Describing, Exploring, and Comparing Data

Glossary

1. **Descriptive Statistics**: summarize or describe relevant characteristics of data. **Inferential Statistics**: to make inferences, or generalizations, about populations
2. A measure of center is a value at the center or middle of a data set.
3. **Mean**: The **mean** (or arithmetic mean) of a set of data is the measure of center found by adding all of the data values and dividing the total by the number of data values.
    a. Sample means drawn from the same population tend to vary less than other measures of center.
    b. The mean of a data set uses every data value.
    c. A disadvantage of the mean is that just one extreme value (outlier) can change the value of the mean substantially.
4. Resistant: A statistic is resistant if the presence of extreme values (outliers) does not cause it to change very much.

5. small n: mean of a set of sample values
   big N: mean of a set of values in population.
6. Median: Measure of the center that is the middle value when originally nal data values are arranged in order of increasing (or decreasing) magnitude.
   a. median doesn't change with extreme outliers
   b. odd → median = middle number
   c. even → median = mean of 2 middle numbers.
7. **Mode**: value that occurs with the greatest frequency.
   a. the mode can be found with qualitative data.
   b. a dataset can have 0,1 or more modes.
   c. binomial → 2 modes multimodal → multiple modes.
8. **Midrange**: (max val - min val)/2
   a. it is very sensitive to extreme values as it uses min and max.
   b. mid range may or may not be the same as median.
9. Mean from Frequency Table

FORMULA 3-2 MEAN FROM A FREQUENCY DISTRIBUTION

First multiply each frequency and class midpoint; then add the products.

$$\bar{x} = \frac{\Sigma(f \cdot x)}{\Sigma f} \quad \text{(Result is an approximation)}$$

Sum of frequencies (equal to n)

TABLE 3-2 McDonald's Lunch Service Times

Time (seconds)	Frequency f	Class Midpoint x	$f \cdot x$
75–124	11	99.5	1094.5
125–174	24	149.5	3588.0
175–224	10	199.5	1995.0
225–274	3	249.5	748.5
275–324	2	299.5	599.0
Totals:	$\Sigma f = 50$		$\Sigma(f \cdot x) = 8025.0$

10. **Weighted Mean:**

FORMULA 3-3

$$\text{Weighted mean: } \bar{x} = \frac{\Sigma(w \cdot x)}{\Sigma w}$$

Formula 3-3 tells us to first multiply each weight w by the corresponding value x, then to add the products, and then finally to divide that total by the sum of the weights, Σw.

EXAMPLE 8 Computing Grade-Point Average

In her first semester of college, a student of the author took five courses. Her final grades, along with the number of credits for each course, were A (3 credits), A (4 credits), B (3 credits), C (3 credits), and F (1 credit). The grading system assigns quality points to letter grades as follows: A = 4; B = 3; C = 2; D = 1; F = 0. Compute her grade-point average.

SOLUTION

Use the numbers of credits as weights: $w = 3, 4, 3, 3, 1$. Replace the letter grades of A, A, B, C, and F with the corresponding quality points: $x = 4, 4, 3, 2, 0$. We now use Formula 3-3 as shown below. The result is a first-semester grade-point average of 3.07. (In using the preceding round-off rule, the result should be rounded to 3.1, but it is common to round grade-point averages to two decimal places.)

$$\begin{aligned}\bar{x} &= \frac{\Sigma(w \cdot x)}{\Sigma w} \\ &= \frac{(3 \times 4) + (4 \times 4) + (3 \times 3) + (3 \times 2) + (1 \times 0)}{3 + 4 + 3 + 3 + 1} \\ &= \frac{43}{14} = 3.07\end{aligned}$$

YOUR TURN Do Exercise 33 "Weighted Mean."

11. Variance: One can increase the quality of operations by reducing variance.
12. Range: Difference between max data value and min data val
    a. Range is sensitive to extreme outliers as it depends on min and max.
    b. doesn't truly reflect the variation among all the data points as it uses just min and max.
13. **Standard Deviation:** Measure of how much data values deviate away from the mean.

    s = sample std dev   mu = population std dev

Important Properties of Standard Deviation

a. The value of standard deviation is never negative. It is zero only when all of the data values are exactly the same.

b. standard deviation can increase dramatically with one or more outliers

c. sample standard deviation s is a biased estimator of the population standard deviation which means that values of the sample standard deviation s do not center around the value mu.

d. Tool : {https://www.calculator.net/standard-deviation-calculator.html} Also gives freq table if needed.

e. Significantly low values are muu - 2(stddev) or lower.

f. Significantly high values muu + 2(stddev)

g. estimating value of standard deviations → range /4

Values not significant: Between $(\mu - 2\sigma)$ and $(\mu + 2\sigma)$

See Figure 3-3, which illustrates the above criteria.

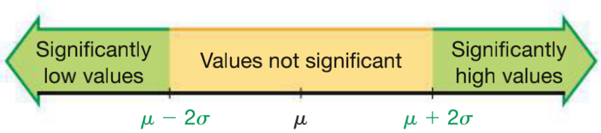


FIGURE 3-3 Range Rule of Thumb for Identifying Significant Values

14. **Variance**: The variance of a set of values is a measure of variation equal to the square of the
standard deviation.
   - Sample variance: s2 = square of the standard deviation s.
   - Population variance: s2 = square of the population standard deviation s.

   important Properties of variance

   a. units are squares eg. ft^2

b. the value of variance can increase dramatically with the inclusion of outliers.

c. The variance is never negative. It is zero when all of the data values are the same.

d. It is an unbiased estimator.

15. Critical Thinking

a. **Adding the deviations isn't good,**
because the sum will always be zero. To get a statistic that measures variation, it's
necessary to avoid the canceling out of negative and positive numbers. One approach is to
add absolute values, as in Σ x - x . If we find the mean of that sum, we get the mean
absolute deviation (or MAD)

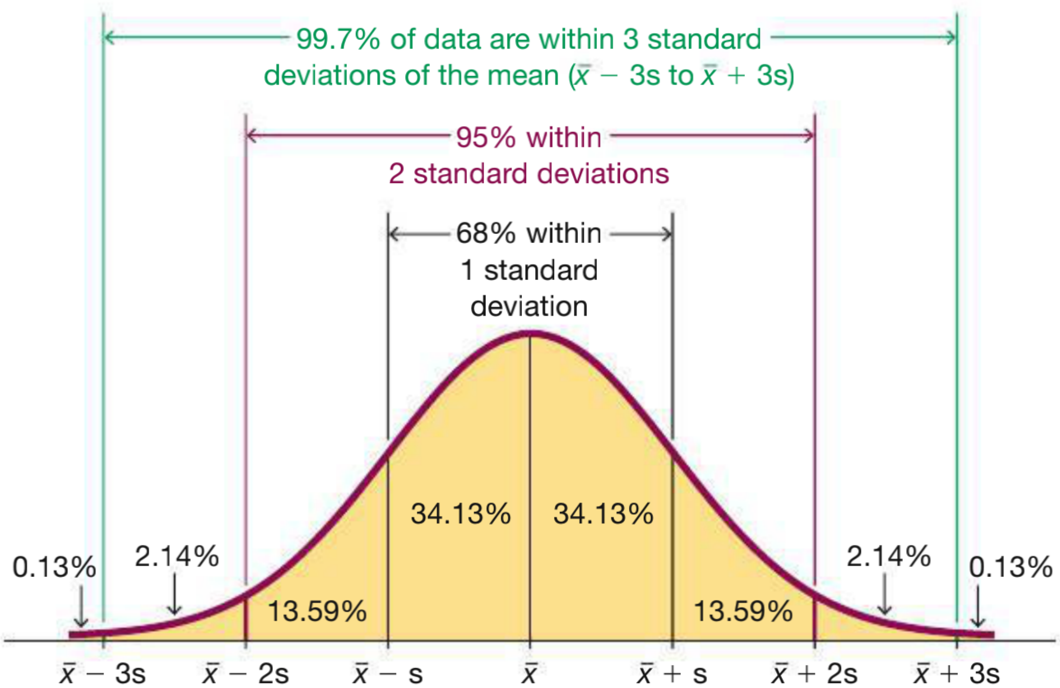
b. **Why Not Use the Mean Absolute Deviation Instead of the Standard Deviation?**
Because it is based on the square root of a sum of squares, the standard deviation closely parallels distance formulas found in algebra. There are many instances where a statistical procedure is based on a similar sum of squares. Consequently, instead of using absolute values, we square all deviations (x - x) so that they are nonnegative, and those squares are used to calculate the standard deviation.

c. **Why Divide by n − 1?** After finding all of the individual values of (x - x)2 we combine them by finding their sum. We then divide by n - 1 because there are only n - 1 values that can assigned without constraint. With a given mean, we can use any
numbers for the first n - 1 values, but the last value will then be automatically determined. With division by n - 1, sample variances s2 tend to center around the value of the population variance s2; with division by n, sample variances s2 tend to underestimate the value of the population variance s2.

d. **Empirical (or 68-95-99.7) Rule for Data with a Bell-Shaped Distribution**
A concept helpful in interpreting the value of a standard deviation is the empirical

rule. This rule states that for data sets having a distribution that is approximately bellshaped, the following properties apply. (See Figure 3-4.)

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.



16. **Chebyshev's Theorem**

    The proportion of any set of data lying within K standard deviations of the mean is

    always at least 1 - 1/K^2, where K is any positive number greater than 1. For K = 2

    and K = 3, we get the following statements:

    - At least 3/4 (or 75%) of all values lie within 2 standard deviations of the mean.
    - At least 8/9 (or 89%) of all values lie within 3 standard deviations of the me

17. coefficient of variation

    The coefficient of variation (or CV) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean, and is given by the following

CV = s/x*100. Round the coefficient of variation to one decimal place (such as 25.3%).

18. The sample standard deviation s is a biased estimator of the population standard deviation s, which means that values of the sample standard deviation s do not tend to center around the value of the population standard deviation s. While individual
values of s could equal or exceed s, values of s generally tend to underestimate the value of s.

    The sample variance s2 is an unbiased estimator of the population variance s2, which means that values of s2 tend to center around the value of s2 instead of systematically tending to overestimate or underestimate s2.

19. **Z-score**

    A z score (or standard score or standardized value) is the number of standard deviations that a given value x is above or below the mean. The z score is calculated by using one of the following:
    z = (x-mean)/stddev

    Important points about Z-score
    a. A z score is the number of standard deviations that a given value x is above or
    below the mean.

    b. z scores are expressed as numbers with no units of measurement.

    c. A data value is significantly low if its z score is less than or equal to -2 or the value is significantly high if its z score is greater than or equal to +2.

    d. If an individual data value is less than the mean, its corresponding z score is a
    negative number.

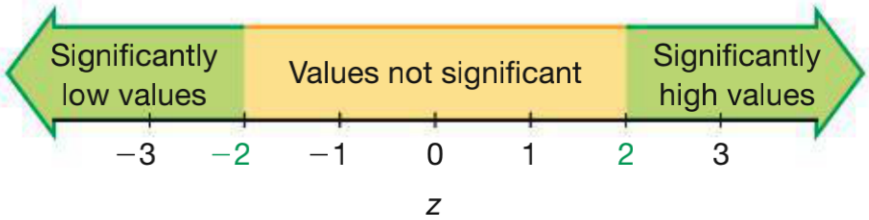


FIGURE 3-5 Interpreting *z* Scores

Significant values are those with z scores ≤ -2.00 or ≥ 2.00.

20. **Percentile**

    Percentile of value x = (no.of values less than x/total no of values)*100

21. **How to find data value from percentile**

From Figure 3-6, we see that the sample data are already sorted, so we can proceed to find the value of the locator L. In this computation we use $k = 25$ because we are trying to find the value of the 25th percentile. We use $n = 50$ because there are 50 data values.

$$L = \frac{k}{100} \cdot n = \frac{25}{100} \cdot 50 = 12.5$$

Since $L = 12.5$ is not a whole number, we proceed to the next lower box in Figure 3-6, where we change L by rounding it up from 12.5 to the next larger whole number: 13. (In this book we typically round off the usual way, but this is one of two cases where we round *up* instead of rounding *off*.) From the bottom box we see that the value of P_{25} is the 13th value, counting from the lowest. In Table 3-4, the 13th value is 7.9. That is, $P_{25} = 7.9$ Mbps. Roughly speaking, about 25% of the data speeds are less than 7.9 Mbps and 75% of them are more than 7.9 Mbps.

If L is 20 find 20th value and 21st value → add → divide by 2 to get data value

22. **Quartiles**

$$\text{Interquartile range (or IQR)} = Q_3 - Q_1$$

$$\text{Semi-interquartile range} = \frac{Q_3 - Q_1}{2}$$

$$\text{Midquartile} = \frac{Q_3 + Q_1}{2}$$

$$\text{10–90 percentile range} = P_{90} - P_{10}$$

23. 5 Numbers
    For a set of data, the 5-number summary consists of these five values:
    a. Minimum
    b. First quartile, Q1
    c. Second quartile, Q2 (same as the median)
    d. Third quartile, Q3
    e. Maximum
24. **Boxplot**

    A boxplot (or box-and-whisker diagram) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile Q1, the median, and the third quartile Q3.
25. Finding outliers from Boxplot

Identifying Outliers for Modified Boxplots

1. Find the quartiles Q_1, Q_2, and Q_3.
2. Find the interquartile range (IQR), where IQR $= Q_3 - Q_1$.
3. Evaluate $1.5 \times$ IQR.
4. **In a modified boxplot, a data value is an *outlier* if it is**
   above Q_3, by an amount greater than $1.5 \times$ IQR
   or below Q_1, by an amount greater than $1.5 \times$ IQR

26. Tool for Boxplot {https://www.desmos.com/calculator/h9icuu58wn}