



Managing Data

- Organizing Data
- Types of Variables
- Frequency Distributions
- Properties of Frequency Distributions
- Methods of Summarizing Data
- Measures of Central Location
- Measures of Spread
- Choosing the Right Measure of Central Location and Spread

Christopher W. Blackwell, Ph.D., ARNP-C
Assistant Professor, College of Nursing
University of Central Florida

NGR 7642: Epidemiology Principles in Advanced Practice Nursing

Managing Data

- The first step in organizing data when investigating an outbreak or conducting a study is to create a line listing
- This is like a spreadsheet with each column containing a variable
- The first column is typically a person's name/initials/ID #
- The rest of the columns may contain demographic information, clinical details, and exposures r/t illness.

Table 2.1: *Line Listing of Hepatitis A Cases, County Health Department, Jan-Feb, 2004*

ID	Date of Dx	Town	Age	Sex	Hosp	Jaundice	Outbreak	IV	IgM +	Highest Alt
01	1/05	B	74	M	Y	N	N	N	Y	232
02	1/06	J	29	M	N	N	N	Y	Y	285
03	1/08	K	37	M	Y	N	N	N	Y	3250
04	1/19	J	3	F	N	N	N	N	Y	1100
05	1/30	C	39	M	N	N	N	N	Y	4146
06	2/02	D	23	M	Y	N	N	Y	Y	1271
07	2/03	F	19	M	Y	N	N	N	Y	300
08	2/05	I	44	M	N	N	N	N	Y	766
09	2/19	G	28	M	Y	N	N	Y	Y	23
10	2/22	E	29	F	N	Y	Y	N	Y	543



Managing Data

- Easily tracked data can be managed simply through hand-record keeping (small cluster of disease)
- Although most data management today is done via computers
- CDC has created a statistical management software package that is free: Epi Info 3
- Variables can be classified into one of four types:
 - Nominal Scale: Values are without numerical values (county of residence; alive/dead; well/ill; vaccinated/unvaccinated; did/did not have characteristic)
 - Ordinal Scale: Values can be ranked but not evenly spaced (cancer staging)
 - Interval Scale: Measured on a scale of equally spaced units but without a zero point (DOB)
 - Ratio-Scale: Interval variable with a true zero point (ht. in CMs or duration of illness)
- Nominal and Ordinal are qualitative or categorical whereas Interval and Ratio are considered quantitative or continuous variables.



Managing Data

Table 2.3 Example of Ordinal-Scale Variable: Stages of Breast Cancer*

Stage	Tumor Size	Lymph Node Involvement	Metastasis (Spread)
I	Less than 2 cm	No	No
II	Between 2 and 5 cm	No or in same side of breast	No
III	More than 5 cm	Yes, on same side of breast	No
IV	Not applicable	Not applicable	Yes

* This table describes the stages of breast cancer. Note that each stage is more extensive than the previous one and generally carries a less favorable prognosis, but you cannot say that the difference between Stages 1 and 3 is the same as the difference between Stages 2 and 4.



Managing Data

- Frequency Distributions:
 - Displays the values a variable can take and the number of persons or records with each value
 - Suppose you want to calculate the number of females with ovarian CA have had children:
 - List all values *parity* can take, from lowest-highest
 - For each value, record the number of births for each value
 - Continuous variables often summarized with measures of central location and measures of spread
 - Categorical variables usually summarized as ratios, proportions, and rates.



Managing Data

Table 2.4 Distribution of Case-Subjects by Parity (Ratio-Scale Variable), Ovarian Cancer Study, CDC

Parity	Number of Cases
0	45
1	25
2	43
3	32
4	22
5	8
6	2
7	0
8	1
9	0
10	1
Total	179

Data Sources: Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The reduction in risk of ovarian cancer associated with oral contraceptive use. *N Engl J Med* 1987;316: 650–5. Centers for Disease Control Cancer and Steroid Hormone Study. Oral contraceptive use and the risk of ovarian cancer. *JAMA* 1983;249:1596–9.



Managing Data

Table 2.5 Distribution of Cases by Stage of Disease (Ordinal-Scale Variable), Ovarian Cancer Study, CDC

Stage	CASES	
	Number	(Percent)
I	45	(20)
II	11	(5)
III	104	(58)
IV	30	(17)
Total	179	(100)

Data Sources: Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The reduction in risk of ovarian cancer associated with oral contraceptive use. *N Engl J Med* 1987;316: 650–5. Centers for Disease Control Cancer and Steroid Hormone Study. Oral contraceptive use and the risk of ovarian cancer. *JAMA* 1983;249:1596–9.



Managing Data

Table 2.6 Distribution of Cases by Enrollment Site (Nominal-Scale Variable), Ovarian Cancer Study, CDC

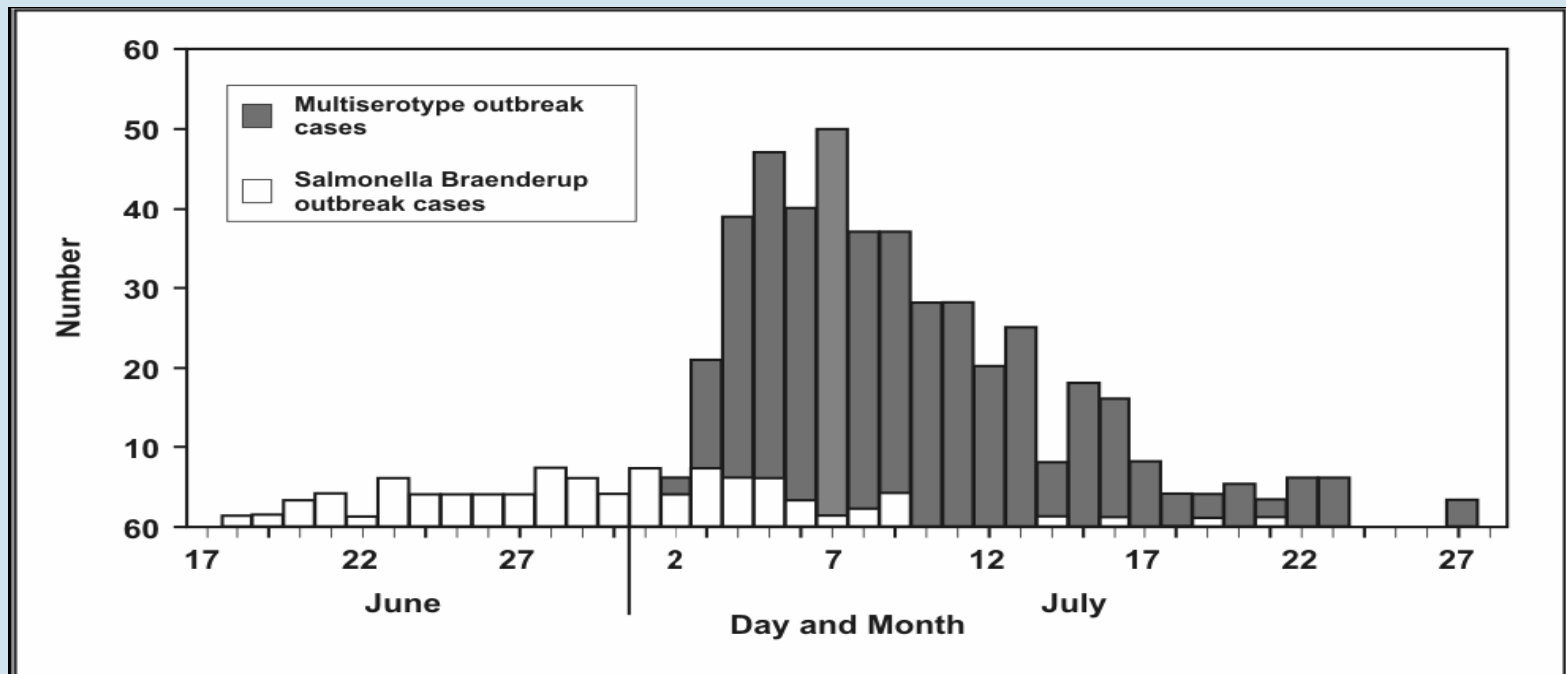
Enrollment Site	CASES	
	Number	(Percent)
Atlanta	18	(10)
Connecticut	39	(22)
Detroit	35	(20)
Iowa	30	(17)
New Mexico	7	(4)
San Francisco	33	(18)
Seattle	9	(5)
Utah	8	(4)
Total	179	(100)

Data Sources: Lee NC, Wingo PA, Gwinn ML, Rubin GL, Kendrick JS, Webster LA, Ory HW. The reduction in risk of ovarian cancer associated with oral contraceptive use. *N Engl J Med* 1987;316: 650–5. Centers for Disease Control Cancer and Steroid Hormone Study. Oral contraceptive use and the risk of ovarian cancer. *JAMA* 1983;249:1596–9.



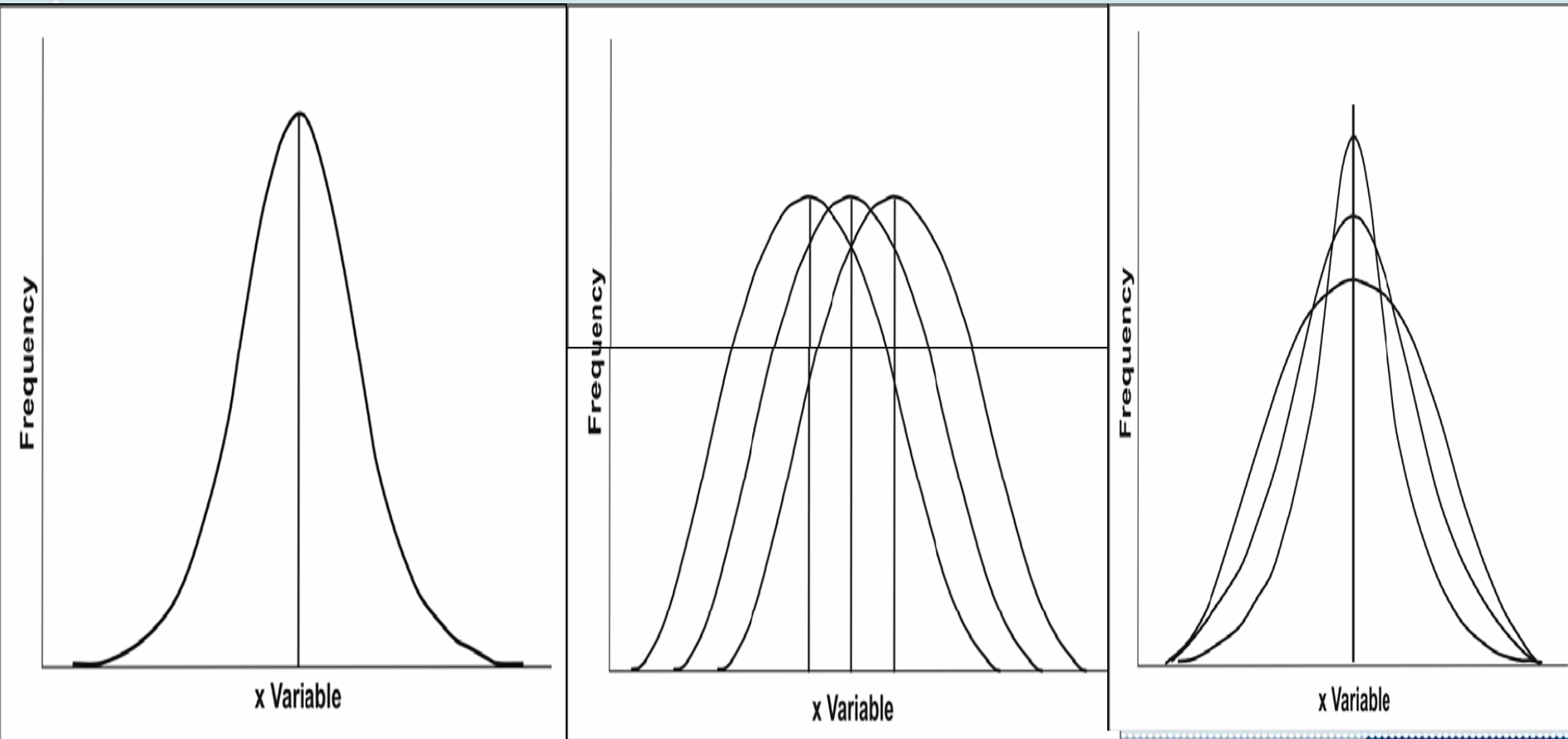
Managing Data

- Properties of Frequency Distributions:
 - Data in a frequency distribution can be graphed as a histogram:



Managing Data

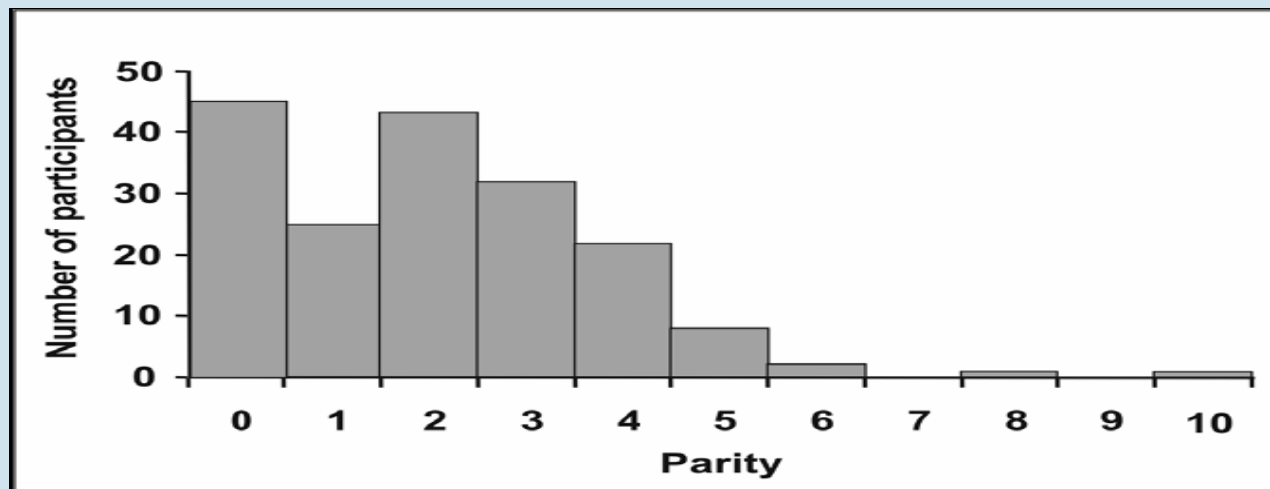
- Variation in Central Locations (Tendencies)



Managing Data

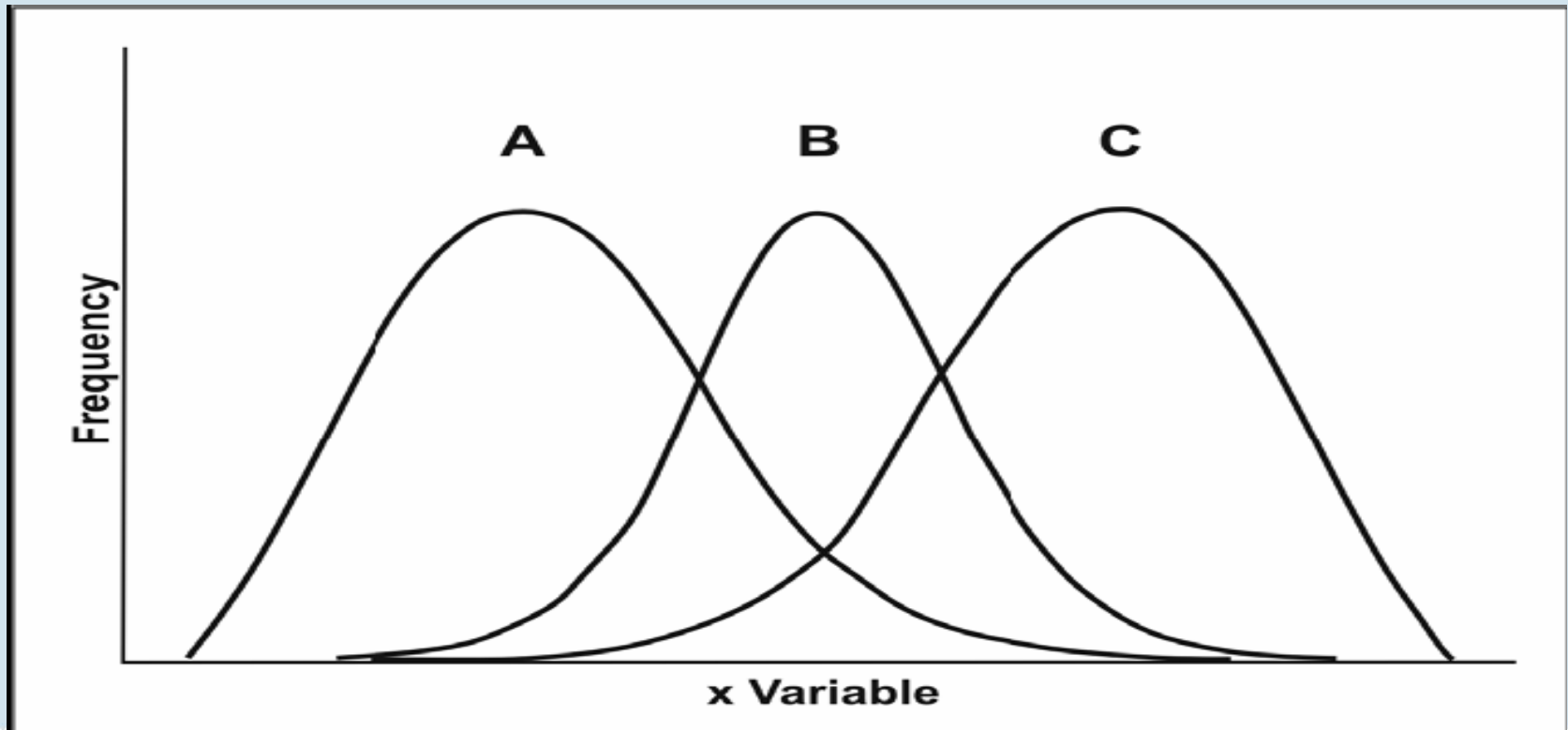
- 3 measures of central location often used in epidemiology:
 - Mean; Median; Mode
- Spread:
 - Distribution out from a central value (most commonly measured as Range and Standard Deviation)
- Shape:
 - Not all distributions are symmetrical (like the ones exemplified)
 - When the distribution to the right or left of the central location are different, the shape is Skewed (asymmetrical):

Figure 2.5 Distribution of Case-Subjects by Parity, Ovarian Cancer Study, CDC



Managing Data

- A distribution that has a central location to the left and tail to the right is positively skewed (A); distribution to the right and tail to the left is negatively skewed (C):



Managing Data

- Normal (Gaussian) Distribution: Classic symmetric, bell-shaped curve where the mean, median, and mode coincide at the central peak while area under the curve determines measures of spread

Table 2.7 Methods for Summarizing Different Types of Variables

	Ratio or Proportion	Measure of Central Location	Measure of Spread
Nominal	yes	no	no
Ordinal	yes	no	no
Interval	yes, but might need to group first	yes	yes
Ratio	yes, but might need to group first	yes	yes



Managing Data

- Measures of Central Location:
 - Provides a single value that summarizes an entire distribution of data
 - Selecting the best measure of central location depends on 2 factors:
 - Shape or skewness of the distribution
 - Intended use of the measure
 - Mode:
 - Value that occurs most often in a set of data; Easiest to calculate
 - Used to determine what is the most “popular” or most common value
 - Not usually affected by a small amount of outliers
 - To ID Mode:
 - Arrange observations into a frequency distribution identifying the variable and frequency the variable occurs
 - ID the value which occurs most often
 - » Note: When no value occurs more than once, there is NO MODE
 - » Note: When a value occurs at the same highest frequency as another, it is said to be bi-modal
 - Practice for yourself on page 2-16.



Managing Data

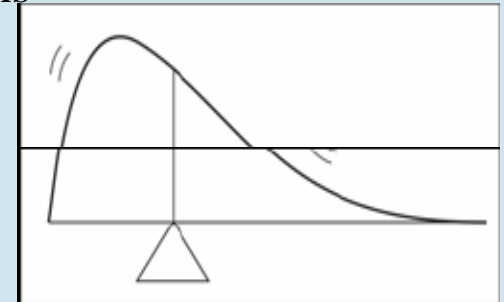
- Median:
 - Middle value of a set that has been put into rank order:
 - Good at locating the central point of skewed data
 - Not always a powerful stat so not used in many analyses
 - To Determine Median:
 - Arrange observations (n) into either ascending or descending order
 - Middle position = $(n + 1) / 2$
 - » If n is odd, the middle position is single and is the median
 - » If n is even, the middle position is between 2 observations and is the average of these 2 observations
 - Let's practice! Page 2-21:
 - A = 27 days
 - B = 28 days



Managing Data

- Arithmetic Mean:
 - Value that is closest to all other values in a distribution (the average)
 - Excellent statistical measure and can be manipulated and analyzed:
 - Great for determining the “center of gravity” if the data are normally distributed (subtracting the mean from *every* observation will sum to 0)
 - VERY weak statistical measure in skewed data with outliers
 - To determine Mean:
 - Sum all the observations of the set together
 - Divide this total by the number of observations

Mean: the center of gravity of the distribution



Managing Data

- Midrange:
 - Usually used as a precursor step in other calculations
 - Half-way point of a set of observations:
 - Identify the smallest and largest observations
 - Add them together and divide by 2
 - For age: You must add the youngest and oldest age and add 1, then divide this by 2
 - Example:
 - » Peds room 1 has 12 patients, all aged 2. What is the midrange age for these patients?
 - » You'd assume 2, right? WRONG:
 - » $2 + 2 + 1 = 5$; $5/2 = 2.5$
 - Let's practice! Page 2-28:
 - A = 23 days
 - B = 19.5 ETOH drinks
 - C = 20 years



Managing Data

- Geometric Mean:
 - Average of a set measured on a logarithmic scale
 - Dampens input of extreme values and is always smaller than arithmetic mean
 - Useful at calculating environmental sample data and dilution assays
 - To determine Geometric Mean (use a scientific calculator):
 - Input each data point in the data set by punching-in the value then log, then + and the next number, then log (repeat until all data points are inputted)
 - Divide the sum of the logs by the total number of data points
 - Take the square root of the product; This is your Geometric Mean
 - Let's practice!:
 - Page 2-32



Managing Data

Practice: Find the geometric mean of 10, 100 and 1000 using a scientific calculator:

Enter:	Calculator Displays:
10	10
LOG	1
+	1
100	100
LOG	2
+	3
1000	1000
LOG	3
=	6
/	6
3	3
=	2
10 ^x	100



Managing Data

- Selecting the appropriate measure:
 - Mode provides the most common value:
 - Good for describing values/variables
 - Median provides the central value:
 - Measure of choice when data are not normally distributed (skewed), often preferred in epidemiology
 - Arithmetic mean provides the average value:
 - Used most often in statistical manipulation
 - Uses all data so sensitive to outliers
 - Midrange provides the midpoint value:
 - Most sensitive to outliers; not very useful
 - Geometric mean provides the logarithmic mean:
 - Used most commonly with lab data or environmental sampling
 - Choose the right measurement based on data characteristics (skewed/normal, outliers, etc.) and reason for calculation (descriptive vs. analytic purpose)



Managing Data

- Measures of Spread:
 - Describe the spread (dispersion) of values from the peak
 - Range:
 - Difference between a data set's largest and smallest value (reported as two numbers, from "X" to "X")
 - Example:
 - » 27, 31, 15, 30.22
 - » Range = 31 – 15 = 16
 - Percentile:
 - Divide the data into 100 equal parts; Pth percentile is the percentage of observations that fall at or below that percentage:
 - 90th percentile = 90% of the observations fall at or below that value
 - Quartiles:
 - Similar to the percentile; but divides data into 25% quadrants:
 - 25% → 50% → 75% → 100%



Managing Data

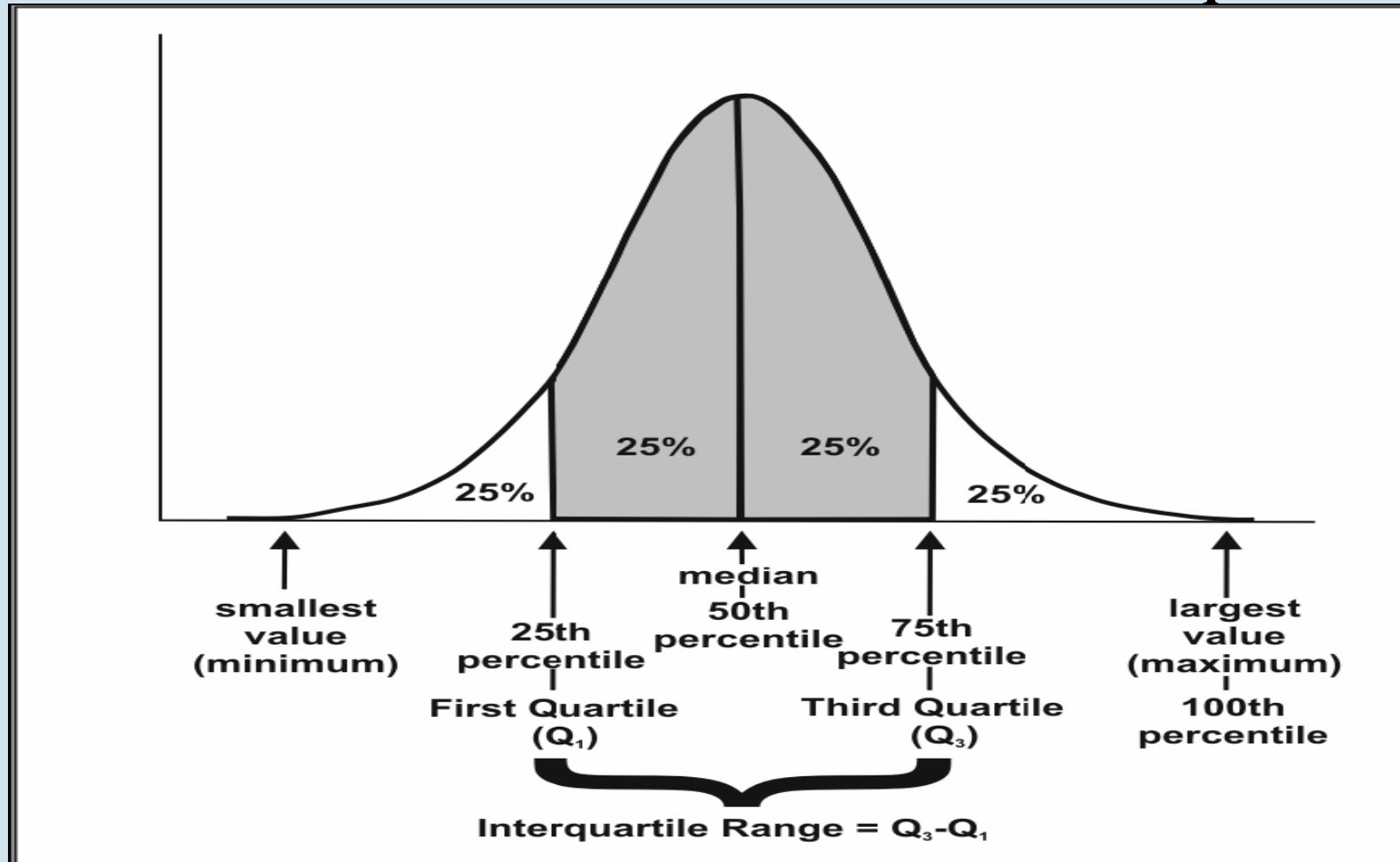
– Interquartile Range:

- Central part of the distribution: between 25%-75%
- Useful for describing the central location and spread of frequency distribution particularly in skewed data
 - Arrange observations in increasing order
 - Find the positions of the 1st and 3rd quartiles (where $n = \# \text{ obs.}$):
 - » $(n + 1) / 4$
 - » $3(n + 1) / 4 = 3 \times Q_1$
 - ID 1st and 3rd quartiles: (if on a whole number, the value of the quartile is the value of that observation; if it is between observations, it is the lower observation plus the specified fraction of the difference between the observations)
 - Report the values at Q_1 and Q_3 ; calculate the interquartile range as $Q_3 - Q_1$
- Let's practice! Page 2-38



Managing Data

Figure 2.7 The Middle Half of the Observations in a Frequency Distribution Lie within the Interquartile Range



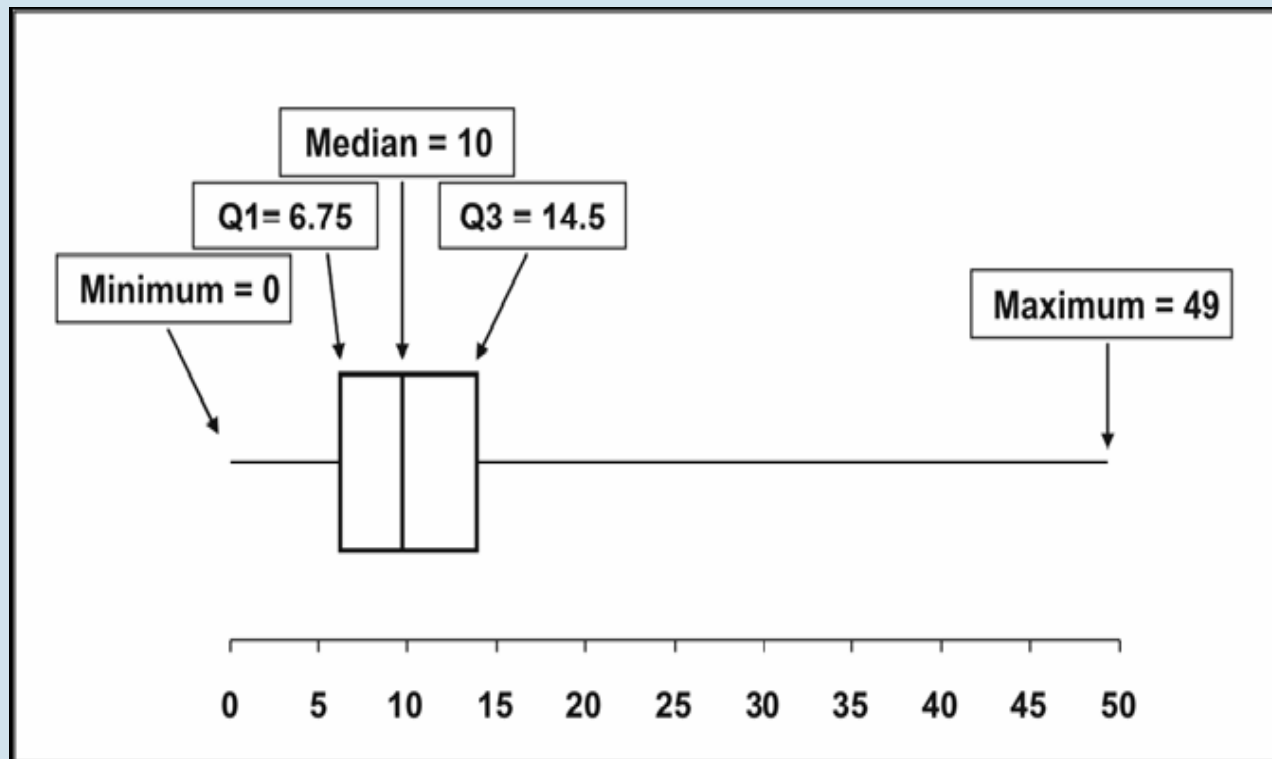
Managing Data

- Practice: Interquartile Range (page 2-38):
 - Arrange observations in increasing order:
 - 0,2,3,4,5,5,6,7,8,9,9,9,10,10,10,10,10,11,12,12,12,13,14,16,18,18,19,22,27,49
 - Find positions of 1st and 3rd quartiles:
 - $Q_1 = (n + 1) / 4 = (30 + 1) / 4 = 7.75$
 - $Q_3 = 3(n + 1) / 4 = 3(30 + 1) / 4 = 23.25$
 - ID value of 1st and 3rd quartiles:
 - $Q_1 = 7.75$ (So, we must take the value of the 7th obs. and add 75% of the value of between 7th and 8th observations to this value:
 - 6 (7th observation is 6) + $.75$ (3/4 of the value between the 7th and 8th observations, which is 1 [7-6]) = **6.75**
 - $Q_3 = 23.25$ (So, we must take the value of the 23rd obs. and add 25% of the value between obs. 23 and 24:
 - 14 (23rd obs. is 14) + $.50$ (25% of the value between the 23rd and 24th obs, which is 2 [16-14]) = **14.5**
 - Subtract Q_3 minus Q_1 :
 - $14.5 - 6.75 = 7.75$



Managing Data

Figure 2.8 Interquartile Range from Cumulative Frequencies



Managing Data

Table 2.10 Frequency Distribution of Length of Hospital Stay, Sample Data, Northeast Consortium Vancomycin Quality Improvement Project

Length of Stay (Days)	Frequency	Percent	Cumulative Percent
0	1	3.3	3.3
2	1	3.3	6.7
3	1	3.3	10.0
4	1	3.3	13.3
5	2	6.7	20.0
6	1	3.3	23.3
7	1	3.3	26.7*
8	1	3.3	30.0
9	3	10.0	40.0
10	5	16.7	56.7*
11	1	3.3	60.0
12	3	10.0	70.0
13	1	3.3	73.3
14	1	3.3	76.7*
16	1	3.3	80.0
18	2	6.7	86.7
19	1	3.3	90.0
22	1	3.3	93.3
27	1	3.3	96.7
49	1	3.3	100.0
Total	30		100.0 (* = closest to 25%, 50%, and 75%)



Managing Data

- Standard Deviation:
 - Spread most commonly used with the arithmetic mean; roughly, it's the difference between each observation and the mean
 - Useful when data are typical and bell-shaped
 - To calculate:
 - Calculate the arithmetic mean
 - Subtract the mean from each observation; square each difference
 - Sum the squared differences
 - Divide the sum of the spread differenced by $n - 1$ (where n is the # of obs.)
 - Take the square root of this value; this is the SD
 - Let's practice! Page 2-44



Managing Data

Figure 2.9 Area Under Normal Curve within 1, 2 and 3 Standard Deviations

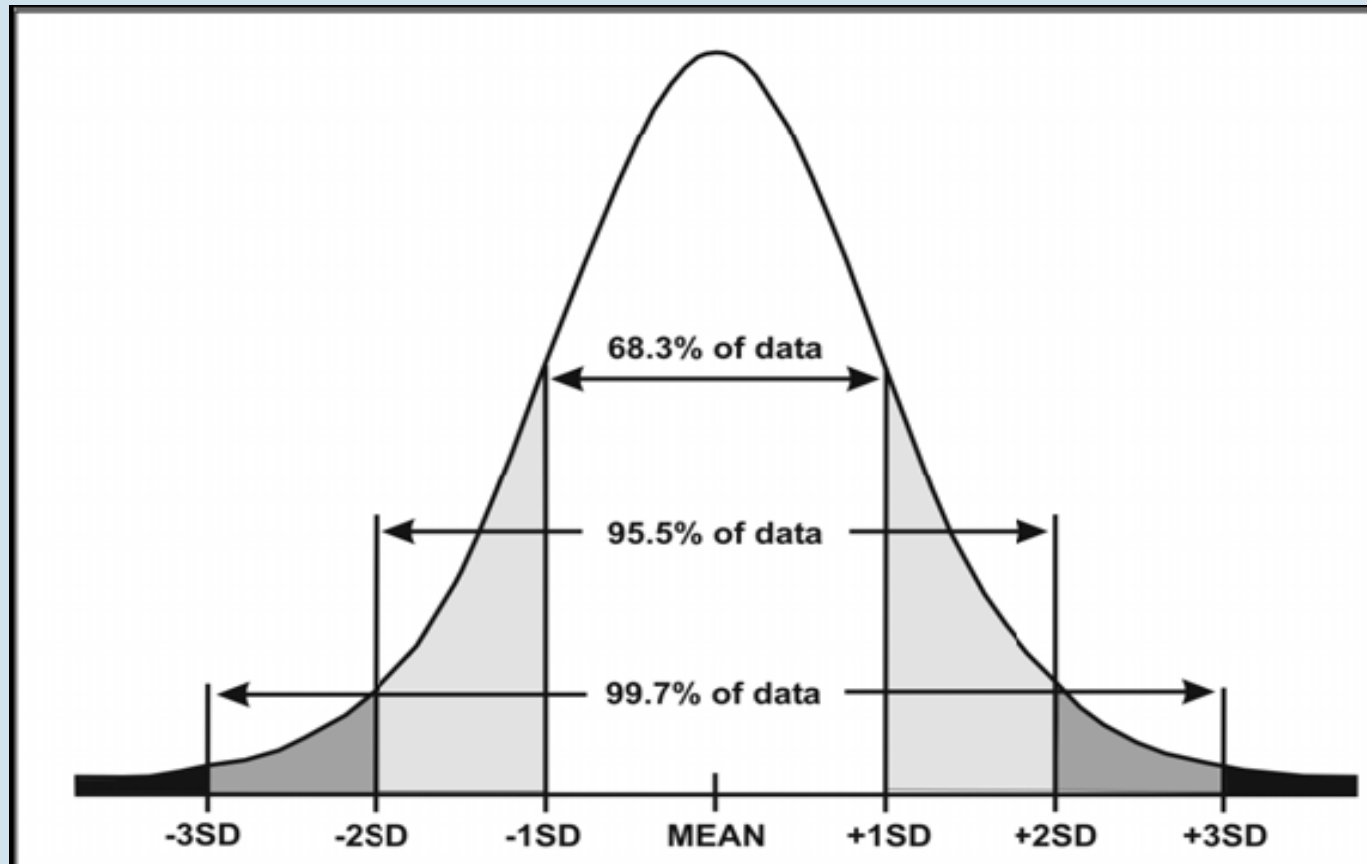
Areas included in normal distribution:

+1 SD includes 68.3%

+1.96 SD includes 95.0%

+2 SD includes 95.5%

+3 SD includes 99.7%



Managing Data

- Standard Error of the Mean:
 - Describes variability around the mean we would see if repeated samples were taken from the same population
 - Used in calculating confidence intervals around the arithmetic mean
 - To calculate:
 - Calculate the SD
 - Divide the SD by the square root of the # of obs. (n)
 - Let's practice:
 - Calculate the Standard Error of the Mean from the previous example



Managing Data

- Confidence Intervals (CI):
 - Used to gauge the precision of a measurement:
 - Narrow CI indicates high precision; a wider CI indicates low precision
 - Useful when drawing conclusions from samples to the population at large
 - The CI is a guide and is not used *too* strictly
 - Clinical example:
 - If the mean cholesterol for adult females is 206, with a standard error of the mean of 3, 1 SD below the mean is 203 and one SD above the mean is 209. Therefore, we can say with 68.3% certainty that these levels would be consistent if we sampled different adult females
 - 68.3% is simply *not* accurate enough, we want 95%



Managing Data

- Confidence Intervals:
 - To calculate CI:
 - Calculate Mean and Standard error (SE) of the mean
 - Multiple the standard error by 1.96
 - Lower limit of the 95% CI = Mean – 1.96 x SE
 - Upper limit of the 95% CI = Mean + 1.96 x SE
 - Let's practice: Page 2-49



Managing Data

- Choosing the Right Measure of Central Location and Spread:

Table 2.11 Recommended Measures of Central Location and Spread by Type of Data

Type of Distribution	Measure of Central Location	Measure of Spread
Normal	Arithmetic mean	Standard deviation
Asymmetrical or skewed	Median	Range or interquartile range
Exponential or logarithmic	Geometric mean	Geometric standard deviation



Managing Data

- Mean is most common used and manipulated statistic
- If the median is greater than the mean, the data is skewed to the right; if the median is less than the mean, the data is skewed to the left
- In skewed data sets, use the median. Great for:
 - Incubation periods, duration of illness, age of study subjects
- Interquartile range and range can be used to show dispersion around the median
- Mode is the least useful measure of central location; used when trying to determine most common value
- Geometric mean used for exponential data, such as lab titers or environmental sampling data



Managing Data

- Summary:
 - Frequency distributions, measures of central location, and measures of spread are effective tools for summarizing numerical variables including:
 - Physical characteristics such as height and diastolic blood pressure,
 - Illness characteristics such as incubation period, and
 - Behavioral characteristics such as number of lifetime sexual partners.
 - Some characteristics, such as IQ, follow a normal or symmetrical bell-shaped distribution in the population. Other characteristics have distributions that are skewed to the right (tail toward higher values) or skewed to the left (tail toward lower values).
 - Some characteristics are mostly normally distributed, but have a few extreme values or outliers.
 - Some characteristics, particularly laboratory dilution assays, follow a logarithmic pattern.



Managing Data

- Finally, other characteristics follow other patterns (such as a uniform distribution) or appear to follow no apparent pattern at all.
- The distribution of the data is the most important factor in selecting an appropriate measure of central location and spread.
- Measures of central location are single values that represent the center of the observed distribution of values.
- The different measures of central location represent the center in different ways:
 - The arithmetic mean represents the balance point for all the data.
 - The median represents the middle of the data, with half the observed values below the median and half the observed values above it.
 - The mode represents the peak or most prevalent value. The geometric mean is comparable to the arithmetic mean on a logarithmic scale.
- Measures of spread describe the spread or variability of the observed distribution. The range measures the spread from the smallest to the largest value.
- The standard deviation, usually used in conjunction with the arithmetic mean, reflects how closely clustered the observed values are to the mean. For normally distributed data, 95% of the data fall in the range from -1.96 standard deviations to $+1.96$ standard deviations.



Managing Data

- The interquartile range, used in conjunction with the median, includes data in the range from the 25th percentile to the 75th percentile, or approximately the middle 50% of the data.
- Data that are normally distributed are usually summarized with the arithmetic mean and standard deviation.
- Data that are skewed or have a few extreme values are usually summarized with the median and range, or with the median and interquartile range.
- Data that follow a logarithmic scale and data that span several orders of magnitude are usually summarized with the geometric mean.

