Chapter 2 Methods for Describing Sets of Data

2.1 Describing qualitative data

Recall qualitative data: non-numerical or categorical data

Basic definitions:

A ____________ is one of the categories into which qualitative data can be classified.

The ______________ is the number of observations in the data set falling into a particular class.

The __________________ is the class frequency divided by the total number of observations in the data set. (in decimal)

Relative Frequency = $\frac{Frequency}{n}$

The __________________ is the class relative frequency multiplied by 100.

Class percentage = (class relative frequency) * 100

Graphical Descriptive Methods for Qualitative Data:

_______________________________: consists of two columns, one is the classes, the other one is the class frequency.

________________________________: consists of two columns, one is the classes, the other one is the class relative frequency or the class percentage.

_______________: The categories (classes) of the qualitative variable are represented by bars, where the height of each bar is either __________ or class relative frequency (or class percentage).

_______________: The categories (classes) of the qualitative variable are represented by slices of a pie (circle). The size of each slice is proportional to the _______________________.

________________: A bar graph with the categories (classes) of the qualitative variable arranged by height in ____________ order from left to right.

Example 1: Twenty-five army soldiers were given a blood test to determine their blood type.

Raw Data: A,B,B,AB,O   O,O,B,AB,B   B,B,O,A,O   A,O,O,O,AB   AB,A,O,B,A

Construct a frequency and relative frequency distribution table for the data.

Example 2: Road Rage: The following table provides the days on which 69 road rage incidents occurred. Use Descriptive Methods to describe this **qualitative data**.

F	Sa	W	M	Tu	F	Th	M
Tu	F	Tu	F	Su	W	Th	F
Th	W	Th	Sa	W	W	F	F
Tu	Su	Tu	Th	W	Sa	Tu	Th
F	W	F	F	Su	F	Th	Tu
F	Tu	Tu	Tu	Sa	W	W	Sa
F	Sa	Th	W	F	Th	F	M
F	M	F	Su	W	Th	M	Tu
Sa	Th	F	Su	W			

1. Summary table: Frequency and relative frequency distribution table.

Day of the week	frequency	Relative frequency
Sunday		
Monday		
Tuesday		
Wednesday		
Thursday		
Friday		
Saturday		

The following table is from SPSS output.

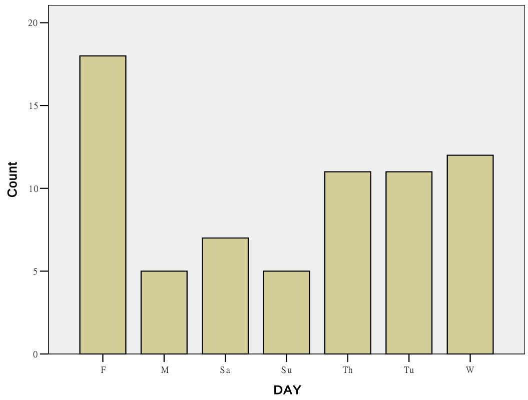
DAY

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	F	18	26.1	26.1	26.1
	M	5	7.2	7.2	33.3
	Sa	7	10.1	10.1	43.5
	Su	5	7.2	7.2	50.7
	Th	11	15.9	15.9	66.7
	Tu	11	15.9	15.9	82.6
	W	12	17.4	17.4	100.0
	Total	69	100.0	100.0	

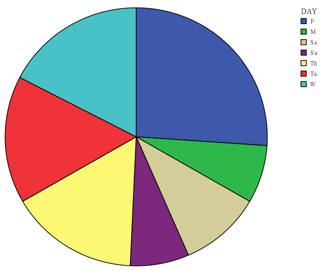
Questions:

1. Which class has the highest relative frequency? (_______________)
2. What is the percentage that road rage incidents occurred on Friday or Saturday?

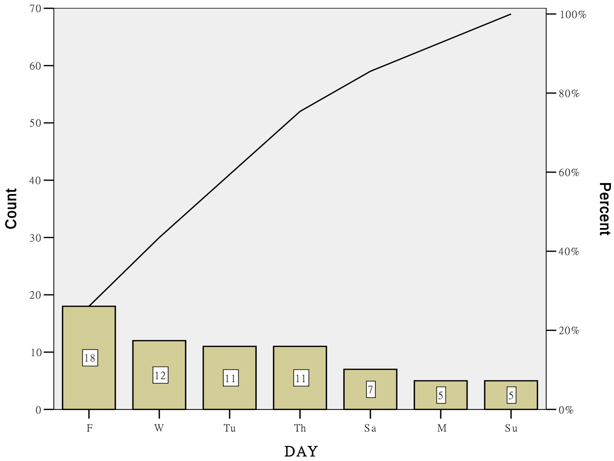
2. Bar graph



3. Pie chart



4. Pareto diagram



Example3. Age and Gender: The following bivariate data on age (in years) and gender were obtained from the 50 students in a freshman calculus course.

Age	Gender	Age	Gender	Age	Gender	Age	Gender	Age	Gender
21	M	29	F	22	M	23	F	21	F
20	M	20	M	23	M	44	M	28	F
42	F	18	F	19	F	19	M	21	F
21	M	21	M	21	M	21	F	21	F
19	F	26	M	21	F	19	M	24	F
21	F	24	F	21	F	25	M	24	F
19	F	19	M	20	F	21	M	24	F
19	M	25	M	20	F	19	M	23	M
23	M	19	F	20	F	18	F	20	F
20	F	23	M	22	F	18	F	19	M

1. Contingency table (summary frequency table)

	Under 21	21-25	Over 25	total
male				
female				
total				

2. Contingency table (summary relative frequency table)

	Under 21	21-25	Over 25	total
male				
female				
total				

Interpret the results.

1. How many female students are under 21 years old out of these 50 students?

2. What percentage of female students are under 21 years old?

3. How many female students are out of these 50 students and what is the percentage?

4. How many students are over 25 years old and what is the relative frequency?

2.2 Graphical methods for describing Quantitative data

____ can be used to describe quantitative data.

Example1. Test score:

score	frequency	Rel. freq.
>=90	8	
80-89	19	
70-79	18	
60-69	5	
<=60	0	

*Note: class can't be very small or very large.

Example2. DVD price: Describe the data by a frequency and relative frequency distribution table.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210

Price	frequency	Relative frequency
195-200		
200-205		
205-210		
210-215		
215-220		
220-225		

*Note: the borderline observation will classify into the next-highest interval. For example: $210 is classified into the 210-215 class.

Three graphical methods for describing quantitative data: **Dot plots, stem-and-leaf display, and histogram.**

________________: the horizontal axis is a scale for the quantitative variable. Each dot represents one observation of the data. (It shows how the data spread.)

Example2. DVD price: Describe the data by a dot plot.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210

The dot plot shows most of the prices fall between ____ and ____.

Questions:

1. How many observations in this data set?
2. What percentage of DVD players prices fall between $200 and $220?
3. define the variable x: the DVD players price, find

   $P(x < 200) =$

   $P(x \leq 200) =$

   $P(200 < x < 220) =$

_________________________: (Here each value is represented by a gross measurement called the _____ and a fine measurement called the ________)

1) Define the stem you wish to choose. (define it in such a way that the number of stems lies between __________________.)

2) List all possible digits of the stem in a column from the ___________________.

3) List the leaves next to the appropriate stem (generally, only ____ digit is displayed in the leaf.)

4) ________ the leaves.

Example2. DVD price: Describe the data by a stem-leaf display.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210

The following is a stem-leaf plot from MiniTab.

```
Stem-and-leaf of price  N  = 16
Leaf Unit = 1.0
   19  7999
   20
   20  89
   21  00224
   21  5999
   22  4
```

Questions:

1. How many observations in this data set?
2. What percentage of DVD players prices fall between $200 and $220?
3. define the variable x: the DVD players price, find

$P(x < 200) =$

$P(x \le 200) =$

$P(200 < x < 220) =$

*Note: 1. Stem-and-leaf display gives the ______________ of data set, and it also gives the actual values of observations;

2. Stem-leaf display is not appropriate for a ______________________ data set.

Example: The following data represents the breaking strengths of 20 linen threads: (in ounces)

32.5	15.2	29.3	24.5
21.2	20.0	23.9	33.0
27.3	41.0	36.8	19.2
20.6	26.9	28.7	34.2
25.4	34.6	33.2	37.0

Notes:

a) If there are too ______ leaves per stem, you can split the stem up, using ____ lines per stem.

b) If there are too many trailing digits, we can _______ some of these digits to maximize clarity.

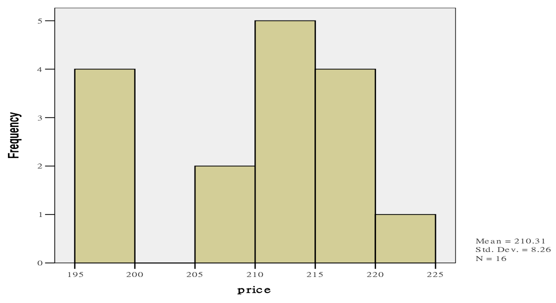
__________________:

________________ histogram: A graph that displays the classes on the horizontal axis and the frequencies of the classes on the vertical axis.

________________________ histogram: A graph that displays the classes on the horizontal axis and the relative frequencies of the classes on the vertical axis.

Example2. DVD price: Describe the data by a histogram.

$210, 219, 214, 197, 224, 219, 199, 199, 208, 209, 215, 199, 212, 212, 219, 210



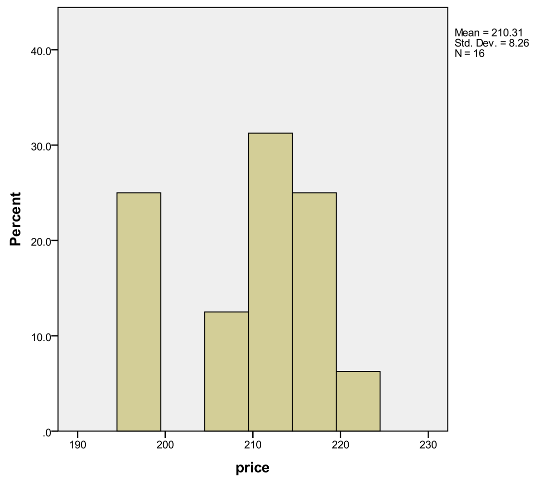
***Note: bar chart and histogram are different.**

1. How many observations are in this data set?

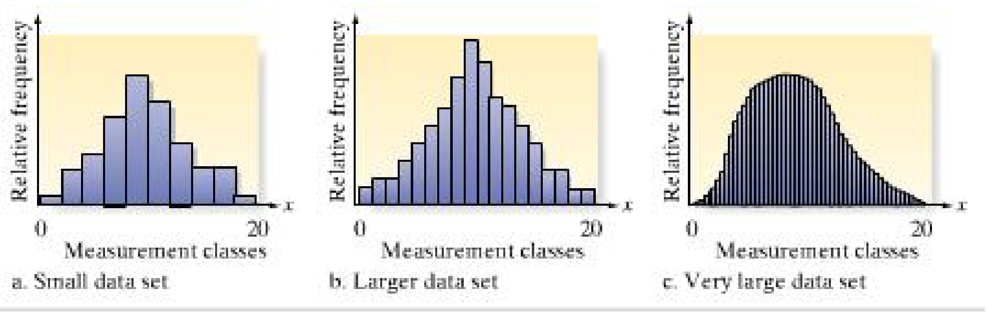
2. How many DVD player prices fall in $195 and $200, and what is the relative frequency?

3. Which interval with the highest relative frequency? How much is it?

4. What is the relative frequency of class interval ($205, $220)? Interpret it.



Relative frequency:



Note:
1. The total area under the curve is __________.
2. The proportion of the total ________ under the histogram that falls above a particular interval of the horizontal axis is equal to __________________________ of measurements falling in the interval.
3. For a very large data set, when the class intervals become small enough, a relative frequency histogram will appear as a __________________________.

2.3 Summation notation

Measurements:

Sum of measurement:

Sum of squares:

Square of sum:

Example: 5, 3, 8, 5, 4,

* Check the column sums in the following table to find the different summations.

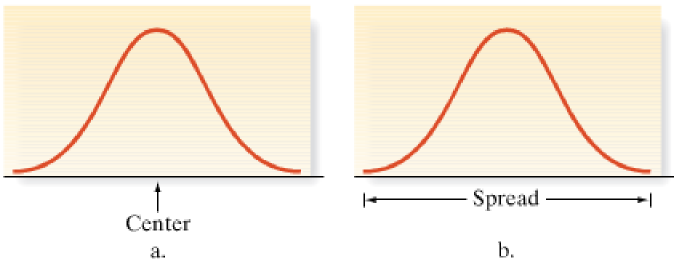
x	x^2	$x-5$	$(x-5)^2$
5			
3			
8			
5			
4			

2.4 Numerical measures of central tendency

Numerical descriptive methods measure two important data characteristics: Central tendency and variability.

The ______________________ of a data set: the tendency of the data to cluster, or center, about certain numerical values.

The ______________________ of a data set: the spread of the data.



Three most numerical measures of central tendency: Mean, Median, Mode

- ____________ of a data set is the arithmetic average of the data set. It measures the central tendency based on the values of observations.

For sample data, sample mean:

Find mean of the two sample data sets.

Example1: 5, 3, 8, 5, 2, 6, 9

Example2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Example3: Here is the survey result of prices of 10 DVD players. What is the average price of these 10 DVD players?

\$210, \$219, \$214, \$197, \$224, \$219, \$209, \$215, \$212, \$219

<u>Find a weighted mean.</u>

You are taking a class in which your grade is determined from five sources: 50% from your test mean, 15% from your midterm, 20% from your final exam, 10% from your computer lab work, and 5% from your homework. Your scores are 86 (test mean), 96 (midterm), 82 (final exam), 98 (computer lab), and 100 (homework). What is the weighted mean of your scores? If the minimum average for an A is 90, did you get an A?

Source	Score, x	Weight, w	x*w
Test mean	86	50%	
midterm	96	15%	
Final exam	82	20%	
Computer lab	98	10%	
homework	100	5%	

<u>Find the mean from a frequency distribution table.</u>

Find the mean number of credit cards the students have.

# of credit cards	students	
0	1	
1	5	
2	14	
3	13	
4	7	

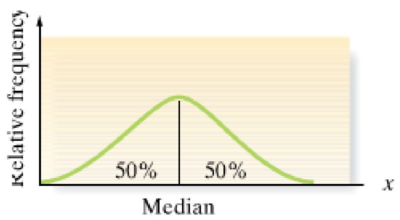
For population data, population mean:

Note: The sample mean ____will play an important role in accomplishing our objective of making inference about population based on sample information. We often use the sample mean ___to estimate the population mean_____.

Note: How accurate using $\bar{x}$ to estimate μ depends on:

1. the ______ of the sample. The larger the sample, the more accurate the estimate will tend to be.
2. the ________________, or spread of the data. If all other factors remaining constant, the smaller the variability, the more accurate the estimate.

- ____________ m **:** the middle number when the measurements are arranged in ascending (or descending) order. Median divides data set into two parts, 50% of the observations below the median and 50% of observations above the median. It is the ___________ of observations.



How to find a sample median m **:**

Arrange the n measurements from _____________ to ______________.

1. If n is odd, media m is the __________ number.

The position of the median is: (______)

2. If n is even, media m is the ___________________________ numbers.

The positions of the two middle numbers are: ____________

Find the median of the two data sets.
Example1: 5, 3, 8, 5, 2, 6, 9

Arrange in increasing order: _____________________________

The sample size ________(odd),

so the **position** of the median is_____________

the 4th measurement is the median: _______.

Example2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Arrange in increasing order: ______________________________________

The sample size ___________(even),

so the **position** of the median is the middle of _______________________,

The 5th measurement is 81 and the 6th is 84, so the median is ____________ .

Example3: Here is the survey result of prices of 10 DVD players. What is the median price of these 10 DVD players?

\$210, \$219, \$214, \$197, \$224, \$219, \$209, \$215, \$212, \$219

Note: In certain situations, when there are some __________observations in a data set, how the extreme observation has effect on the mean and median?

Example1: 5, 3, 8, 5, 2, 6, 9 (mean = 5.43 and median = 5)
Example2: 5, 3, 8, 5, 2, 6, 90 (mean = ___ and median = ___)

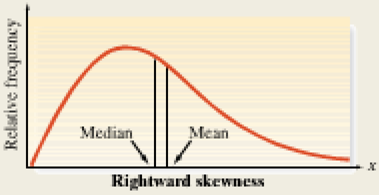
Example 3: (math test score) 89, 91, 73, 76, 69, 88, 79, 84, 85, 81(mean=81.5, median=82.5)
Example 4: (math test score) 7, 91, 73, 76, 69, 88, 79, 84, 85, 81 (mean=_____, median=____)

From above examples, you can find that mean is more ________ to extreme value than median. The _______may be a better measure of central tendency than the mean when there are some extreme observations in a data set.

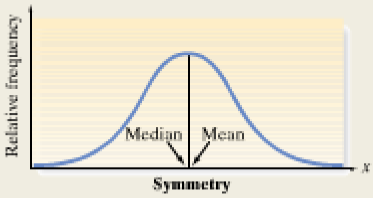
The relationship between mean and median:

Detecting Skewness by Comparing the Mean and the Median

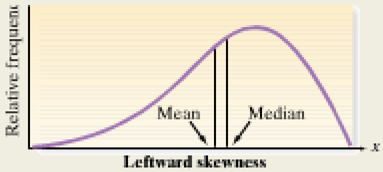
If the data set is skewed to the right, then the median is less than the mean.



If the data set is symmetric, then the mean equals the median.



If the data set is skewed to the left, then the mean is less than the median.



Example1: The mean test score for a class is 72 while the median is 80, what type of distribution most likely describes the shape of the test score?

(______________________)

Example2. The mean time for the patients stay in the hospital is 15 days while the median is 10 days, what type of distribution most likely describes the shape of the days staying in hospital?

(___________________________)

- **__________:** is the measurement that occurs most frequently in the data set.

Find the mode for the following data.

Example1: 5, 3, 8, 5, 2, 6, 9

The mode is the observation _________.

Example2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Example3: Here is the survey result of prices of 10 DVD players. What is the mode price of these 10 DVD players?

\$210, \$219, \$214, \$197, \$224, \$219, \$209, \$215, \$212, \$219

Example4. Road Rage:

Day of the week	frequency	Relative frequency
Sunday	5	0.0725
Monday	5	0.0725
Tuesday	11	0.159
Wednesday	12	0.174
Thursday	11	0.159
Friday	18	0.261
Saturday	7	0.101

______________ is the mode of the data. (since Friday with the highest frequency)

Note: 1. the mode is particularly useful for describing _________________ data. The modal category is simply the category (or class) that occurs most often.

2. Mean, median, or mode: The choice of which measure of central tendency to use will depend on the properties of the data set analyzed and on the application. So, it is vital that you understand how the mean, median and mode are computed.

2.5 Numerical measures of variability (range, Standard deviation)

- The ______________ is the largest measurement minus the smallest measurement.

$$Range = \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_$$

Find the range for the following data set:

Example1: 5, 3, 8, 5, 2, 6, 9,

Example2: Math test score: 89, 91, 73, 76, 69, 88, 79, 84, 85, 81

Example3: Here is the survey result of prices of 10 DVD players. What is the range of these 10 DVD players' prices?

$210, $219, $214, $197, $224, $219, $209, $215, $212, $219

Note: range is easy to calculate, and easy to interpret; but it is insensitive when the data set is large. It is only use the information of the smallest and the largest.

- The _________________________ for a sample of n measurements is equal to the sum of the squared distances from the mean divided by (n – 1).

Note: A shortcut formula for calculating s^2 is:

- The ________________________________ is the square root of the sample variance.

Example1: calculate mean, the sample variance and sample standard deviation for data 1, 2, 3, 4, 5.

Example2: calculate mean, the sample variance and sample standard deviation for data 2, 3, 3, 3, 4.

Example3: Here is the survey result of prices of 10 DVD players. What is the variance and standard deviation of these 10 DVD players' prices?

$210, $219, $214, $197, $224, $219, $209, $215, $212, $219

Note: 1. The standard deviation is expressed in the _____________ of measurement.
2. The ______the standard deviation, the _______________ the data.

- **The symbols for sample statistic and population parameter:**

Sample mean ____ population mean ____
Sample variance ____ population variance____
Sample standard deviation ____ population standard deviation ____

Note: 1. The __________________ can be calculated based on sample data, so it is known while the parameter usually is unknown.
2. _________________ can be used to estimate the corresponding _______________________.
3. The ________ the sample size, the better the estimation.

2.6 Interpreting the standard deviation

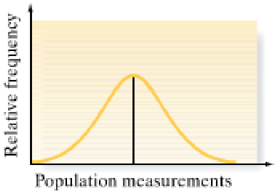
How does the standard deviation provide a measure of variability of a data set?
Interpreting the standard deviation: **Chebyshev's rule** and **Empirical rule**

- ____________________ applies to ___ data set, regardless of the shape of the frequency distribution of the data. (at least________of the measurements will fall within k standard deviation of the mean.)

a. At least ___________ of the measurements will fall within ____ standard deviations of the mean (within interval ($\overline{x}-2s, \overline{x}+2s$) for samples; ($\mu-2\sigma, \mu+2\sigma$) for populations).

b. At least ___________ of the measurements will fall within ____ standard deviations of the mean (within interval ($\overline{x}-3s, \overline{x}+3s$) for samples; ($\mu-3\sigma, \mu+3\sigma$) for populations).

- _____________________ is a rule of thumb that applies to data sets with frequency distributions that are mound shaped and _________.



a. Approximately ______ of the measurements will fall within ____ standard deviation of the mean (within interval ($\overline{x}-s, \overline{x}+s$) for samples; ($\mu-\sigma, \mu+\sigma$) for populations).

b. Approximately ______of the measurements will fall within ___ standard deviations of the mean (within interval ($\bar{x}-2s, \bar{x}+2s$) for samples; ($\mu-2\sigma, \mu+2\sigma$) for populations).

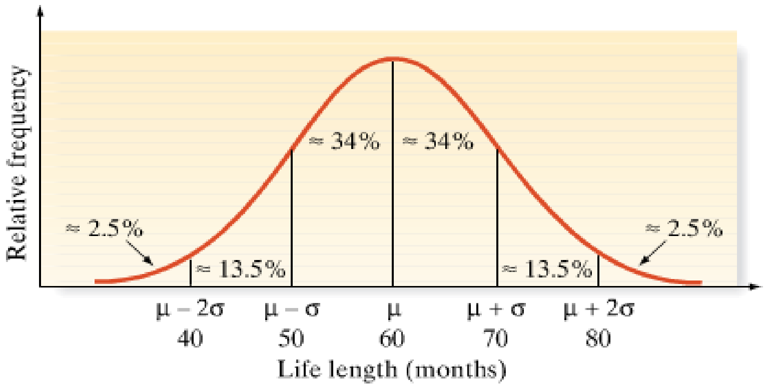
c. Approximately _____ of the measurements will fall within ____ standard deviations of the mean (within interval ($\bar{x}-3s, \bar{x}+3s$) for samples; ($\mu-3\sigma, \mu+3\sigma$) for populations).

Example1: A study was designed to investigate the effects of teaching method on a student's achievement in a mathematics course. These students obtained a mean score of 320 with a standard deviation of 50 on a standardized test. Assuming no information concerning the shape of the distribution is known, what percentage of the students scored between 220 and 420?

Example2: A study was designed to investigate the effects of teaching method on a student's achievement in a mathematics course. These students obtained a mean score of 320 with a standard deviation of 50 on a standardized test. Assuming a mound-shaped and symmetric distribution of the score, what percentage of the students scored between 170 and 470?

Example3: A manufacturer of automobile batteries claims that the average length of life time for its battery is 60 months and the standard deviation is 10 months, what is the minimum percentage of this brand battery that will last within a time interval of 30 months to 90 months?

Example4: A manufacturer of automobile batteries claims that the average length of life time for its battery is 60 months and the standard deviation is 10 months. If it turns out that the distribution of life times for this brand battery is normally distributed (or bell shaped),



1. What percentage of the batteries will last more than 50 months?

2. What percentage of the batteries will last less than 50 months?

3. What is the percentage of the battery that life time will last between 40 and 80 months?

4. What percentage of the batteries will last more than 50 months?

5. What percentage of the batteries will last more than 70 months?

6. What percentage of the batteries will last less than 40 months?

7. Suppose you buy one this brand battery. It lasts less than 40 months. What could you infer about the manufacturer's claim?

Note:
1.Both rules apply to either _____ data sets or _____________data set.

2. From these two rules, ______ (at least 75% or appro. 95%) of the measurements will within ___ standard deviations of mean, and ____________ (at least 89% or appro. 99.7%) of the measurements will fall within ____ standard deviations of the mean.

2.7 Numerical Measures of relative standing (percentile and z-score)

______________: descriptive measures of the relationship of a measurement to the rest of the data.

Percentile and z-score are two used to measure of relative standing.

- For any set of n measurements (arranged in __________ order), the ____________________ is a number such that p% of the measurements fall below that number and (100-p)% fall above it.

Q1: Ana's reading score is ranked as 99^{th} percentile in her school. What percentage of students has higher score than Ana?
Q2: This infant's weight is ranked as 60^{th} percentile. What percentage of infants is lighter than this infant?

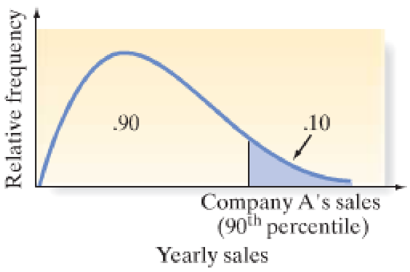


Figure 2.23
Location of 90th percentile for yearly sales of oil companies

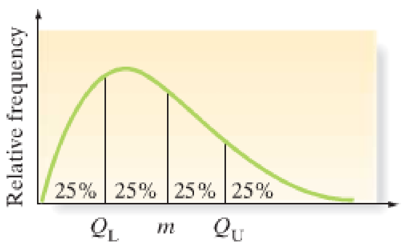


Figure 2.27
The quartiles for a data set

Three important Percentiles are:
___ percentile (______________): the median of the data set that lies at or below the median of the entire data set
____ percentile = 2nd Quartile = median
_____percentile (______________): median of the data set that lies at or above the median of the entire data set.

Example: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Median = 50th percentile = ____

___________ = 25th percentile = ____

___________ = 75th percentile = _____

- A ______________ represents the distance between a given measurement x and the mean, expressed in standard deviations.

Sample z-score for a measurement x is:

Population z-score for a measurement x is:

Example1: Find the z-score for the value 96, when the mean is 93 and the standard deviation is 3. (interpret it)

Example2: A manufacturer of automobile batteries claims that the average length of life time for its battery is 60 months and the standard deviation is 10 months. Calculate the z-scores for one this kind battery lasts 40 months and another one lasts 90 months.

Interpret: 40 months is____ standard deviations ______ mean 60 months.

Interpret: 90 months is____ standard deviations ________ mean 60 months.

Example3: A student took three Statistics exams last semester.

	score	Mean of class	standard deviation of class
Exam1	76	70	10
Exam2	83	75	9
Exam3	76	66	8

Compare to the whole class, which exam did he do best? (compare z-score)

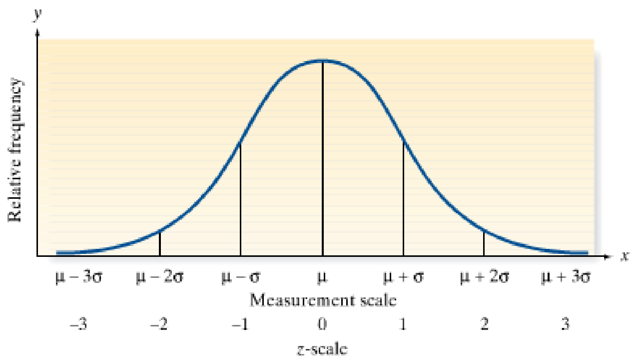
Note: 1. if $z > 0$, it means the observation is ___________ the mean.

2. if $z < 0$, it means the observation is ___________ the mean.

3. if $z = 0$, it means the observation ___________ the mean.

Interpretation of z-score for bell-shaped distributions of data:

Note that this interpretation of z-scores is identical to that given by the empirical rule for mound-shaped distributions (Table 2.7). The statement that a measurement falls into the interval from $(\mu - \sigma)$ to $(\mu + \sigma)$ is equivalent to the statement that a measurement has a population z-score between −1 and 1, since all measurements between $(\mu - \sigma)$ and $(\mu + \sigma)$ are within one standard deviation of μ. These z-scores are displayed in Figure 2.28.



1. Approximately _________ of the measurements will have a z-score between -1 and 1.

2. Approximately _________ of the measurements will have a z-score between -2 and 2.

3. Approx. _______ (almost all) of the measurements will have a z-score between -3 and 3.

4. If _________, it means the observation is highly suspected to be an outlier (rare observations).

Understand z-score:

Q1: Ana's reading score is 1.57 standard deviations above the school average.

Q2: This infant's weight is 3 standard deviations above the average.

Q3: This baby's height is 0.16 standard deviation below the average.

Q: The following table is the descriptive statistics for a class test score. (two intervals?)

Mean	80.32203
Median	84
Mode	92
Standard Deviation	10.84442
Sample Variance	117.6014
Range	48
Minimum	52
Maximum	100
Sum	4916
Count	59

Q: Calculate the z-scores for two students: one scored 98, the other one scored 62?

Learning Objective of Chapter 2:

1. Graphical methods to describe qualitative data: frequency and relative frequency distribution table, bar graph, pie chart and pareto diagram. (construct and interpret)
2. Graphical methods to describe quantitative data: frequency and relative frequency distribution table, dot plot, stem-leaf display and histogram. (construct and interpret)
3. Numerical measures to describe the central tendency of quantitative data: mean, median and mode (calculate and interpret; relationship between mean and median (skewness); if there is extreme value, median is less sensitive than mean; mode can describe the qualitative data.)
4. Numerical measures to describe the variability of quantitative data: range, sample variance and sample standard deviation (calculate and interpret, two rules—application restriction, percentage and corresponding intervals)
5. Numerical measures to describe the relative standing of quantitative data: percentile and z-score (percentile---understand and interpret; z-score-- calculate and interpret)