2-5 Measures of Variation

NOTE: Although not given in the text, the symbol R will be used for the range throughout this manual. Remember that the range is the difference between the highest and the lowest scores, and not necessarily the difference between the last and the first values as they are listed. Since calculating the range involves only the subtraction of 2 original pieces of data, that measure of variation will be reported with the same accuracy as the original data.

NOTE OF EXPLANATION: Problem #1 will be worked in detail using formulas 2-4 and 2-5. Problems #2-8 will be worked using formula 2-5. Whether using formulas or Excel, always verify that the answer for s is reasonable -- it should measure the amount that a typical score deviates from the mean. In Excel, using the **descriptive statistics** command to find a several values instead of finding only the value of interest helps prevent errors by providing a quick check that the values are correct for n, maximum, minimum, etc.

1.

x	$x-\bar{x}$	$(x-\bar{x})^2$	x^2
8	-8	64	64
11	-5	25	121
13	-3	9	169
14	-2	4	196
14	-2	4	196
14	-2	4	196
15	-1	1	225
16	0	0	256
17	1	1	289
18	2	4	324
25	9	81	625
27	11	121	729
192	0	318	3390

$\bar{x} = (\Sigma x)/n = 192/12 = 16.0$

$R = 27 - 8 = 19$

by formula 2-4,

$s^2 = \Sigma(x-\bar{x})^2/(n-1)$
$= 318/11$
$= 28.9091$
$= 28.9$

by formula 2-5,

$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [12(3390) - (192)^2]/[12(11)]$
$= [3816]/[132]$
$= 28.9091$
$= 28.9$

$s = \sqrt{28.9091} = 5.4$

NOTE: When finding the square root of the variance to obtain the standard deviation, use <u>all</u> the decimal places of the variance, and not the rounded value reported as the answer. The best way to do this is either to keep the value on the calculator display or to place it in the memory. Do <u>not</u> copy down all the decimal places and then re-enter them to find the square root, as that could introduce round-off and/or copying errors.

When using formula 2-4, constructing a table having the first three columns shown above helps to organize the calculations and makes errors less likely. In addition, verify that $\Sigma(x-\bar{x}) = 0$ before proceeding -- if such is not the case, there is an error and further calculation is fruitless. For completeness, and as a check, both formulas 2-4 and 2-5 were used above. In general, formula 2-5 is to be preferred for the following reasons:

(1) When the mean does not "come out even," formula 2-4 involves round-off error and/or many messy decimal calculations.

(2) The quantities Σx and Σx^2 needed for formula 2-5 can be found directly from the original data without having to construct a table like the one above.

(3) The formula includes an automatic check. The quantity $[n(\Sigma x^2) - (\Sigma x)^2]$ cannot be less than zero. A negative value indicates an error and that further calculation is fruitless.

3. preliminary values: $n = 12$, $\Sigma x = 972$, $\Sigma x^2 = 86424$
$R = 119 - 35 = 84$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [12(86424) - (972)^2]/[12(11)]$
$= (92304)/132 = 699.3$
$s = 26.4$

5. Jefferson Valley
$n = 10$, $\Sigma x = 71.5$, $\Sigma x^2 = 513.27$
$R = 7.7 - 6.5 = 1.2$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [10(513.27) - (71.5)^2]/[10(9)]$
$= 20.45/90 = 0.23$
$s = 0.48$

Providence
$n = 10$, $\Sigma x = 71.5$, $\Sigma x^2 = 541.09$
$R = 10.0 - 4.2 = 5.8$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [10(541.09) - (71.5)^2]/[10(9)]$
$= 298.65/90 = 3.32$
$s = 1.82$

Exercise #5 of section 2-4 indicated that the mean waiting time was 7.15 minutes at each bank. The Jefferson Valley waiting times, however, are considerably less variable. The range measures the difference between the extremes. The longest and shortest waits at Jefferson Valley differ by a little over 1 minute (R=1.2), while the longest and shortest waits at Providence differ by almost 6 minutes (R=5.8). The standard deviation measures the typical difference from the mean. A Jefferson Valley customer usually receives service within about ½ minute (s=0.48) of 7.15 minutes, while a Providence customer usually receives service within about 2 minutes (s=1.82) of the mean.

6. Regular
$n = 6$, $\Sigma x = 4.9144$, $\Sigma x^2 = 4.02528224$
$R = .8247 - .8150 = .0097$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [6(4.02528224) - (4.9144)^2]/[6(5)]$
$= .00036608/30 = .000012202$
$s = .00349$

Diet
$n = 6$, $\Sigma x = 4.7000$, $\Sigma x^2 = 3.68181990$
$R = .7896 - .7758 = .0138$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [6(3.68181990) - (4.7000)^2]/[6(5)]$
$= .00091940/30 = .000030633$
$s = .00554$

There appears to be slightly more variation among the Diet Coke weights than among those for Regular Coke.

7. Coke
$n = 6$, $\Sigma x = 4.9144$, $\Sigma x^2 = 4.02528224$
$R = .8247 - .8150 = .0097$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [6(4.02528224) - (4.9144)^2]/[6(5)]$
$= .00036608/30 = .000012202$
$s = .00349$

Pepsi
$n = 6$, $\Sigma x = 4.9313$, $\Sigma x^2 = 4.05310181$
$R = .8302 - .8156 = .0146$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [6(4.05310181) - (4.9313)^2]/[6(5)]$
$= .00089117/30 = .000029706$
$s = .00545$

There appears to be slightly more variation among the Pepsi weights than among those for Coke. In other words, Coke seems to be doing slightly better when it comes to producing a uniform product.

NOTE: Following the usual round-off rule of giving answers with one more decimal than the original data produces variances of .00001 and .00003, which have only one significant digit. This is an unusual situation occurring because values less than 1.0 become smaller when they are squared. Since the original data had 4 significant digits, we provide 5 significant digits for the variance.

9.

x	f	f·x	f·x²
44.5	8	356.0	15842.00
54.5	44	2398.0	130691.00
64.5	23	1483.5	95685.75
74.5	6	447.0	33301.50
84.5	107	9041.5	764006.75
94.5	11	1039.5	98232.75
104.5	1	104.5	10920.25
	200	14870.0	1148680.00

$s^2 = [n(\Sigma f\cdot x^2) - (\Sigma f\cdot x)^2]/[n(n-1)]$
$= [200(1148680.00) - (1487.0)^2]/[200(199)]$
$= (8619100.00)/29800 = 216.56$
$s = 14.7$

11.

x	f	f·x	f·x²
43.5	25	1087.5	47305.25
47.5	14	665.0	31587.50
51.5	7	360.5	18565.75
55.5	3	166.5	9240.75
59.5	1	59.5	3540.25
	50	2339.0	110239.50

$s^2 = [n(\Sigma f\cdot x^2) - (\Sigma f\cdot x)^2]/[n(n-1)]$
$= [50(110239.50) - (2339.0)^2]/[50(49)]$
$= (41054.00)/2450 = 16.76$
$s = 4.1$

13. author's college: n=35, s=14.683
UMASS: n=40, s=23.076
There appears to be more variation among prices of new textbooks at UMASS than at the author's college.

15. Thursday: n=52, s=.167
Sunday: n=52, s=.200
The values are close, although there may be slightly more variation among the amounts for Sunday than among the amounts for Thursdays.

17. Assuming the graduate ages range from 17 to 19, the Range Rule of Thumb suggests $s \approx \text{range}/4 = (19-17)/4 = 2/4 = 0.5$.

19. Given $\bar{x}=75$ and $s=12$, the Range Rule of Thumb suggests
minimum "usual" value $= \bar{x} - 2s$
$= 75 - 2(12) = 75 - 24 = 51$
maximum "usual" value $= \bar{x} + 2s$
$= 75 + 2(12) = 75 + 24 = 99$
Yes, in this context a score of 50 would be considered unusually low.

21. a. The limits 61.1 and 66.1 are 1 standard deviation from the mean. The Empirical Rule for Data with a Bell-shaped Distribution states that about 68% of the heights should fall within those limits.
b. The limits 56.1 and 76.1 are 3 standard deviations from the mean. The Empirical Rule for Data with a Bell-shaped Distribution states that about 99.7% of the heights should fall within those limits.

23. The limits 58.6 and 68.6 are 2 standard deviations from the mean. Chebyshev's Theorem states that there must be at least $1 - 1/k^2$ of the scores within k standard deviations of the mean. Here k = 2, and so the proportion of the heights within those limits is at least $1 - 1/2^2 = 1 - 1/4 = 3/4 = 75\%$.

25. A standard deviation of s = 0 is possible only when $s^2 = 0$, and $s^2 = \Sigma(x-\bar{x})^2/(n-1) = 0$ only when $\Sigma(x-\bar{x})^2 = 0$. Since each $(x-\bar{x})^2$ is non-negative, $\Sigma(x-\bar{x})^2 = 0$ only when every $(x-\bar{x})^2 = 0$ -- i.e., only when every x is equal to $\bar{x}$. In simple terms, no variation occurs only when all the scores are identical.

27. The Everlast brand is the better choice. In general, a smaller standard deviation of lifetimes indicates more consistency from battery to battery -- signaling a more dependable production process and a more dependable final product. Assuming a bell-shaped distribution of lifetimes, for example, that empirical rule states that about 68% of the lifetimes will fall within one standard deviation of the mean. Here, those limits would be

   for Everlast: 50 ± 2 or 48 months to 52 months
   for Endurance: 50 ± 6 or 44 months to 56 months

While a person might be lucky and purchase a long-lasting Endurance battery, an Everlast battery is much more likely to last for the advertised 48 months.

29. section 1

$n = 11, \Sigma x = 201, \Sigma x^2 = 4001$
$R = 20 - 1 = 19$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [11(4001) - (201)^2]/[11(10)]$
$= 3610/110 = 32.92$
$s = 5.7$

section 2

$n = 11, \Sigma x = 119, \Sigma x^2 = 1741$
$R = 19 - 2 = 17$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [11(1741) - (119)^2]/[11(10)]$
$= 1990/110 = 45.36$
$s = 6.7$

The range values give the impression that section 1 had more variability than section 2. The range can be misleading because it is based only on the extreme scores. In this case, the lowest score in section 1 was so distinctly different from the others that to include it in any measure trying to give a summary about the section as a whole would skew the results. For the mean, where the value is only one of 11 used in the calculation, the effect is minimal; for the range, where the value is one of only 2 used in the calculation, the effect is dramatic. The standard deviation values give the impression that section 2 had slightly more variability.

NOTE: In this case, section 2 seems considerably more variable (or diverse), and even the standard deviation by itself fails to accurately distinguish between the sections.

31. preliminary values: $n = 11, \Sigma x = 56, \Sigma x^2 = 396$

$\bar{x} = (\Sigma x)/n = 56/11 = 5.09$
$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$
$= [11(396) - (56)^2]/[11(10)]$
$= (1120)/110 = 11.09$
$s = 3.33$

a. The coefficient of variation is $(s/\bar{x})\cdot 100\% = (3.33/5.09)\cdot 100\% = 65.4\%$.

b. The signal-to-noise ratio is $\bar{x}/s = 5.09/3.33 = 1.529$.

33. The largest standard deviation (i.e., the most spread) occurs when all the values are evenly distributed between the two extremes. Here that would be: 1, 1, 1, 9, 9, 9.
preliminary values: $n = 6$, $\Sigma x = 30$, $\Sigma x^2 = 246$

$\bar{x} = (\Sigma x)/n = 30/6 = 5.0$

$s^2 = [n(\Sigma x^2) - (\Sigma x)^2]/[n(n-1)]$

$= [6(246) - (30)^2]/[6(5)]$

$= (576)/30 = 19.20$

$s = 4.4$

35. For greater accuracy and understanding, use 3 decimal places and avoid shortcut formulas.

a. the original population

x	x-μ	(x-μ)²
1	-1	1
2	0	0
3	1	1
6	0	2

$\mu = (\Sigma x)/N = 6/3 = 2$

$\sigma^2 = \Sigma(x-\mu)^2/N = 2/3 = .667$

$\sigma = .816$

b. the nine samples: using $s^2 = \Sigma(x-\bar{x})^2/(n-1)$ [for each sample, n = 2]

sample	x̄	s²	s
1,1	1.0	0	0
1,2	1.5	0.5	0.707
1,3	2.0	2.0	1.414
2,1	1.5	0.5	0.707
2,2	2.0	0	0
2,3	2.5	0.5	0.707
3,1	2.0	2.0	1.414
3,2	2.5	0.5	0.707
3,3	3.0	0	0
	18.0	6.0	5.656

mean of the 9 calculated variances

$(\Sigma s^2)/9 = 6.0/9 = 2/3 = .667$

c. the nine samples: using $\sigma^2 = \Sigma(x-\mu)^2/N$ [for each sample, N = 2]

sample	μ	σ²
1,1	1.0	0
1,2	1.5	0.25
1,3	2.0	1.00
2,1	1.5	0.25
2,2	2.0	0
2,3	2.5	0.25
3,1	2.0	1.00
3,2	2.5	0.25
3,3	3.0	0
	18.0	3.00

mean of the 9 calculated variances

$(\Sigma\sigma^2)/9 = 3.0/9 = 1/3 = .333$

d. The approach in (b) of dividing by n-1 when calculating the sample variance gives a better estimate of the population variance. On the average, the approach in (b) gave the correct population variance of 2/3 = .667. The approach in (c) of dividing by n underestimated the correct population variance. When computing sample variances, divide by n-1 and not by n.

e. No. An unbiased estimator is one that gives the correct answer on the average. Since the average value of s^2 in part (b) was .667, which was the correct value calculated for σ^2 in part (a), s^2 is an unbiased estimator of σ^2. Since the average value of s in part (b) is $(\Sigma s)/9 = 5.656/9 = .628$, which is not the correct value of .816 calculated for σ in part (a), s is not an unbiased estimator of σ.

NOTE: Since the average value of $\bar{x}$ in part (b) is $(\Sigma\bar{x})/9 = 18.0/9 = 2.0$, which is the correct value calculated for μ in part (a), $\bar{x}$ is an unbiased estimator of μ.