Introductory Statistics

2024 Lectures
Part 3 - Descriptive Statistics

Institute of Economic Studies
Faculty of Social Sciences
Charles University in Prague

Categorical data

Example 3: Sample of 50 soft drink purchases

Coke Classic	Sprite	Pepsi
Diet Coke	Coke Classic	Coke Classic
Pepsi	Diet Coke	Coke Classic
Diet Coke	Coke Classic	Coke Classic
Coke Classic	Diet Coke	Pepsi
Coke Classic	Coke Classic	Dr. Pepper
Dr. Pepper	Sprite	Coke Classic
Diet Coke	Pepsi	Diet Coke
Pepsi	Coke Classic	Pepsi
Pepsi	Coke Classic	Pepsi
Coke Classic	Coke Classic	Pepsi
Dr. Pepper	Pepsi	Pepsi
Sprite	Coke Classic	Coke Classic
Coke Classic	Sprite	Dr. Pepper
Diet Coke	Dr. Pepper	Pepsi
Coke Classic	Pepsi	Sprite
Coke Classic	Diet Coke	

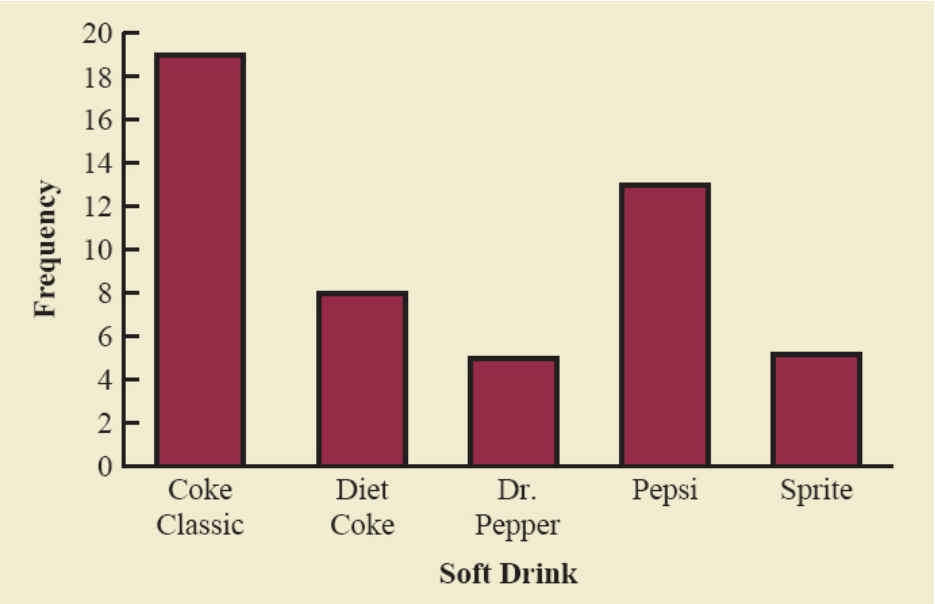
Categorical data

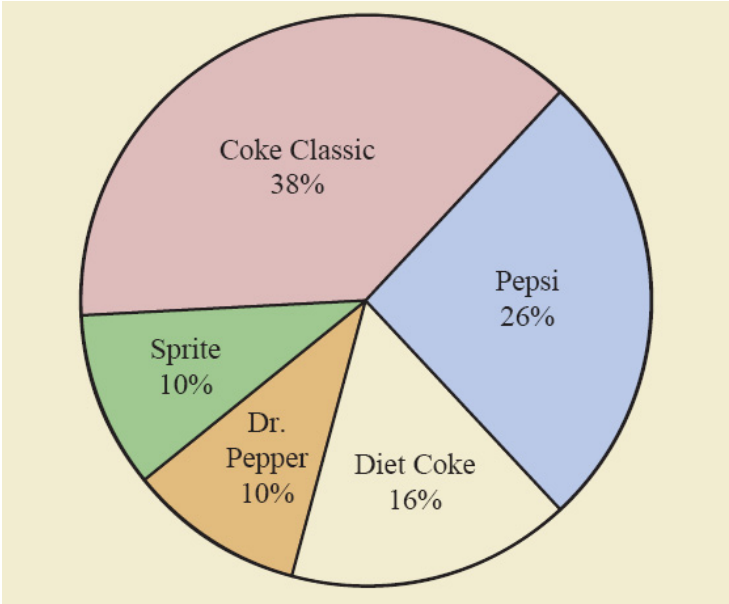
- frequency distribution is a tabular summary of data showing the number of each items in each of several non-overlapping classes
- relative frequency of a class equals the proportion of items belonging to a class; percent frequency of a class is the relative frequency expressed in percents (multiplied by 100)
- relative (percent) frequency distribution gives a tabular summary of relative (percent) frequency for each class

Soft Drink	Relative Frequency	Percent Frequency
Coke Classic	.38	38
Diet Coke	.16	16
Dr. Pepper	.10	10
Pepsi	.26	26
Sprite	.10	10
Total	1.00	100

Categorical data

- bar chart is a graphical representation of categorical data summarized in a frequency, relative frequency or percent frequency distribution; the bars should be separated to emphasize that each class is separate.
- pie chart provides another such graphical device





Quantitative data

Example 4: Year-end audit times in days

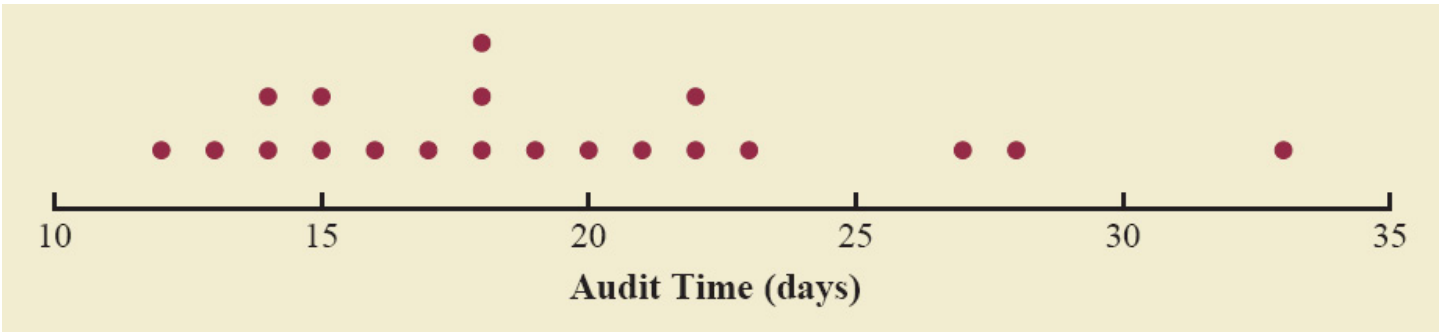
YEAR-END AUDIT TIMES (IN DAYS)			
12	14	19	18
15	15	18	17
20	27	22	23
22	21	33	28
14	18	16	13

- frequency distributions as for categorical data; one has to define the non-overlapping classes
- number of classes - depending on size of the data set; use enough classes to show the variation in the data; general guideline to use 5-20 classes
- width of the classes - usually same for each class; approximate width as range of data divided by number of classes rounded to a convenient value
- class limits chosen such that each data item belongs to a unique class

Quantitative data

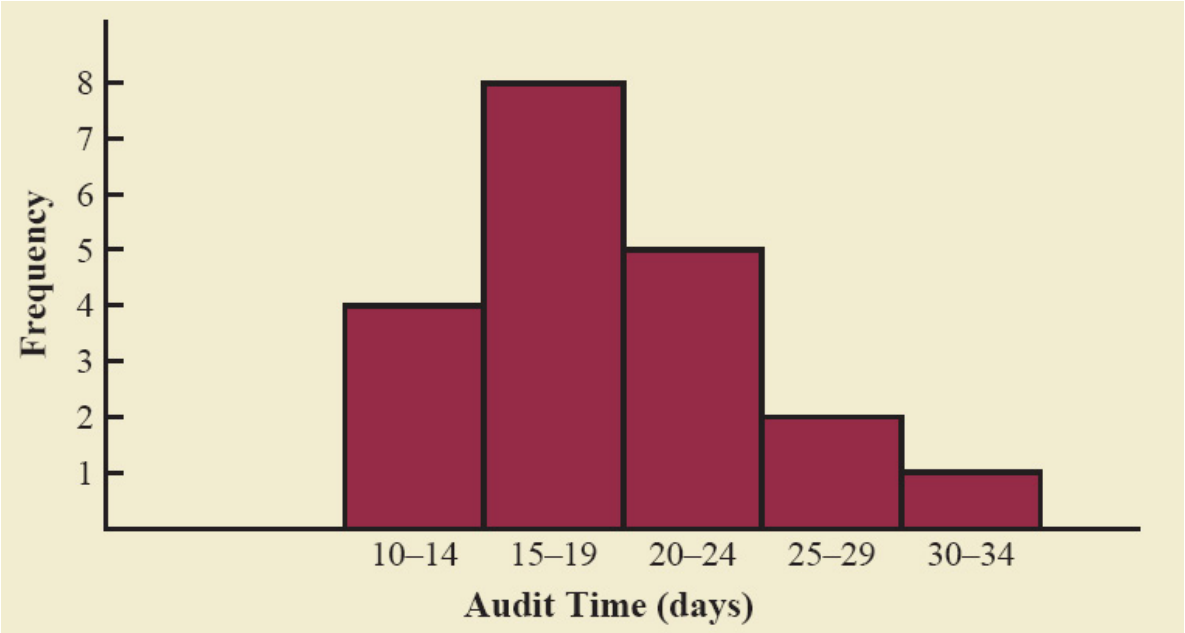
Audit Time (days)	Relative Frequency
10–14	.20
15–19	.40
20–24	.25
25–29	.10
30–34	.05
Total	1.00

- class midpoint useful in some applications; value halfway between the lower and upper class limits
- in our example the five midpoints are 12,17,22,27 and 32
- dot plot is the simplest graphical summary



Quantitative data

- histogram is a common graphical presentation of quantitative data summarized in some frequency distribution; spaces between classes are eliminated to show that all values in the range of data are possible



- histogram provides information about shape of a distribution (audit data are moderately skewed right, similarly data on housing prices, salaries etc.)

Quantitative data

- bar chart and histogram are essentially the same thing; a histogram is a bar chart with no separation between bars. For discrete data, separation is also possible. With continuous data a separation between bars is not appropriate.
- to determine the class limits, take into account the units of data. If audit data were rounded to a tenth of a day, the appropriate first class would be 10.0 to 14.9 days.
- one can consider also open-ended class. In case of audit data, the last class can be considered “30 or more” instead of 30-34. Similarly the first class can be “14 or less” instead of 10-14.

Stem-and-leaf display

Example 5: Number of correct answers on an aptitude test

112	72	69	97	107
73	92	76	86	73
126	128	118	127	124
82	104	132	134	83
92	108	96	100	92
115	76	91	102	81
95	141	81	80	106
84	119	113	98	75
68	98	115	106	95
100	85	94	106	119

- stem-and-leaf display is a technique of exploratory data analysis to show both shape and rank order of a data set

Stem-and-leaf display

6	9 8
7	2 3 6 3 6 5
8	6 2 3 1 1 0 4 5
9	7 2 2 6 2 1 5 8 8 5 4
10	7 4 8 0 2 6 6 0 6
11	2 8 5 9 3 5 9
12	6 8 7 4
13	2 4
14	1

- first arrange to the left of a vertical line the leading digits of each data (stems) and to the right of the vertical line record the last digit (leaves); here leaf unit is 1

Stem-and-leaf display

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

- next we sort the digits on each line into rank order
- Question: have you seen table of this sort in Prague?

Stem-and-leaf display

6	8 9
7	2 3 3 5 6 6
8	0 1 1 2 3 4 5 6
9	1 2 2 2 4 5 5 6 7 8 8
10	0 0 2 4 6 6 6 7 8
11	2 3 5 5 8 9 9
12	4 6 7 8
13	2 4
14	1

- use a rectangle to contain the leaves of each stem
- rotating provides similar graphical representation to histogram

Crosstabulation

Example 6: Analysis of verdicts for judges Luckett and Kendall

	Judge		
Verdict	Luckett	Kendall	Total
Upheld	129 (86%)	110 (88%)	239
Reversed	21 (14%)	15 (12%)	36
Total (%)	150 (100%)	125 (100%)	275

- crosstabulation or contingency table is a tabular summary of data for two variables. Left and top margin labels define the classes for the two variables. Each observation is associated with a cell.
- useful for inspecting the strength or symmetry of the relationship between the two variables

Simpson's paradox

- the data in two or more crosstabulations are often combined or aggregated. Conclusions based on unaggregated data can be reversed - Simpson's paradox

Judge Luckett

Verdict	Common Pleas	Municipal Court	Total
Upheld	29 (91%)	100 (85%)	129
Reversed	3 (9%)	18 (15%)	21
Total (%)	32 (100%)	118 (100%)	150

Judge Kendall

Verdict	Common Pleas	Municipal Court	Total
Upheld	90 (90%)	20 (80%)	110
Reversed	10 (10%)	5 (20%)	15
Total (%)	100 (100%)	25 (100%)	125

- inspect a presence of a hidden variable

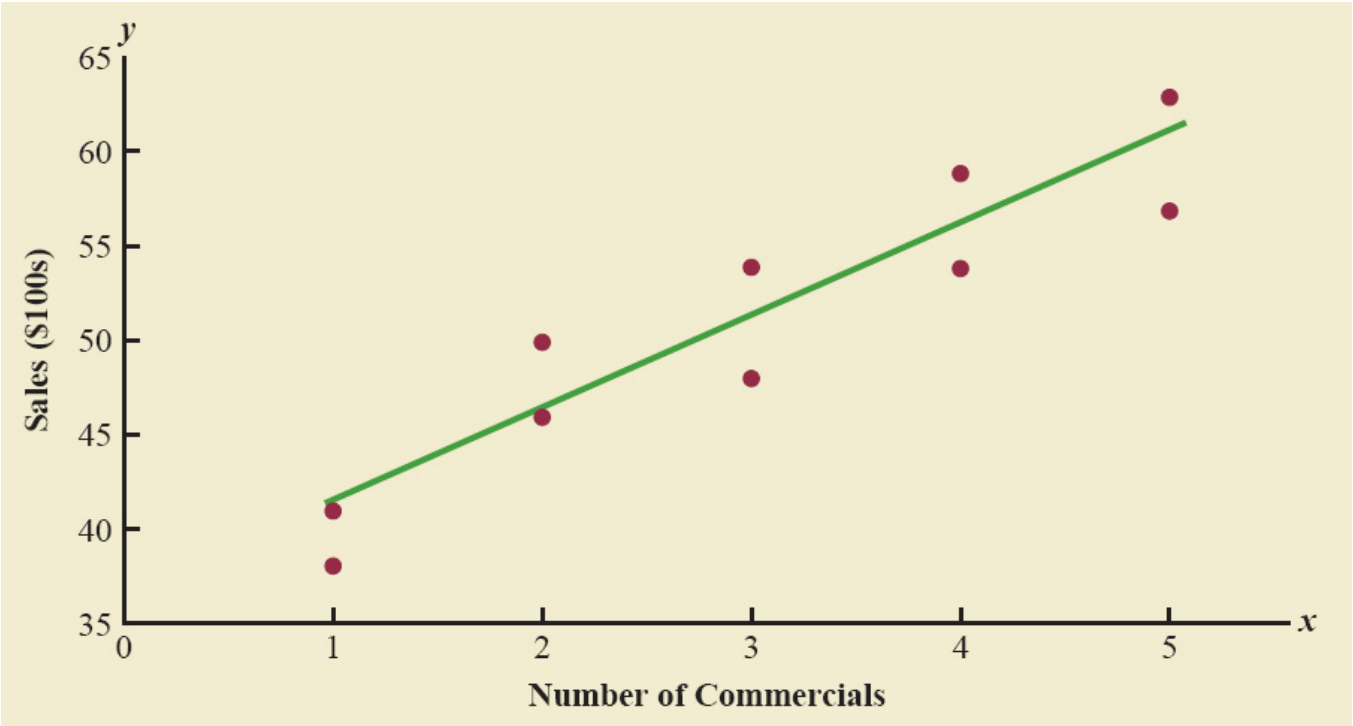
Scatter diagram

Example 7: Sample data for the Stereo and sound system store

Week	Number of Commercials x	Sales ($100s) y
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	46

Scatter diagram

- scatter diagram or scatter plot is a graphical representation of the relationship between two quantitative variables
- trendline or regression line is the line that provides an approximation of the relationship



Sample statistics

- sample statistics are numerical summary measures computed for data from a sample
- numerical measures of location, dispersion, shape and association provide additional alternatives to tabular and graphical representations

Example 8: Starting salaries of 12 graduates

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3450	7	3490
2	3550	8	3730
3	3650	9	3540
4	3480	10	3925
5	3355	11	3520
6	3310	12	3480

Measures of location

- We can consider measures of central location (mean, median, mode) or other locations (percentiles)
- Sample mean or average value provides a measure of central location for the data
- Denoting values of observations $x_i, i = 1, \ldots, n$, the sample mean is given as

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- The mean monthly starting salary for the sample of 12 graduates is 3540.
- In some applications each observation has its own weight reflecting its importance and we compute weighted sample mean

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- E.g. computation of student's grade point average.

Measures of location

- (sample) median is another measure of central location
- arrange data in the ascending order from smallest to largest value; for an odd number of observations, the median is the middle value; for an even number of observations, the median is the average of the two middle values.

3310 3355 3450 3480 3480 3490 3520 3540 3550 3650 3730 3925

Middle Two Values

- The median value of monthly salary for the sample of 12 graduates is 3505.
- The mean can be influenced by extremely small or large data; e.g. consider replacing the largest value \$3925 by \$10,000, then the mean changes to \$4046. However, the median is unchanged (robust).

Measures of location

- Mode is the value that occurs with greatest frequency
- The only monthly salary that occurs more than once is \$3480 which is thus the mode.
- Sample quantile or percentile provides information about how the data are spread over the interval from the smallest value to the largest value
- The $p \cdot 100$th percentile, $p \in (0, 1)$, is a value such that at least $p \cdot 100$ percent of the observations are less than or equal to this value and at least $(1 - p) \cdot 100$ percent of the observations are greater than or equal to this value.
- Multiple calculating methods: E.g., arrange the n data in ascending order. If pn is not an integer, the next integer greater than pn denotes the position of the pth percentile. If pn is an integer, the pth percentile is the average of the values in positions pn and $pn + 1$.
- The 85th percentile for the salary data is the data value in the 11th position, \$3730, as $0.85 \cdot 12 = 10.2$.

Measures of location

- First quartile or lower quartile Q_1 is the 25th percentile.
- Second quartile is the median, i.e. the 50th percentile, sometimes denoted as Q_2.
- Third quartile or upper quartile Q_3 is the 75th percentile.

3310 3355 3450 | 3480 3480 3490 | 3520 3540 3550 | 3650 3730 3925

$Q_1 = 3465$ $Q_2 = 3505$ (Median) $Q_3 = 3600$

- E.g. university accepts only students with 20% best score in a test. How many points do you need to collect to be admitted?

Measures of variability

- Measures of variability provide information about dispersion of values of data in the data set
- Range is the difference between the largest and the smallest value; rarely used as it depends on two extreme values only
- Interquantile range (IQR) is the difference between upper and lower quartiles, $Q_3 - Q_1$, i.e. the range for the middle 50% of the data.
- For the salary data, the range is $3925 - 3310 = 615$, while the IQR is $3600 - 3465 = 135$, When replacing the largest value \$3925 by $\$10,000$, then the range changes to 6690, however, IQR remains unchanged.

Measures of variability

- (unbiased) sample variance is the average square distance from the mean

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- It is expressed in units squared, thus its positive square root is often considered - sample standard deviation

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- the standard deviation is commonly used measure of the risk associated with investing in stock and stock funds. It provides measure of how monthly returns fluctuate around the long-run average return

Measures of variability

Monthly Salary (x_i)	Sample Mean ($\bar{x}$)	Deviation About the Mean ($x_i - \bar{x}$)	Squared Deviation About the Mean $(x_i - \bar{x})^2$
3450	3540	−90	8,100
3550	3540	10	100
3650	3540	110	12,100
3480	3540	−60	3,600
3355	3540	−185	34,225
3310	3540	−230	52,900
3490	3540	−50	2,500
3730	3540	190	36,100
3540	3540	0	0
3925	3540	385	148,225
3520	3540	−20	400
3480	3540	−60	3,600
		0	301,850
		$\Sigma(x_i - \bar{x})$	$\Sigma(x_i - \bar{x})^2$

- The sample variance of salaries is $\frac{301850}{11} = 27440.91$ (dollars^2) while the sample standard deviation is \$165.65

Properties of measures of location and variability

- Shifting all values of data by a (common) constant results in the same shift of a measure of location

$$m(x + a) = m(x) + a, a \in \mathbb{R}$$

- Multiplying all values of data by a (common) nonnegative constant results in proportional shift of a measure of location

$$m(bx) = bm(x), b > 0$$

- Shifts of values of data do not affect measures of variability

$$s(x + a) = s(x), a \in \mathbb{R}$$

- Multiplying all values of data by a (common) nonnegative constant results in proportional shift of a range, IQR or sample standard deviation

$$s(bx) = bs(x), b > 0$$

and quadratic in the case of sample variance

$$s(bx) = b^2 s(x), b > 0$$

Measures of shape

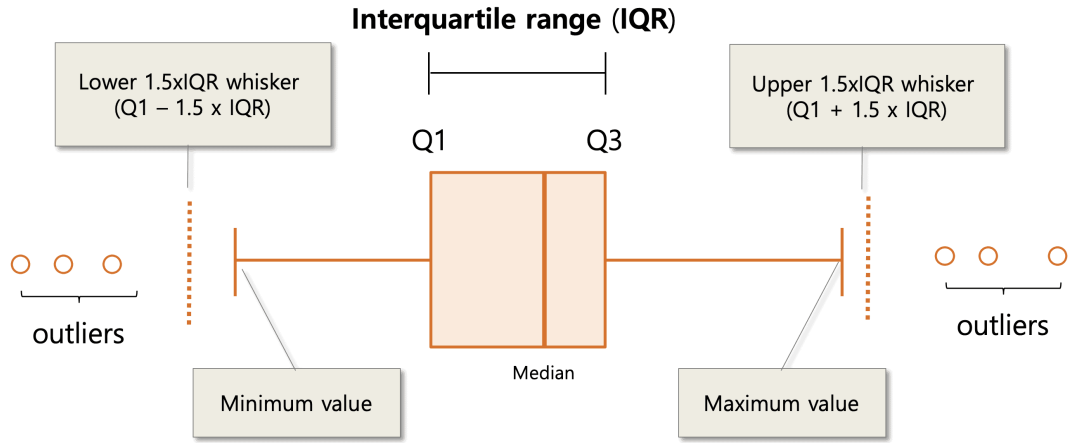
- z-score is a measure of relative location of values within a data set - how far a particular value is from the mean in terms of standard deviation as a unit

$$z_i = \frac{x_i - \bar{x}}{s}$$

- Sometimes called standardized value; positive (negative) z-scores indicate values larger (smaller) than mean. A z-score of zero indicates that the observation is equal to the sample mean.
- Useful for detection of outliers - data with unusual small or large values with respect to the data set. Can be an incorrectly recorded value or an observation incorrectly included in the data set (can be corrected or removed after careful inspection). Or it is just a correct value of an unusual observation that belongs to the data set.
- Empirical rule: outlier is an observation with z-score less than -3 or greater than 3.

Measures of shape

- For a five-number summary we specify the following five values to summarize the data: smallest value, lower quartile, median, upper quartile, maximum value
- Box plot is a graphical summary based on five-number summary. Interquantile range and outliers are also used.
- The endpoints of a box represent lower and upper quartile, a vertical line inside the box the median. Whiskers are drawn as lines from the quartiles to the last value within $\frac{3}{2}$ of the IQR. Data by more than $\frac{3}{2}$ of the IQR below Q_1 or above Q_3 (dashed lines depict the boundaries) are considered outliers.



Measures of association between two variables

- Previous measures summarized the data for one variable, we may be interested in descriptive measures of relationship between two (or more) variables
- Sample covariance measures the strength of the linear relationship between two variables

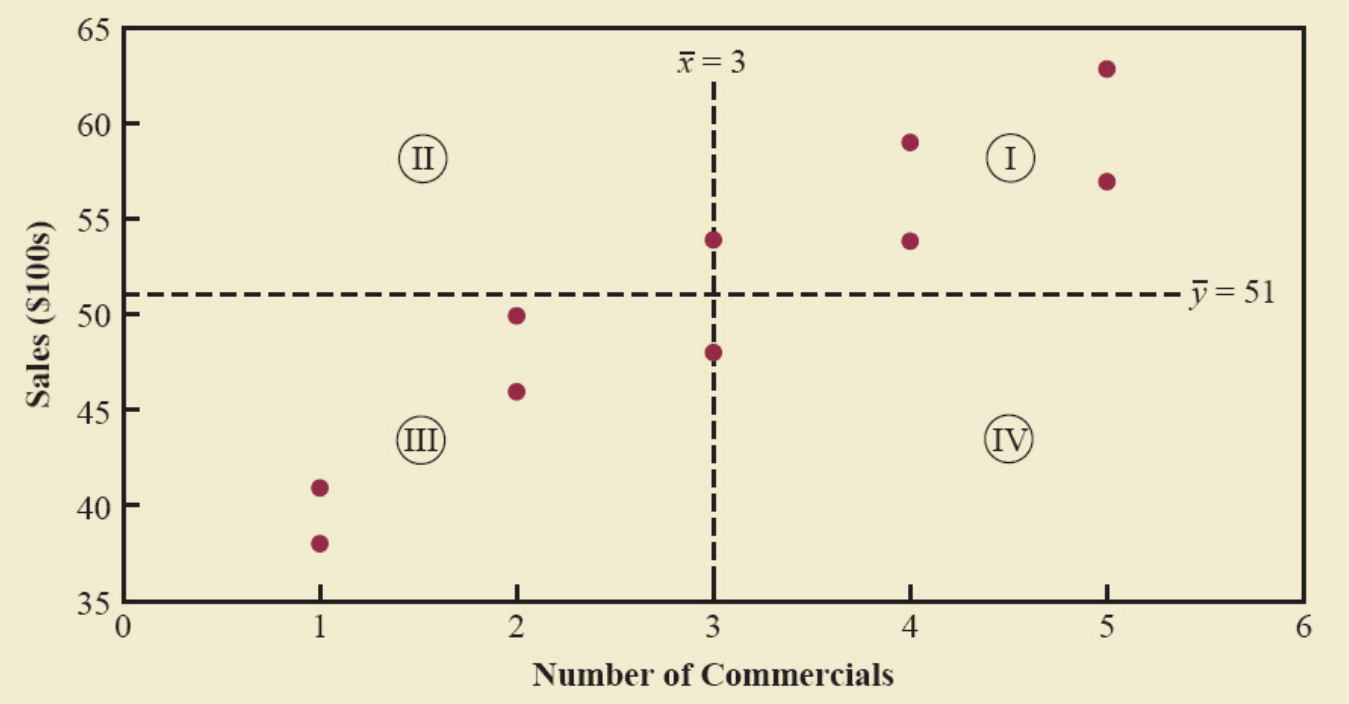
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Positive values indicate positive linear relation between x and y, however, the degree of linear relationship is affected by units of measurements of x and y
- Sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

takes values in $[-1, 1]$. The closer in absolute value to 1, the stronger the linear relationship (perfect for $\pm$ 1).

Example 7 cont. Sample data for the Stereo and sound system store



- Computing sample covariance between x (number of commercials) and y (sales volume): $s_{xy} = 11$.
- The sample correlation coefficient equals $r_{xy} = 0.93$.

Sample statistics and scales of measurement

- for nominal type of data, only mode is allowed, but median and mean make no sense to consider
- for ordinal type, mode and median are allowed, mean still makes no sense; interquartile range can be used as a measure of dispersion but other measures of dispersion make no sense
- for interval type, mode, median and mean are allowed to measure central tendency, measures of dispersion include range, interquartile range and standard deviation (those which do not require some ratios of values)
- for ratio type, all statistical measures are allowed because all necessary mathematical operations are well defined