STAT 3600 Reference: Chapter 1 of Devore's 8th Ed. Maghsoodloo

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 ≅ 33000 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 (College of engineering) students, where the frame can be formed if need be.

Examples 1.1 and 1.2 on pp. 4-6, 1.5 on p. 11, 1.11 on pp. 20-21, and 1.14 on p. 29 of Devore's 8th edition provide sampling from conceptual (or virtual) populations, while Example 1.20 on p. 41 of Devore is from a concrete population.

A variable, X, is any (performance) 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: 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 1.11 on pp. 20-21, sample size n = 48), Specific Gravity of Exercise 12 on p. 24 and Shear Strength (lb) of Exercise 24 on p. 26 of Devore, etc.

Note that the late W. Edwards Deming (perhaps the most prominent of Quality gurus in the 20th century who also was responsible for the Japan's Quality evolution after World War II, starting with late 1940's to early 1960's) 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

(1) 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, median, the mode, percentiles, etc.), of Variability, of Skewness, and of Kurtosis.

1(i) Stem-and-leaf Plot for the Exercise 12 on page 24 of Devore's (8e)

The data is already in order-statistics format with $x_{(1)} = 0.31$ and $x_{(n)} = x_{(36)} = 0.75$. The sample size, universally denoted by n, is equal to 36. The variable X = Specific Gravity (a quantitative measure). Stem = 0.10 (The same as Minitab's increment = 0.10) and Minitab's Leaf unit = 0.01.

($Cumf_i$)	Stem	Leaf	(n/2 = 18)
6	3	156678	
(19)	4	0001122222345667888	
11	5	14458	
6	6	26678	
1	7	5	

I will name the increment "0.4" as the median stem for the above data because its sample median lies in the interval $0.40 \le \tilde{x} = \hat{x}_{0.50}$ = Sample Median = 0.4450 < 0.50.

Histograms. (See the Example 1.10 on pp. 18-19 of Devore's 8th 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. Sample range $R = x_{(90)} - x_{(1)} = 15.29$, C = No. of subgroups (or classes, or bins) for which there are 3 statistical guidelines: $C_1 = 1 + 3.3 \times \log_{10}(n)$ [which is called Sturges' practical guideline], $C_1 = 1+3.3 \log_{10}(90) = 7.45$ (use for n < 125), $C_2 \cong \sqrt{n}$, $(125 \le n < 600) \rightarrow C_2 = 9.49$, 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 > 600. Thus, it is best to select between 7 to 10 subgroups. So, we choose C = 9 subgroups. As a result, Δ_j = j^{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 1.10 of Devore is given Table 1 below; the $\sum_{j=1}^{C} f_j$ must always add to n (= 90 in this case). This

Table 1. The Frequency distribution of Example 1.10, on pp. 18-19

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

is why the subgroup intervals must be non-overlapping. The histogram from Minitab is provided in Figure 1. In Figure 1, the area inside 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" $= \sum_{j=1}^{C} Relf_j \equiv 1 = \sum_{j=1}^{C} h_j \Delta_j = \sum_{j=1}^{C} d_j \Delta_j$, and because both ($Relf_j$, $d_j\Delta_j$) represent the same jth 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, ..., C. For the histogram of Figure 1, the density $d_1 = (2/90)/1.70 = 0.02222/1.70 = 0.013072 = h_1$, $d_2 = h_2 = 0.05555/1.70 = 0.032676$/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 jth rectangular area $a_j = Relf_j = d_j \times \Delta_j$. Further, all histograms should be constructed with equal Δ_j's; if this is impossible, then the histogram-ordinate must

always be expressed in terms of densities. Finally, 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. For example, $m_1 = (4.66+2.97)/2 =$

Minitab Project Report

The Histogram for the Exp1.10 on pages 18-19 of Devore's 8th edition based on midpoints (m_j) and 9 subgroups each of length 1.70. The values inside each bar represent the *Relf*$_j$

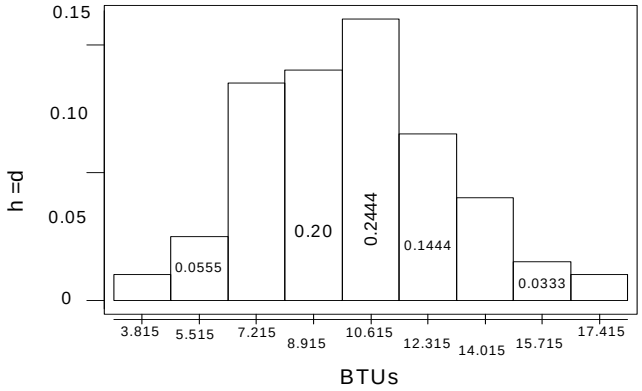


Figure 1

3.815, $m_2 = (6.36+4.67)/2$,…, and $m_9 = (18.26+16.57)/2 = 17.415$. The 3rd pictorial summary, the Boxplot, will be discussed on pp. 11-12 of these notes. Note that the *Relf*$_j$'s are unit-less while d_j's always have units.

1 (ii) (Quantitative) Measures in Descriptive Statistics From Data

(a) Measures of Location are: The sample Mean (or arithmetic average), median, sample geometric mean, harmonic mean, trimmed mean, mode, and sample quantiles (or percentiles). A bar is almost universally used to denote sample averages such as the arithmetic mean $\overline{x}$ (or $\overline{y}$). The arithmetic mean is defined as $\overline{x}$ =

$\sum_{i=1}^{n} x_i / n$. For the example 1.11 of Devore's 8th edition, p. 20, n = 48, $\sum_{i=1}^{n} x_i$ = 387.80 $\rightarrow$ $\bar{x}$ = 8.0792; the corresponding Bond-Strength data, in order-statistics format, are reproduced below for your convenience. Note that the sample mean represents

3.40	3.60	3.60	3.60	3.60	3.70	3.80	3.80	3.90	4.00	4.10	4.20
4.80	4.90	5.00	5.10	5.10	5.20	5.20	5.20	5.40	5.50	5.60	5.70
6.20	6.60	7.00	7.60	7.80	8.20	8.50	8.90	9.30	9.30	9.90	10.70
10.70	11.50	12.10	12.60	13.10	13.40	13.80	14.20	15.20	17.10	20.60	25.50

the center of gravity of a data set and must carry at least one more decimal than the original data; see Figure 1.15 on p. 29 of Devore's 8th edition. Please note that in the above data set 3.40 is called the 1st-order statistic, i.e., $x_{(1)}$ = 3.40, 3.60 = the 2nd-order statistic = $x_{(2)}$, …, 25.50 = $x_{(n)}$= $x_{(48)}$. Clearly, a 1st-order statistic is the smallest element of a sample, and the nth-order statistic is the largest observation in a sample of size n for all random samples in the universe.

The median, $\tilde{x} = \hat{x}_{0.50}$, is another measure of central location (or tendency) 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 0.50 by n +1. If this result is an exact integer, say r, then $\hat{x}_{0.50}$ = $x_{(r)}$; if 0.50(n+1) is not an exact integer, then $\hat{x}_{0.50}$ = 0.50$x_{(r)}$ + 0.50$x_{(r+1)}$. When n is an even integer, then exactly half the data will lie below and the other half above the sample median $\tilde{x} = \hat{x}_{0.50}$. Thus, for the Example 1.11, 0.50(n+1) = 24.50 so that $\hat{x}_{0.50}$ = 0.50$x_{(24)}$ + 0.50$x_{(25)}$ = 0.50(5.70) + 0.50(6.20) = 5.950; note that exactly 24 data points lie below, and 24 data points (or observations) lie above 5.950. However, for the data of Example 1.14 on p. 29 of Devore's 8th edition, the sample size n = 21 (odd integer) gives 0.50×(n+1) = 11 = r, which is an exact integer. Thus, $\hat{x}_{0.50}$ = $\tilde{x}$ = $x_{(11)}$ = 21.20, while $\bar{x}$ = 21.181. Note that in this case 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 = $x_{(11)}$.

The geometric mean is defined as $\bar{x}_g$ = $(x_1x_2…x_n)^{1/n}$, i.e., $\bar{x}_g$ is the nth root of

($\prod_{i=1}^{n} x_i$) only if all x_i's > 0 for all i, and in general $\bar{x}_g \le \bar{x}$. For the data of example 1.14 of Devore on p. 29 of his 8th edition, $\bar{x}_g$ = 19.379764 < 21.180952 = $\bar{x}$. Geometric mean has limited applications in DOX (Design of Experiments; some authors use the acronym DOE) 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 1.14 on page 29 of Devore, n = 21, and $\frac{1}{21}\sum_{i=1}^{21}(1/x_i)$ = [(1/16.1) + (1/9.6) + ... + (1/28.5)]/21 = 1.1902504/21 = 0.0566786, which yields $\bar{x}_h = 1/0.0566786$ = 17.643346 < $\bar{x}_g$ < $\bar{x}$. 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 always $\bar{x}_h \le \bar{x}_g \le \bar{x}$. In general, the geometric and harmonic means are not as important measures of central tendency as $\bar{x}$ and $\tilde{x}$. The sample mean, $\bar{x}$, is the most common measure of central tendency and always is the central gravity of a data set.

TRIMMED MEANS

A 10% trimmed mean, $\bar{x}_{tr(10)}$ = TrMean, is computed by deleting the smallest and largest 0.10×n of the order-statistics from the two 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 $x_1, x_2, \ldots, x_n$. However, our author, Devore, is notationally consistent with other statistical literature, and therefore, we will also use Devore's notation of $\bar{x}_{tr(10)}$. To illustrate, consider the Bond-Strength data of Example 1.11, of size n = 48, on page 20 of Devore (8e), for which 0.10×n = 4.8.

Step 1. Trim or remove the order-statistics $x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)}, x_{(48)}, x_{(47)}, x_{(46)}$, and

$x_{(45)}$. Next, compute $\sum_{i=5}^{44} x_{(i)}$ / 40 = 7.380 = $\bar{x}_{tr4}$.

Step 2. Trim $\bar{x}_{tr4}$ further by removing $x_{(5)}$ and $x_{(44)}$ in order to obtain $\bar{x}_{tr5}$.

$$\bar{x}_{tr5} = \sum_{i=6}^{43} x_{(i)} / 38 = 7.300.$$

Step 3. Interpolate between $\bar{x}_{tr4}$ and $\bar{x}_{tr5}$ to obtain $\bar{x}_{tr(10)}$, i.e., the exact trimmed mean for the Example 1.11 should be computed from the following convex combination.

$$\bar{x}_{tr(10)} = 0.2\times\bar{x}_{tr4} + 0.8\times\bar{x}_{tr5} = 7.3160$$

Had n been equal to 44, then 0.10×n = 4.4 and the above formula would change to $\bar{x}_{tr(10)}$ = $0.6\times\bar{x}_{tr4}$ + $0.4\times\bar{x}_{tr5}$. Note that most statistical packages, such as Minitab, only give the $\bar{x}_{tr(5)}$ and they round the value of 0.05×n to the nearest integer in order to obtain the 5% trimmed mean $\bar{x}_{tr(5)}$.

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 $\hat{x}_{0.50}$ (= 5.950 for the Example 1.11) than is $\bar{x}$ = 8.0792.

The MODE

The mode is the observation with the highest frequency. For the data of Example 1.11 on p. 20 of Devore, MO = 3.6 and modal frequency *f* = 4 (this is the highest frequency). 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, y, occurs.

Computing Sample Percentiles (or Quantiles)

The p^{th} sample quantile (or 100×pth percentile), $\hat{x}_p$, is obtained using the following steps.

(1) First rearrange the data in ascending order of $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$, where $x_{(1)}$ is called the 1^{st}-order statistic, $x_{(2)}$ the 2^{nd}-order statistic, …., $x_{(n)}$ is called the n^{th}-

order statistic. The nth-order statistic is always the maximum of the sample.

(2) Multiply p by n+1: if $(n+1)\times p$ is an exact integer, say I, then $\hat{x}_p = x_{(I)}$.

(3) If $(n+1)\times p$ is not an exact integer such that $I < (n+1)p < I+1$, then the sample p^{th}-quantile is given by the convex combination $\hat{x}_p = a x_{(I)} + (1-a)x_{(I+1)}$, where $0 < a = (I+1) - (n+1)p < 1$. For the data of Example 1.11 on p. 20 of Devore's 8(e), where X represents Bond Strength (BNDS in psi), the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 80^{th}, and 90^{th} sample quantiles (or percentiles) are computed below: (Note that only for convenience hats have been removed from sample percentiles, and the sample size is n = 48 so that n +1 = 49)

$x_{0.10}$: $0.10\times 49 = 4.9 \rightarrow x_{0.10} = 0.10x_{(4)}+0.90x_{(5)} = 3.60$

$x_{0.25}$: $0.25\times 49 = 12.25 \rightarrow x_{0.25} = 0.75x_{(12)} + 0.25x_{(13)} = 4.35$

$x_{0.50}$: $0.50\times 49 = 24.5 \rightarrow x_{0.50} = 0.5x_{(24)} + 0.5x_{(25)} = 5.950$

$x_{0.75}$: $0.75\times(n+1) = 36.75 \rightarrow x_{0.75} = 0.25x_{(36)} + 0.75x_{(37)} = 10.70$

$x_{0.80}$: $0.80\times 49 = 39.2 \rightarrow x_{0.80} = 0.80x_{(39)} + 0.20x_{(40)} = 12.20$

$x_{0.90}$: $0.90\times(n+1) = 44.1 \rightarrow x_{0.90} = 0.9x_{(44)} + 0.10x_{(45)} = 14.30$.

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 1^{st} decile, and the 0.90 quantile is called the 9^{th} decile. Every data set has 9 sample deciles.

Minitab's Descriptive Statistics: BNDS (Example 1.11 pp. 20-21)

Variable	Mean	SE Mean	TrMean	StDev	Variance	CoefVar	Sum
BNDS	8.079	0.703	7.607	4.868	23.702	60.26	387.800

Sum of Squares	Minimum	Q1	Median	Q3	Maximum	Range	IQR
4247.080	3.400	4.350	5.950	10.700	25.500	22.100	6.350

Mode	N for Mode	Skewness	Kurtosis
3.6	4	1.54	2.64

(b) Measures of Variability (Three Quantitative Measures)

(1) Standard deviation (Stdev) = S, **(2)** Sample Range/d_2 = R/d_2, and **(3)** the IQR = $x_{0.75} - x_{0.25}$, where the sample 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, and for the most common QC sample size n = 5 the corresponding value of d_2 is

approximately equal to 2.325929, and d_2 is an increasing function of n (at n = 10, d_2 = 3.077505). 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 first; there are no other alternatives.

Definition. The sample variance, var, 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, Sample range R = 2.7, USS = $\sum_{i=1}^{n} x_i^2$ = 84.30, CF = Correction Factor = $(\sum_{i=1}^{n} x_i)^2 / n = n(\bar{x})^2 = 20^2/5 = 80 \rightarrow$

$x_i - \bar{x} = x_i - 4 = -1.30, -0.50, -0.20, 0.60, 1.40 \longrightarrow \sum_{i=1}^{5}(x_i - \bar{x}) = \sum_{i=1}^{5}(x_i - 4) \equiv 0$

$(x_i - 4)^2$: 1.69, 0.25, 0.04, 0.36, 1.96, $\longrightarrow$ $\sum_{i=1}^{5}(x_i - \bar{x})^2 = 4.30$ $\longrightarrow$ Sample variance $var_1 = 4.3/5 = 0.86$. 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 $\rightarrow$ $\sum_{i=1}^{5}(x_i - \bar{x}) = \sum_{i=1}^{5}(x_i - 4.0) = 0$

$(x_i - \bar{x})^2$: 3.61, 0.64, 0.16, 0.25, 6.76 $\longrightarrow$ CSS = Corrected Sum of Squares = S_{xx}

= 3.61+ 0.64 + 0.16 + 0.25 + 6.76 = $\sum_{i=1}^{5}(x_i - \bar{x})^2 = \sum_{i=1}^{5}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) = \sum_{i=1}^{5} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i +$

$\sum_{i=1}^{n}\bar{x}^2 = USS - 2\bar{x}(n\bar{x}) + n\bar{x}^2 = USS - n\bar{x}^2 = USS - n(\sum_{i=1}^{n} x_i / n)^2 = USS - (\sum_{i=1}^{n} x_i)^2 / n = USS$

– CF = 91.42 – 80 = 11.42; thus, $var_2 = 11.42/5 = 2.284$.

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 $\longrightarrow$ CSS = S_{xx} = 13.16 $\rightarrow$ $var_3 = 13.16/5 = 2.632$.

Note that in general as the overall spread of the data increases, so does the variance, i.e., variance is a measure of variability. Further, the divisor of var is n, i.e., var = $(1/n)\times\Sigma (x_i -$

$\bar{x})^2$. Note that the n deviations from the mean $(x_1 - \bar{x})$, $(x_2 - \bar{x})$, $(x_3 - \bar{x})$,…, $(x_n - \bar{x})$ 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 number 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 $x_1 - \bar{x}$, $x_2 - \bar{x}$, …, $x_n - \bar{x}$ have $(n-1)$ degrees of freedom (*df*) not n, i.e., before the sample is taken, we have freedom to specify any of the $(n-1)$ of them, and the nth deviation from the mean is automatically determined from $\sum_{i=1}^{n}(x_i - \bar{x}) \equiv 0$. Therefore, we define the most common measure of variability with the divisor of $(n-1)$ for $S_{xx} = \Sigma(x_i - \bar{x})^2$, given by

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

For data above sets 1, 2 and 3 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), $S_1^2 = 5 \times var_1/4 = 5 \times 0.86/4 = 1.075$, and so forth. The reader should deduce from above examples that the USS plays a much more important role in determining the value of S^2 than does the CF.

The exact name for S^2 is not the sample variance as defined by Devore on his p. 36. In actuality the sample variance is $var = \Sigma (x_i - \bar{x})^2 / n$ as defined by me herein, but var 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 $CSS = S_{xx} = \Sigma(x_i - \bar{x})^2$ by a smaller number than n, namely its *df* (*degrees of freedom*) = $n - 1$, 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 sample 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, but most commonly, 2 decimals.

The IQR (InterQuartile Range) is defined as IQR = $\hat{x}_{0.75} - \hat{x}_{0.25}$ = Q3 – Q1 = the Devore's 4th spread, while the interdecile range is defined by me as $\hat{x}_{0.90} - \hat{x}_{0.10}$. The 4th spread, f_s = Q3 – Q1, is Devore's uncommon terminology explained near the bottom of his p. 39, and his terminology should be avoided. For the Example 1.11 of Devore (8e), the value of Interquartile-range is equal to IQR =10.70 – 4.35 = 6.350.

Identifying Outliers

If Q1 – 3×IQR < $x_{(i)}$ < Q1 – 1.5 × IQR, or Q3 + 1.5×IQR < $x_{(i)}$ < Q3 + 3×IQR, then the i[th] order-statistic, $x_{(i)}$, is a mild outlier. If the value of $x_{(i)}$ < Q1 – 3×IQR, or $x_{(i)}$ > Q3 + 3×IQR, then $x_{(i)}$ is an extreme outlier. For the Example 1.11 of Devore, since Q1 – 1.5×6.35 = 4.35 – 9.525 < 0 and Q3 +1.5×IQR = 20.225, then the data contain 2 outliers on the RHS (or the upper tail). Further, because Q3+ 3×IQR = 29.75, then the data has no extreme outliers.

Graphical Measure of Variability (The Boxplot)

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

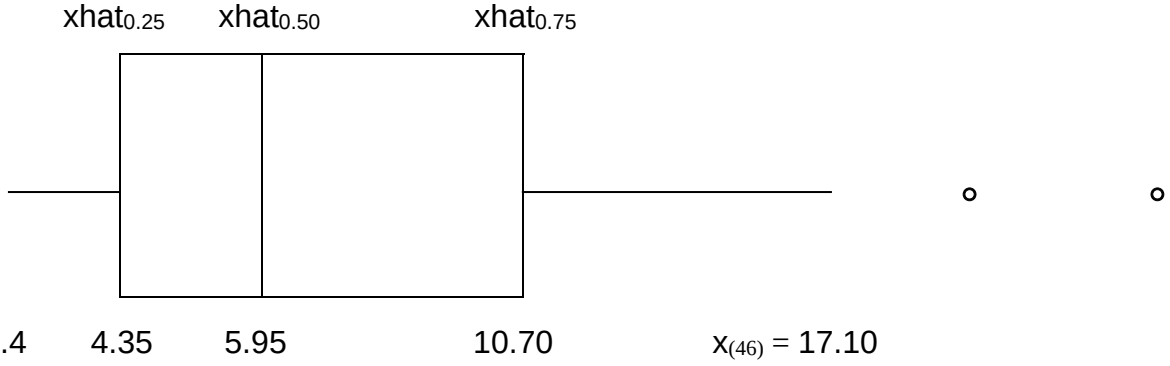
Step 2. Draw vertical lines thru Q1 = $\hat{x}_{0.25}$ and Q3 = $\hat{x}_{0.75}$ and connect at the bottom and the top to make a rectangular box. For the data of Example 1.11 on p. 20 of Devore (8e), the box is shown atop the next page, where hats are removed from sample percentiles only for convenience.

Step 3. Compute both 1.5×IQR and 3×IQR. For the Example 1.11 on p. 20, 1.5×IQR = 9.525, which yields the mild interval (Q1 – 1.5×IQR = – 5.175, Q3 +1.5×IQR = 20.225). If the entire data lies in this last interval, then the data has no outliers. Thus, the data of Example 1.11 contain 2 outliers on the right tail, namely 20.60 & 25.50. Because, Q3+3×IQR = 10.70 +19.050 = 29.750, then both outliers are mild. Note that the dots on the RHS of the Boxplot represent the mild outlier $x_{(47)}$ = 20.6 and $x_{(48)}$ = 25.50.

Step 4. Draw whiskers from Q1 and Q3 to the smallest and largest order statistics that are not outliers. The Box-plot is given atop the next page.

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

The Box-Plot for the Example 1.11 on page 20 of Devore's 8th Edition



$= \bar{x}$. **(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 = CSS = $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ = USS – CF , where the Uncorrected Sum of Squares USS= $\sum_{i=1}^{n} x_i^2$, and the correction factor CF = $(\sum_{i=1}^{n} x_i)^2 / n = n(\bar{x})^2$. **(d)** By definitions the mean and variance of a grouped (gr) data (or an empirical distribution) are given by

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

Prove that for a histogram (or a frequency distribution) $\sum_{j=1}^{C}(m_j - \bar{x}_{gr}) \times f_j \equiv 0$,

and that the computing formula for S_{gr}^2 is given by $S_{gr}^2 = CSS_{gr} / (n-1) = \frac{1}{n-1}[\sum_{j=1}^{C} m_j^2 \times f_j - \frac{(\sum_{j=1}^{C} m_j \times f_j)^2}{n}]$, where $\sum_{j=1}^{C} m_j^2 \times f_j = USS_{gr}$, and $\frac{(\sum_{j=1}^{C} m_j \times f_j)^2}{n} = CF_{gr}$ = Grouped Correction Factor.

My Chapter 1 notes have been edited mostly by Rong Huangfu (rzh0024) and also by Mohammad-Ali Alamdar-Yazdi (mza0052). S. Maghsoodloo (08/23/2014)