a "whisker". If the minimum does exceed that value, then draw a perpendicular line to mark that value. Any values lower than the value of (the 25th percentile – (1.5*IQR)) are represented by open circles or asterisks.
- Extend vertical (if scaled vertically) or horizontal (if scaled horizontally) line segments to the maximum if the maximum does not exceed the value of (the 75th percentile + (1.5*IQR)). Your book calls this a "whisker". If the maximum does exceed that value, then draw a perpendicular line to mark that value. Any values higher than the value of (the 75th percentile + (1.5*IQR) are represented by open circles or asterisks .

Your textbook identifies values as outliers if they are more that 1.5*IQR away from the nearest edge of the "box" and extreme outliers are values that are more than 3.0*IQR away from the nearest edge of the "box". These values are represented by circles and asterisks. Stata just uses circles and doesn't make the distinction.

Values exceeding 1.5*IQR is a common criterion for outliers.

D. Remarks

Note that the standard deviation is in the same units as the data. In our case, it's hours. This is why the standard deviation involves the square root, it allows for easier interpretation than hours-squared for example. The SD measures how close the numbers in the list are to the average; i.e., not all numbers are equal to the mean; the SD is a measure of the "average" distance between each point and the average. The SD is tied to the mean, usually people talk about them together. It also has some of the same problems as the mean, that is, it is very sensitive to outliers.

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 4 different zip codes:

```
-> zipcode = 90011
    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+-----------------------------------------------------
         age |     410    61.71821    24.74859          0   99.08282

-> zipcode = 90024
```

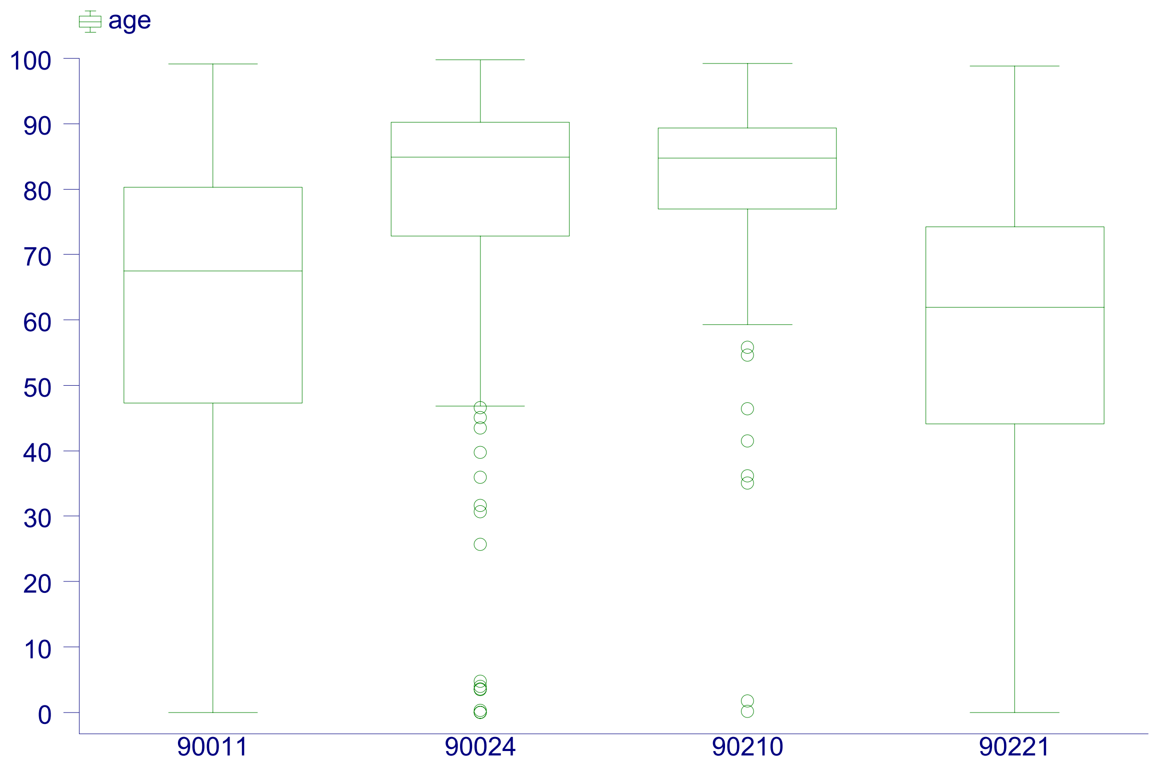
```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       238    78.38765    19.28202          0   99.73169

-> zipcode = 90210

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       200    81.64175    13.44267   .1259411   99.20876

-> zipcode = 90221

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       245    57.18768    23.36865          0   98.80356
```



Let’s take a closer look at two of the zip codes to see how very different they are.

```
. summarize age if zip==90210, detail

                             age
-------------------------------------------------------------
      Percentiles      Smallest
 1%     18.36413       .1259411
 5%     61.57563       1.724846
10%     68.55031       35.00342       Obs                 200
25%     76.90349       36.18617       Sum of Wgt.         200

50%     84.70774                      Mean           81.64175
                        Largest       Std. Dev.      13.44267
75%     89.34702        97.8371
```

```
90%      93.44969       98.08624       Variance       180.7055
95%      96.10678       98.23135       Skewness      -2.848412
99%      98.15879       99.20876       Kurtosis       15.72241

. summarize age if zip==90221, detail

                              age
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%     4.788501              0
10%     22.17112              0       Obs                 245
25%     44.06297              0       Sum of Wgt.         245

50%     61.90554                      Mean           57.18768
                        Largest       Std. Dev.      23.36865
75%     74.23956       94.64476
90%     81.11156       94.84463       Variance       546.0939
95%     89.17728       94.98152       Skewness      -.7722799
99%     94.84463       98.80356       Kurtosis       2.962964
```