

Definition. A population is a collection (or an aggregate) of objects or elements that, generally, have at least one characteristic in common. If all the elements can be well defined and placed (or listed) onto a frame from which the sample can be drawn, then the population is said to be concrete and existing; otherwise, it is a hypothetical, conceptual, or a virtual population.

Example 1. (a) All Auburn University students ($N \cong 28000$ members on 2 campuses). Here the frame may be AU Telephone Directories. (b) All households in the city of Auburn. Again the frame can be the Auburn-Opelika Tel. Directory. (c) All AU COE students, where the frame can be found on the web at <http://www.eng.auburn.edu/info/listing-all.html>

Examples 1.1 and 1.2 on pp. 4-6, 1.4 on p. 8, 1.8 on p. 14, 1.10 on pp. 17-18, 1.12 on p. 25, 1.15 on p. 33, 1.18 on p. 37 of Devore's 7th edition provide sampling from conceptual (or virtual) populations.

A variable, X , is any characteristic whose value changes from one element of a population to the next and can be categorical, or quantitative.

Example 2. (a) Categorical or Qualitative variable X : Examples are Grade performance in a college course; Success/Failure; Freshman, Sophomore, Junior, and Senior on a campus; Pass/Fail, Defective/ Conforming, Male/Female, etc.

(b) Quantitative Variable X : O-Ring Temperature (Example 1.1 on page 4 of Devore); Flexural Strength in MPa (Example 1.2 of Devore, p. 5), Diameter of a Cylindrical Rod, Length of steel pipes, Bond Strength of Concrete (Example 1.10 on pp. 17-18, sample size $n = 48$), Shear Strength (lb) of Exercise 24, etc.

Note that W. Edwards Deming (perhaps the most prominent of Quality gurus in the 20th century) generally refers to studies made on concrete populations as enumerative and those made on conceptual populations as analytic.

Branches of Statistics

(1) Descriptive, (2) Inductive or Inferential

Descriptive Statistics comprises of all methods that summarize collected data and is subdivided into 2 categories: (i) Pictorial and Tabular : Stem-and-leaf plot, Histogram, and Boxplots. (ii) Numerical (or quantitative) Measures: of Location (i.e., the mean, the median), of Variability, of Skewness, and of Kurtosis.

(i) Stem-and-leaf Plot for the Example 1.1 on page 4 of Devore's (7e)

Order-statistics $x_{(i)}$: 31°F, 40, 45, 49, 52, 53, 57, 58, 58, 60, 61, 61, 63, 66, 67, 67, 67, 68, 69, 70, 70, 70, 70, 72, 73, 75, 75, 76, 76, 78, 79, 80, 81, 83, 84

The sample size, universally denoted by n , is equal to 36.

$x_1 = 84$, $x_{(1)} = 31$, $x_2 = 49$, $x_{(2)} = 40$, $x_3 = 61$, $x_{(3)} = 45$, ..., $x_{36} = 31$, $x_{(n)} = x_{(36)} = 84$.

Stem = 10 (Minitab increment =10) , Leaf = 1 °F.

(Cumf _i)	Stem	Leaf	(n/2 = 18)
1	3	1	
4	4	0 5 9	
9	5	2 3 7 8 8	
(11)	6	0 1 1 3 6 7 7 7 8 9	
16	7	0 0 0 0 2 3 5 5 6 6 8 9	
4	8	0 1 3 4	

I will name the stem "6" as the median stem for the above example because its sample median lies in the interval $60 \leq \tilde{x} = \hat{x}_{0.50} = \text{Sample Median} = 67.50 \leq 69$ °F.

Histograms. (See the Example 1.9 on p.16 of Devore's 7th edition)

The 1st-order statistic is $x_{(1)} = 2.97$, the n th-order statistic is $x_{(n)} = 18.26$, and the sample size $n = 90$. $R = x_{(90)} - x_{(1)} = 15.29$, $C = \text{No. of subgroups (or classes, or bins)}$, $C_1 = 1 + 3.3 \times \log_{10}(n)$ [Sturges' practical guideline], $C_1 = 1 + 3.3 \log_{10}(90) = 7.45$. Or $C_2 \cong \sqrt{n} \rightarrow C_2 = 9.49$, or Shapiro's recommendation $C_3 = 4[0.75(n - 1)^2]^{0.20} = 4[0.75(89)^2]^{0.20} = 22.7420$; this last guideline is generally too large and should be

used only when $n > 500$. Thus, it is best to select between 7 to 10 subgroups. So, we choose $C = 9$ subgroups. As a result, $\Delta_j = j^{\text{th}}$ subgroup width = $R/C = 15.29/9 = 1.6988 \uparrow 1.70 \rightarrow \Delta = 1.70$. (Always round up to obtain Δ to the same number of decimals as the original data.) Note that class limits must have the same number of decimals as the original data, but boundaries must carry one more decimal. In Table 1, the upper class limit of the 1st subgroup is 4.66 while the upper boundary of the 1st class is $Ub_1 = 4.665$. The lower class limit of the 4th subgroup is 8.07 while the lower boundary of the 4th subgroup is $Lb_4 = 8.065$, etc. Further, $\Delta_j = Ub_j - Lb_j$ for all j . The frequency distribution for the Example 1.9 of Devore is given Table 1

below; the $\sum_{j=1}^C f_j$ must always add to n ($= 90$ in this case). This is why the subgroup intervals must be non-overlapping.

Table 1 The Frequency distribution of Example 1.9, p. 16

Subgroups	2.97 – 4.66	4.67 – 6.36	6.37 – 8.06
f_j	2	5	17
Classes	8.07 – 9.76	9.77 – 11.46	11.47 – 13.16
f_j	18	22	13
Subgroups	13.17 – 14.86	14.87 – 16.56	16.57 – 18.26
f_j	8	3	2

The histogram from Minitab is provided in Figure 1. In Figure 1, the area in each rectangle (or bar) represents Relative Frequency (f_j/n), and the ordinate represents the height or density $h_j = d_j$ of each rectangle. Because every histogram in the

universe must have the “Total Area Under the Histogram” = $\sum_{j=1}^C Relf_j = 1.0000 =$

$\sum_{j=1}^C h_j \Delta_j = \sum_{j=1}^C d_j \Delta_j$, and because both $(Relf_j, d_j \Delta_j)$ represent the same j th

rectangular area of the histogram, then it follows that $Relf_j = d_j \Delta_j$, and hence $d_j = Relf_j/\Delta_j$ for all $j = 1, 2, 3, \dots, C$. For the histogram of Figure 1, $h_1 = (2/90)/1.70 =$

$0.02222/1.70 = 0.013072 = d_1$, $d_2 = h_2 = 0.05555/1.70 = 0.032676/BTU$, etc. Note that Δ must have the same number of decimals as the original data.

It is extremely paramount to understand that the densities d_j have very little (if any) statistical or geometrical meaning but it is their product with the corresponding class-width, Δ_j , that gives the corresponding j^{th} rectangular area $a_j = \text{Rel}f_j = d_j \times \Delta_j$.

Minitab Project Report

The Histogram for the Exp1.9 on page 16 of Devore's 7th edition based on midpoints (m_j) and 9 subgroups

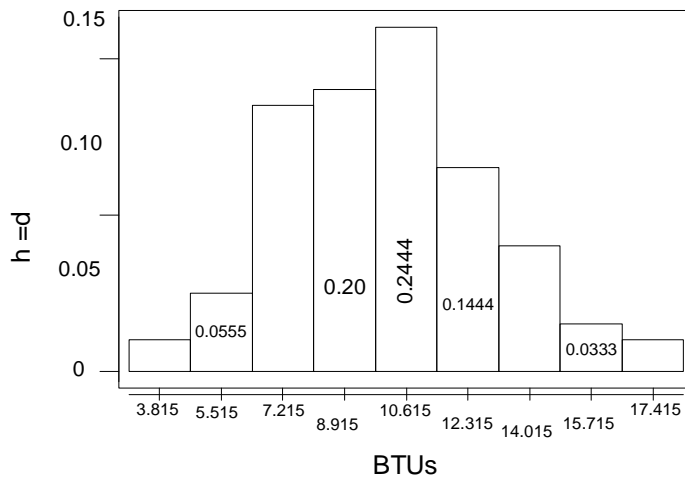


Figure 1

Further, the midpoint of each subgroup (or bin) is simply $m_j = (Ub_j + Lb_j)/2 = (UclassL_j + LclassL_j)/2$, where Ub_j is the upper boundary and $LclassL_j$ is the lower class limit of the j^{th} subgroup.

The 3rd pictorial summary, the Boxplot, will be discussed on pp. 11-12 of these notes.

1(ii) (Quantitative) Measures in Descriptive Statistics.

(a) Measures of Location : Mean (or arithmetic average), median, geometric

mean, harmonic mean, trimmed mean, mode, and percentiles.

A bar is universally used to denote averages such as the arithmetic mean \bar{x} (or \bar{y}).

The arithmetic mean is defined as $\bar{x} = \sum_{i=1}^n x_i / n$. For the example 1.1 of Devore's 7th

edition, p. 4, $n = 36$, $\sum_{i=1}^n x_i = 2371 \rightarrow \bar{x} = 65.86111$ °F. Note that the sample mean

(\bar{x} , \bar{y} , etc.) represents the center of gravity of a data set; see Figure 1.14 on p. 26 of Devore's 7th edition.

The median, $\tilde{x} = \hat{x}_{0.50}$, is another measure of central location of data such that exactly (or at most) half of the data are below $\hat{x}_{0.50}$ and at most half of the data exceed $\hat{x}_{0.50}$. To obtain $\hat{x}_{0.50}$ for any data (whether n is odd or even), 1st multiply n by 0.50. If this result is an exact integer, say r , then $\hat{x}_{0.50} = [X_{(r)} + x_{(r+1)}]/2$. Only in this case exactly half the data will lie below and the other half above the sample median \tilde{x} . If $0.50 \times n$ is not integer, then always round it up to the next higher integer, say m . Then, for this last case $\hat{x}_{0.50} = x_{(m)}$ = the m^{th} order-statistic.

For the Example 1.1 on p. 4 of Devore, $0.50 \times n = 18$, which is an exact integer. Thus, $\tilde{x} = \hat{x}_{0.50} = [x_{(18)} + x_{(19)}]/2 = (67 + 78)/2 = 67.50$. Note that exactly 18 data points lie below 67.50, and 18 points lie above 67.50. For the data of Example 1.12 on p. 25 of Devore's 7th edition, the sample size $n = 21$ gives $0.50 \times n = 10.5$, which is not an integer. Thus, round 10.5 up to the next higher order statistic $m = 11$, and as a result $\hat{x}_{0.50} = \tilde{x} = x_{(11)} = 21.20$, while $\bar{x} = 21.181$. Note that only 47.61905% of the data are below 21.20, and 47.61905% of the data are above $\hat{x}_{0.50} = \tilde{x} = 21.20$.

The geometric mean is defined as $\bar{x}_g = (x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n}$, i.e., \bar{x}_g is the n^{th} root of $\prod_{i=1}^n x_i$ only if all x_i 's > 0 for all i , and in general $\bar{x}_g \leq \bar{x}$. For the data of example 1.12 of Devore, $\bar{x}_g = 19.379764 < 21.181 = \bar{x}$. Geometric mean has applications in DOE (Design of Experiments) where at least 2 responses from each experimental unit is observed.

The harmonic mean is defined as $\bar{x}_h = \left[\sum_{i=1}^n (1/x_i) / n \right]^{-1} = \frac{n}{\sum (1/x_i)}$,

i.e., \bar{x}_h is the inverse of the average reciprocals of x_i 's. For the data of Example 1.12

on page 25 of Devore, $n = 21$, and $\frac{1}{21} \sum_{i=1}^{21} (1/x_i) = [(1/16.1) + (1/9.6) + \dots + (1/28.5)]/21$
 $= 1.1902504/21 = 0.0566786$, which yields

$$\bar{x}_h = \frac{1}{0.0566786} = 17.64335 < \bar{x}_g < \bar{x}$$

The harmonic mean has applications in ANOVA (Analysis of Variance) when the design is unbalanced. It gives the average sample size over all levels of a factor, and in general $\bar{x}_h \leq \bar{x}_g \leq \bar{x}$. For the time being, the geometric and harmonic means are not as important measures of central tendency as \bar{x} and \tilde{x} .

TRIMMED MEANS

A 10% trimmed mean, $\bar{x}_{tr(10)}$, is computed by deleting the smallest and largest $0.10 \times n$ of the order-statistics from the tails of data and computing the arithmetic average of the remaining 80% of the data. It seems that such a mean should be called the 20% trimmed mean because 20% of the data is actually removed from the original n observations x_1, x_2, \dots, x_n . However, I am certain that our author, Devore, is notationally consistent with other statistical literature, and therefore, we will use Devore's notation of $\bar{x}_{tr(10)}$. To illustrate, consider the O-ring temperature data of Example 1.1 on page 4 of Devore, for which $0.10 \times n = 3.6$.

Step 1. Trim or remove the order-statistics $x_{(1)}, x_{(2)}, x_{(3)}, x_{(36)}, x_{(35)}$, and $x_{(34)}$.

$$\text{Next, compute } \sum_{i=4}^{33} x_{(i)} / 30 = 66.90 \text{ } ^\circ\text{F} = \bar{x}_{tr3}.$$

Step 2. Trim \bar{x}_{tr3} further by removing $x_{(4)}$ and $x_{(33)}$.

$$\bar{x}_{tr4} = \sum_{i=5}^{32} x_{(i)} / 28 = 67.07143.$$

Step 3. Interpolate between \bar{x}_{tr3} and \bar{x}_{tr4} to obtain $\bar{x}_{tr(10)}$. Note that most statistical packages, such as Minitab, only give the $\bar{x}_{tr(5)}$ and round the $0.05 \times n$ to the nearest integer in order to obtain the 5% trimmed mean.

$$\bar{x}_{tr(10)} = 66.90 + 0.60 (67.07143 - 66.90) = 67.00286 \cong 67.003.$$

$$\begin{aligned} \bar{x}_{tr(10)} &= 67.07143 - 0.40(0.171430) = 67.0029 \\ &= 0.4 \times \bar{x}_{tr3} + 0.6 \times \bar{x}_{tr4} \end{aligned}$$

Had n been equal to 34, then the above formula would change to $\bar{x}_{tr(10)} = 0.6 \times \bar{x}_{tr3} + 0.4 \times \bar{x}_{tr4}$.

The trimmed mean, \bar{x}_{tr} , has applications when data contain outliers (or when the data originate from an underlying distribution with heavy tail probabilities), and \bar{x}_{tr} is always as close or closer to $\hat{x}_{0.50}$ (= 67.50 for the Example 1.1 than is \bar{x}).

The MODE

The mode is the observation with the highest frequency. For the data of Example 1.1 on p. 4 of Devore, $MO1 = 67$ and $MO2 = 70$ because both observations 67 and 70 have a frequency $f = 4$ (i.e., the data is bimodal). Most populations have a single mode; however, if a population has two or more modes, then it should be stratified for the purpose of sampling. In calculus, Mode is referred to as the point on the abscissa at which the maximum of the ordinate occurs.

Computing Sample Percentiles (or Quantiles)

The $100 \times p^{\text{th}}$ sample percentile, \hat{x}_p , is obtained using the following steps.

(1) First rearrange the data in ascending order of $x_{(1)}, x_{(2)}, \dots, x_{(n)}$.

(2) Multiply p by n : if $n \times p$ is an exact integer, say i , then

$$\hat{x}_p = [x_{(i)} + x_{(i+1)}] / 2.$$

(3) If $n \times p$ is not an exact integer, then round $n \times p$ always up to the next higher integer, i.e., $n \times p \uparrow i$. Then $\hat{x}_p = x_{(i)}$. Note that this procedure is

consistent with SAS (Statistical Analysis System), which is the oldest prominent Statistical Package but not so with other statistical packages.

For the data of Example 1.18 on p. 37 of Devore, where X represents pulse width, the 10th, 25th, 50th, 75th, 80th, and 90th sample percentiles are computed below: (Note that only for convenience hats have been removed from sample percentiles, and the sample size n = 25.)

$$x_{0.10} : 0.10 \times 25 = 2.5 \uparrow 3 \quad \rightarrow \quad x_{0.10} = x_{(3)} = 13.8$$

$$x_{0.25} : 0.25 \times 25 = 6.25 \uparrow 7 \quad \rightarrow \quad x_{0.25} = x_{(7)} = 90.2$$

$$x_{0.50} : 0.50 \times 25 = 12.5 \uparrow 13 \quad x_{0.50} = x_{(13)} = 94.8.$$

$$x_{0.75} : 0.75 \times n = 18.75 \quad \rightarrow \quad x_{0.75} = x_{(19)} = 96.7$$

$$x_{0.80} : 0.80 \times n = 20 \quad \rightarrow \quad x_{0.80} = [x_{(20)} + x_{(21)}]/2 = (98.1 + 99.0)/2 = 98.55$$

$$x_{0.90} : 0.90 \times n = 22.5 \quad \rightarrow \quad x_{0.90} = x_{(23)} = 103.7$$

The above sample percentiles are also called the 0.10, 0.25, 0.50, 0.75, 0.80, and 0.90 sample quantiles, respectively. The 0.10 quantile is also called the 1st decile, and the 0.90 quantile is called the 9th decile. Note that Minitab's quantile estimates do not always match those of SAS, both of which are given below.

Variable	N	Mean	SE Mean	TrMean	StDev	Variance	CoefVar
PW	25	84.86	5.89	87.07	29.47	868.53	34.73%
Minimum		5.30					
Variable	Q1	Median	Q3	Maximum	IQR	Skewness	Kurtosis
PW	89.10	94.80	97.40	113.50	8.30	-2.22	3.78

Minitab, and many other sources, use the procedure of multiplying p by (n+1), instead of n; if this product is an exact integer i, then they set the pth quantile equal to $x_{(i)}$. Otherwise, the pth quantile is the convex combination of the ensuing 2 order-statistics. This method is used in at least 50% of statistical literature. For the Example 1.18, Minitab's Q1 is computed by first multiplying 0.25 by n+1 = 26, which results in 6.5, so that $Q1 = 0.50 \times x_{(6)} + 0.50 \times x_{(7)} = 0.50 \times (88.0) + 0.50 \times (90.2) = 89.10$, as listed above. Note that our text is not consistent in computing sample quantiles (see the error in Example 1.17, p. 36 of Devore, where Q1 must equal 70, not 72.5, using either methods.)

SAS Output

11:32 Wednesday, August 12, 2009

Example 1.18 on page 37 of Devore's 7th Edition

The UNIVARIATE Procedure

Variable: Pulse Width

Quantiles (definition 5)

Quantile	Estimate
100% Max	113.5
99%	113.5
95%	106.0
90%	103.7
75% Q3	96.7
50% Median	94.8
25% Q1	90.2
10%	13.8
5%	8.2
1%	5.3
0% Min	5.3

In my section of STAT 3600, you may use either of the two above methods to estimate the p^{th} population quantile Q_p .

The IQR (interquartile range) is defined as $IQR = \hat{x}_{0.75} - \hat{x}_{0.25} = Q3 - Q1 =$ the 4th spread $= f_s$. The 4th spread, f_s , is Devore's uncommon notation, which is explained near the bottom of p. 35. For the Example 1.18 of Devore, its value is equal to $96.7 - 90.20 = 6.50$.

If $Q1 - 3 \times IQR < x_{(i)} < Q1 - 1.5 \times IQR$, or $Q3 + 1.5 \times IQR < x_{(i)} < Q3 + 3 \times IQR$, then the i^{th} order-statistic, $x_{(i)}$, is a mild outlier. If the value of $x_{(i)} < Q1 - 3 \times IQR$, or $x_{(i)} > Q3 + 3 \times IQR$, then $x_{(i)}$ is an extreme outlier. For the Example 1.1 of Devore, since $Q1 - 1.5 \times 16 = 59 - 24 = 35$ and $Q1 - 3 \times IQR = 59 - 3 \times 16 = 11$, only $11 < x_{(1)} = 31 < 35$, then the data contain a single mild outlier on the LHS (or the lower tail).

(b) Measures of Variability (Three Quantitative Measures)

(1) Standard deviation (Stdev) = S , (2) Range/ $d_2 = R/d_2$, and (3) the IQR = $x_{0.75} - x_{0.25}$, where the range $R = x_{(n)} - x_{(1)}$ and the IQR (or f_s) have already been defined. The parameter d_2 is a Quality Control constant that will be defined in INSY 4330. The most common measure of variability is the standard deviation followed

by R/d_2 . In order to compute S , we must always compute the variance 1st.

Definition. The sample variance, v , is the average of deviations of n observations from their-own-mean squared. (USS = Uncorrected Sum of Squares)

Data Set 1: 2.7, 3.5, 3.8, 4.6, 5.4. $n = 5 \rightarrow \bar{x} = 4.0, R = 2.7, USS = 84.30, CF = 80 \rightarrow$

$x_i - \bar{x} = -1.3, -0.50, -0.20, 0.60, 1.4 \rightarrow$

$$(x_i - \bar{x})^2: 1.69, 0.25, 0.04, 0.36, 1.96, \rightarrow \sum_{i=1}^5 (x_i - \bar{x})^2 = 4.30$$

$v_1 = 4.3/5 = 0.86$. Note that $\sum_{i=1}^n (x_i - \bar{x}) \equiv 0$ for all data sets in the universe.

Data Set 2: 2.1, 3.2, 3.6, 4.5, 6.6, ($\bar{x} = 4.0, R = 4.5, USS = 91.42, CF = 80$),

$x_i - \bar{x}: -1.9, -0.8, -0.40, 0.50, 2.6$

$\rightarrow (x_i - \bar{x})^2: 3.61, 0.64, 0.16, 0.25, 6.76 \rightarrow$

$$CSS = \text{Corrected Sum of Squares} = S_{xx} = \sum_{i=1}^5 (x_i - \bar{x})^2 = 11.42$$

$\rightarrow v_2 = 11.42/5 = 2.284$

Data sets 3: 1.9, 2.9, 4.0, 4.5, 6.7, ($\bar{x} = 4.0, R = 4.8, USS = 93.16, CF = 80$).

$(x_i - \bar{x}): -2.1, -1.1, 0, 0.50, 2.7 \rightarrow CSS = S_{xx} = 13.16 \rightarrow v_3 = 13.16/5 = 2.632$.

Note that as the overall spread of the data increases, so does the variance, i.e.,

variance is a measure of variability. Further, the divisor of v is n , i.e., $v = (1/n) \times \sum (x_i - \bar{x})^2$. Note that the n deviations from the mean $(x_1 - \bar{x}), (x_2 - \bar{x}), (x_3 - \bar{x}), \dots, (x_n - \bar{x})$

are not independent because of the constraint $\sum_{i=1}^n (x_i - \bar{x}) \equiv 0$ for all data sets in the

universe. For the data set no. 3 above, if we are given $x_1 - \bar{x} = -2.1, x_2 - \bar{x} = -1.1, x_3$

$- \bar{x} = 0$, and $x_5 - \bar{x} = 2.7$, then the value of $x_4 - \bar{x}$ is automatically constrained to $x_4 -$

$\bar{x} = -(2.1) - (-1.1) - (0) - 2.7 = 3.2 - 2.7 = 0.50$, i.e., the variables $x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n$

$- \bar{x}$ have $(n - 1)$ degrees of freedom (df) not n . Therefore, we define most common

measure of variability with the divisor of $(n - 1)$ for $S_{xx} = \sum (x_i - \bar{x})^2$, given by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = n \times v / (n-1) = S_{xx} / (n-1) = CSS / (n-1) \quad (1)$$

For data sets 1, 2 and 3 above the values of $S_1^2 = 4.30/4 = 1.075$, $S_2^2 = 2.855$, and $S_3^2 = 13.16/4 = 3.29$ because $df = 4$ (not 5). Further, as stated in equation (1), $S_1^2 = 5 \times v_1/4 = 5 \times 0.86/4 = 1.075$, and so forth. The reader should follow from above examples that the USS plays a more important role in determining the value of S^2 than the CF.

The exact name for S^2 is not the sample variance as defined by Devore on his p. 32. In actuality the sample variance is $v = \sum (x_i - \bar{x})^2 / n$ as defined herein, but v generally underestimates the population variance σ^2 because $\sum_{i=1}^n (x_i - c)^2$, where c is any real constant, attains its minimum value iff $c = \bar{x} = \sum_{i=1}^n x_i / n$. To compensate for this underestimation, we divide the CSS = $S_{xx} = \sum (x_i - \bar{x})^2$ by a smaller number than n , namely its $df = n - 1$, in order to obtain an “unbiased estimate” of σ^2 . The positive square root of S^2 provides the standard deviation, S , and dividing S by \sqrt{n} gives the standard error of the mean, i.e., $se(\bar{x}) = S / \sqrt{n}$. Further, the ratio S/\bar{x} is called the coefficient of variation (or variation coefficient), and generally the sample $cv = S/\bar{x}$ is expressed in % with at least 2 decimals.

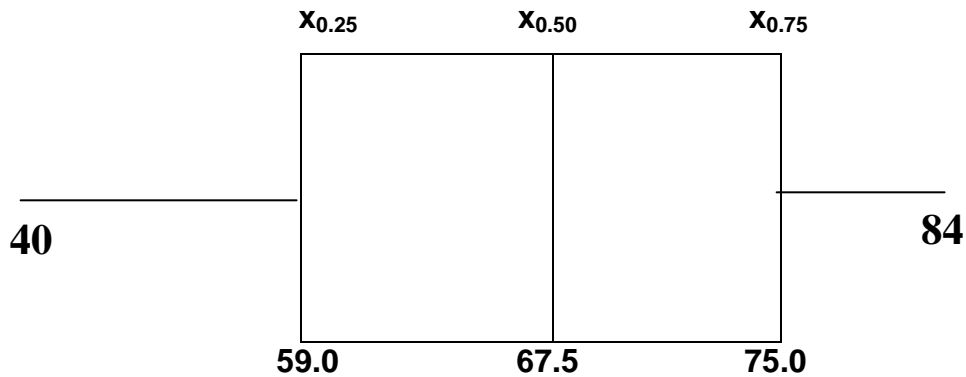
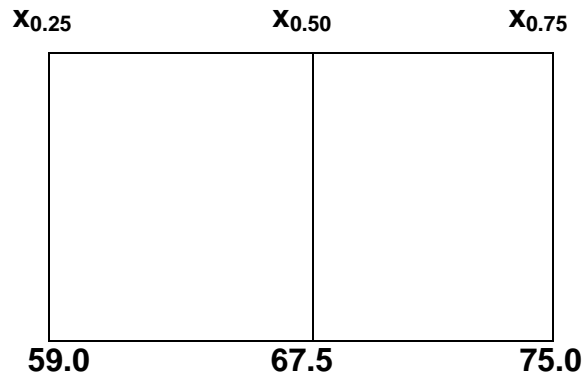
Graphical Measure of Variability (The Boxplot)

Step 1. Draw a vertical line thru the median $\tilde{x} = \hat{x}_{0.50}$.

Step 2. Draw vertical lines thru $Q1 = \hat{x}_{0.25}$ and $Q3 = \hat{x}_{0.75}$ and connect at the bottom and the top to make a rectangular box. For the data of Example 1.1, the box is shown atop the next page, where hats are removed from sample percentiles only for convenience.

Step 3. Compute both $1.5 \times IQR$ and $3 \times IQR$. For the Example 1.1, $1.5 \times IQR = 24$ and $3 \times IQR = 48$. Then all data points less than $Q1 - 1.5 \times IQR = 59 - 24 = 35$ but larger than $Q1 - 3 \times IQR = 11$, or larger than $75 + 24 = 99$ but less than 123 are mild outliers. Example 1.1 on page 4 of Devore has no extreme outliers because all data points lie within the interval $(Q1 - 3 \times IQR, Q3 + 3 \times IQR)$.

Step 4. Draw whiskers from Q1 and Q3 to the smallest and largest order statistics that are not outliers.



Note that the dark dot on the LHS of the above Boxplot represents the mild outlier $x_{(1)} = 31$. Extreme outliers are represented by clear dots.

Bonus Homework 1. (Worth 5 Points, either none or all, i.e., no partial bonus points). Please note that the solution to all Bonus problems must be only yours and no one else's. It will be considered cheating to discuss any aspects of bonus problem solutions with anyone else.

(a) Prove that $\sum_{i=1}^n (x_i - c) \equiv 0$ iff the real constant $c = \bar{x}$.

(b) Prove that the $SS = \sum_{i=1}^n (x_i - c)^2$ attains its minimum

value only if real the constant $c = \bar{x}$.

(c) Prove that for any data set of size n the Corrected Sum of Squares =

$$\text{CSS} = S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \text{USS} - \text{CF}, \text{ where the Uncorrected Sum of}$$

$$\text{Squares USS} = \sum_{i=1}^n x_i^2, \text{ and the correction factor } \text{CF} = \left(\sum_{i=1}^n x_i \right)^2 / n.$$

(d) By definitions the mean and variance of a grouped (gr) data (or a frequency distribution) are given by

$$\bar{gr} = \frac{1}{n} \sum_{j=1}^C m_j \times f_j = \bar{x}_{gr}, \text{ and } S_{gr}^2 = \frac{1}{n-1} \sum_{j=1}^C (m_j - \bar{g})^2 \times f_j.$$

Prove that for a histogram (or an empirical distribution) $\sum_{j=1}^C (m_j - \bar{g}) \times f_j = 0$,

and that the computing formula for S_{gr}^2 is given by

$$S_{gr}^2 = \frac{1}{n-1} \left[\sum_{j=1}^C m_j^2 \times f_j - \frac{\left(\sum_{j=1}^C m_j \times f_j \right)^2}{n} \right], \text{ where } \sum_{j=1}^C m_j^2 \times f_j = \text{USS}_{gr}, \text{ and}$$

$$\frac{\left(\sum_{j=1}^C m_j \times f_j \right)^2}{n} = \text{CF}_{gr} = \text{Grouped Correction Factor.}$$

Bonus Homework 2 (7 Points). Compute the 0.25, 0.75 quantiles and the resulting IQRs of the Example 1.9 atop page 16 of Devore's 7th edition from 3 different methods: (1) SAS's, (2) Minitab's and verify your answers by submitting the Minitab output, (3) Excel's using its percentile function. Further, fully explain how MS Excel computes the two quartiles, i.e., what formulation (not just the syntax) Excel uses to obtain sample percentiles. (4) Provide two data sets each of size $n = 6$ such that data set 1 has a larger range (R) than data set 2, but data set 1 has smaller standard deviation, S , than data set 2. Note that it will be impossible to create such 2 samples iff both sizes are $n = 2$.