# Statistics 1024

Chapter 7

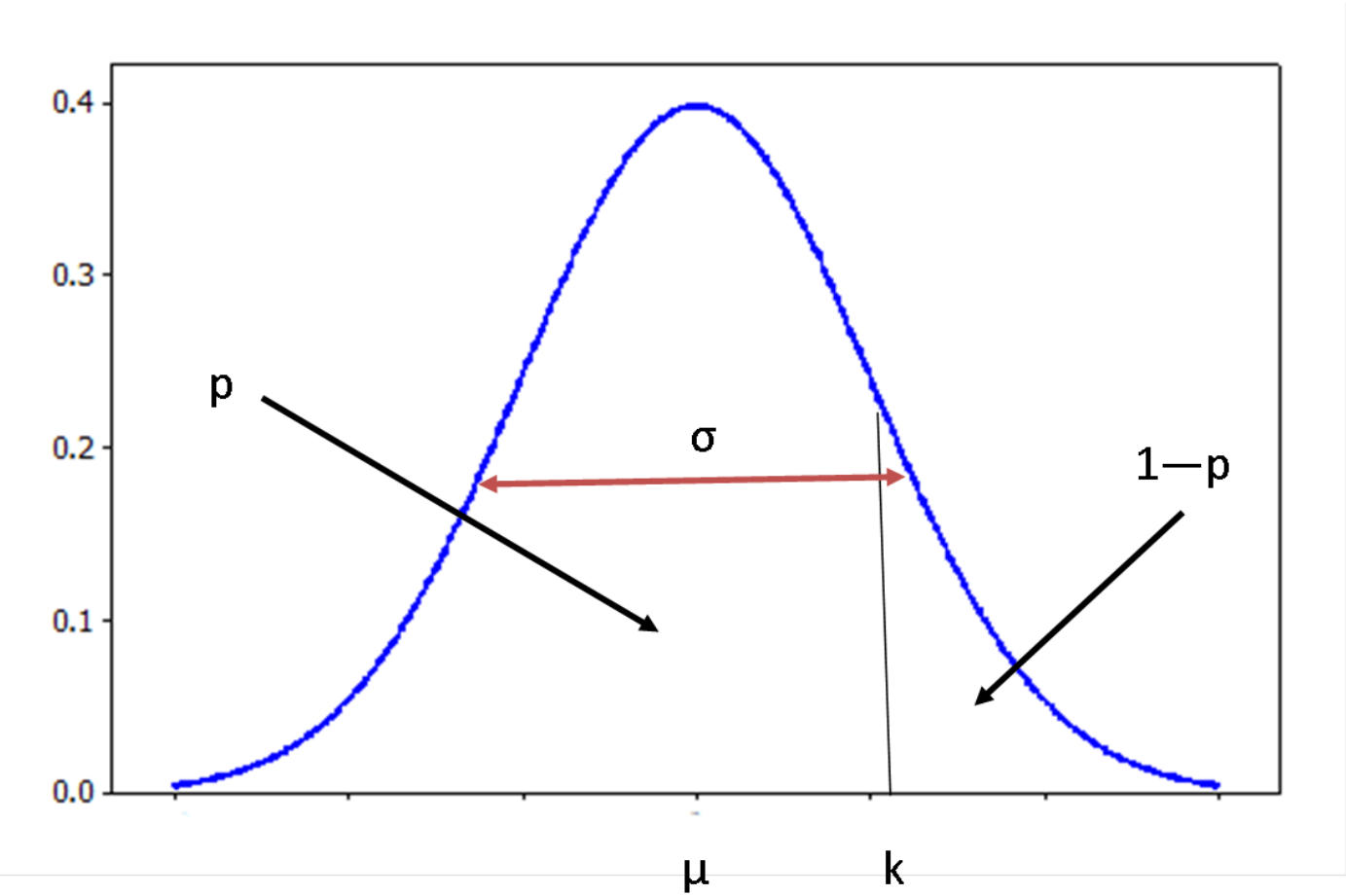
Review

# Normal Curve Problems

Reduces to two basic problems

- given the mean $\mu$, standard deviation $\sigma$ and the range of values through a constant $k$, find the appropriate area under the curve $p$
- given the area under the curve $p$ and any two of $\mu$, $\sigma$ and the range of values $k$, find the remaining one

# Looked at Graphically

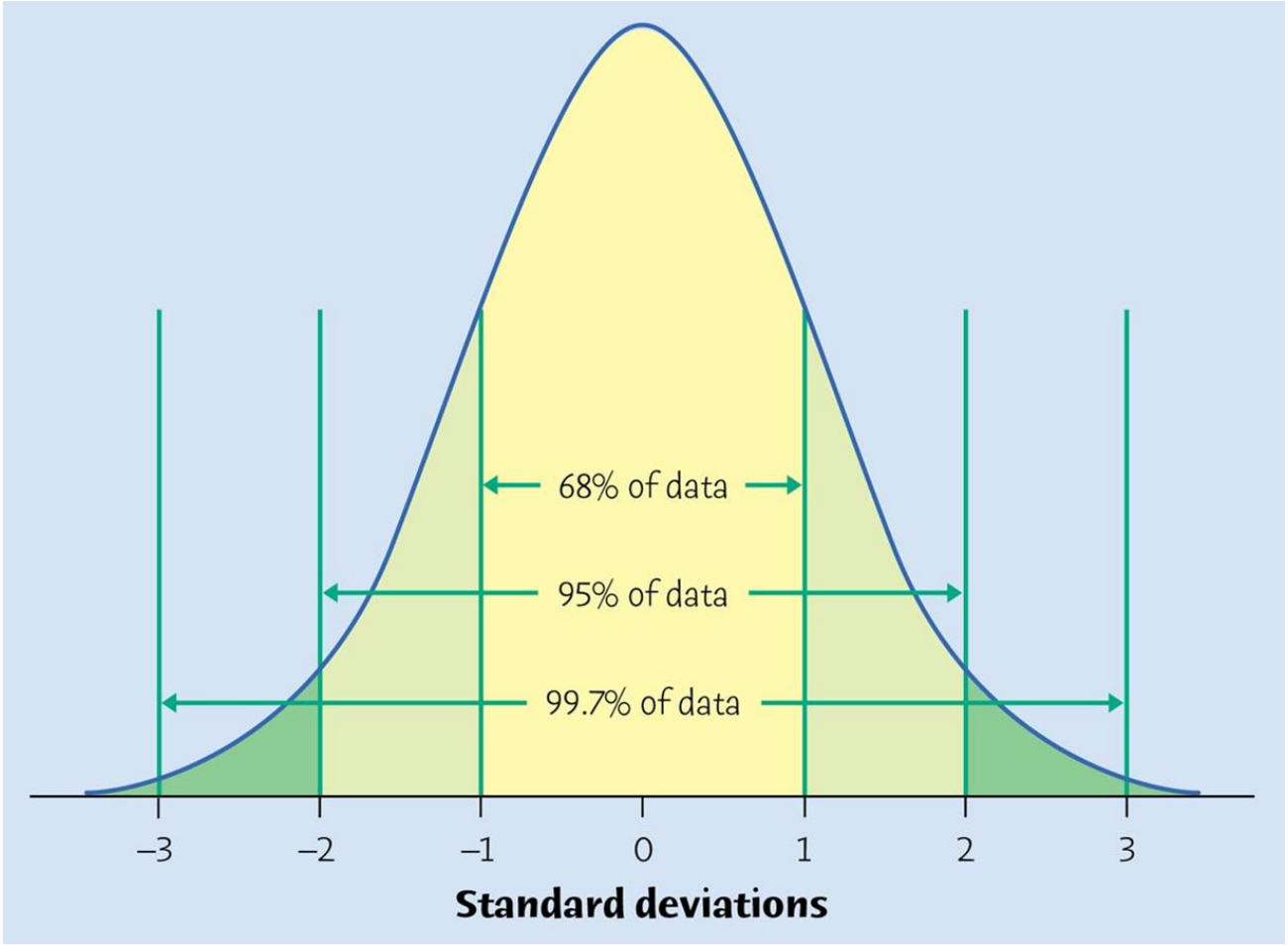


# Standardize

$$\frac{k-\mu}{\sigma} = z_p$$

- $z_p$ is that number from the standard normal distribution such that the area to the left of $z_p$ is *p*
- given any three of *k*, *μ*, *σ* and *p* (or $z_p$), the last one can be determined

# The 68-95-99.7 Rule as a Special Case




$z_{0.84} = 1$

$z_{0.16} = -1$

$z_{0.975} = 2$

$z_{0.025} = -2$

$z_{0.9985} = 3$

$z_{0.0015} = -3$

approximately

# Example

Children's gross motor development is assessed with an overall motor quotient, MQ, which follows an approximate normal distribution

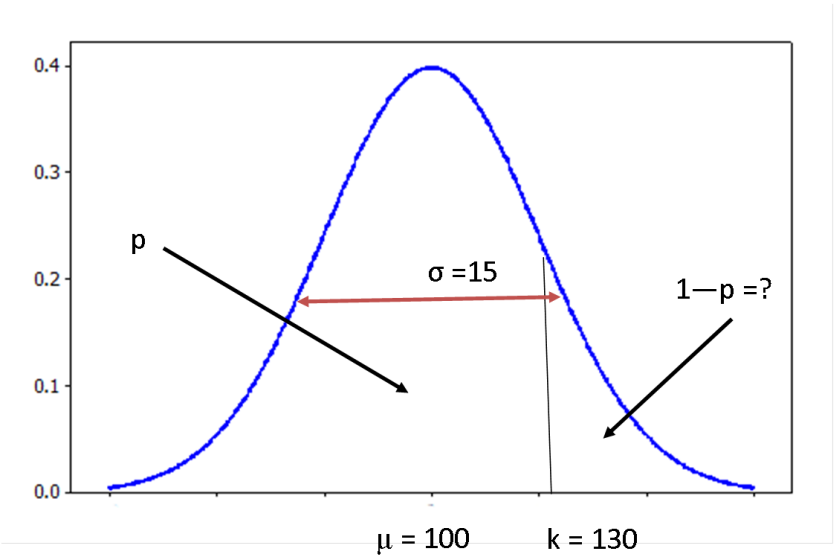
# Possible Questions

1) Given that the average motor quotient is 100 and the standard deviation is 15, what fraction of children have a motor quotient greater than 130?
2) Given that the average motor quotient is 100 and the standard deviation is 15, what is the lower quartile of children in terms of their motor quotient?

# More Questions

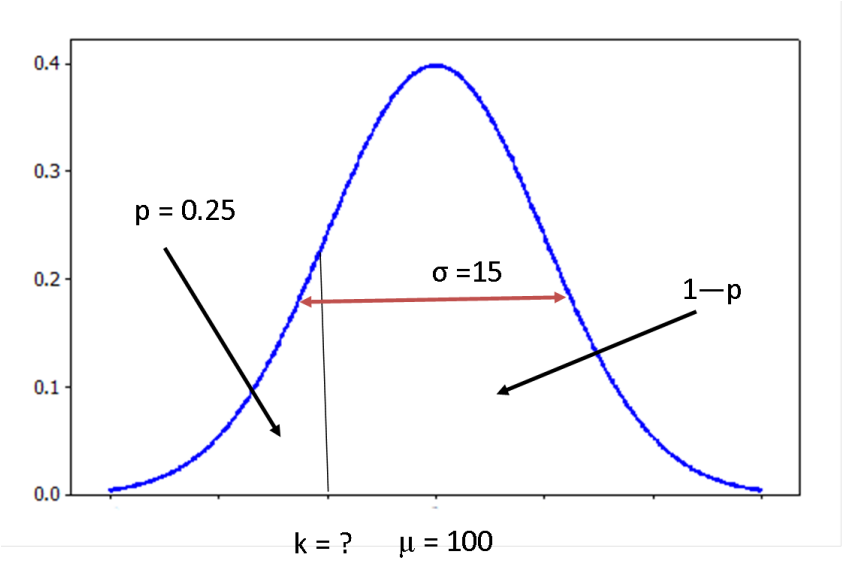
3) Given that the standard deviation of the motor quotient is 15, it is known that 10% percent of children have motor quotients greater than 119.2. What is the mean?
4) Given that the mean of the motor quotient is 100, approximately 68% of the children have motor quotients between 85 and 115. What is the standard deviation of motor quotients?

# Answer (1)



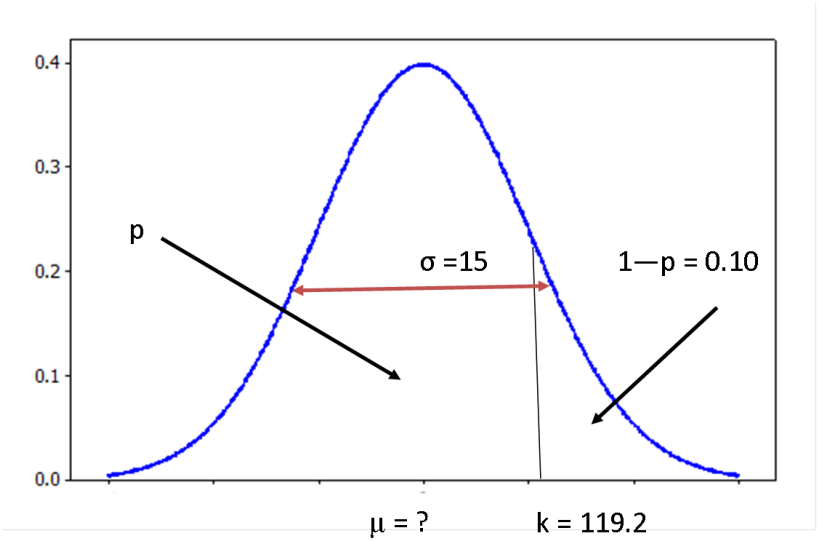

- $z_p = \frac{130-100}{15} = 2$
- Using the 68-95-99.7 rule 95% of the curve is between – 2 and 2 so that 2.5% is above 2 or $1 - p = 0.0.25$.
- From the tables, at 2 $p = 0.9772$ or $1 - p = 0.0228$.

# Answer (2)



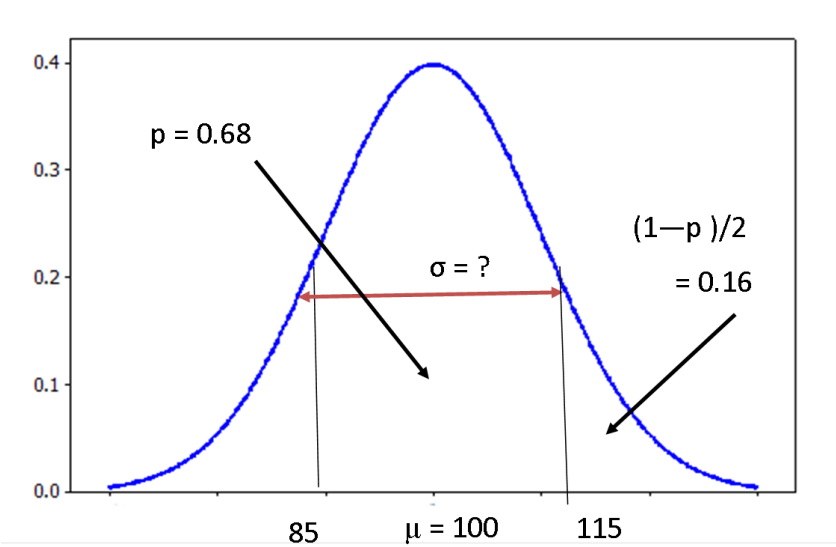

- $z_{0.25} = -0.67$
- $\frac{k-100}{15} = -0.67$
- $k = 89.95$

# Answer (3)



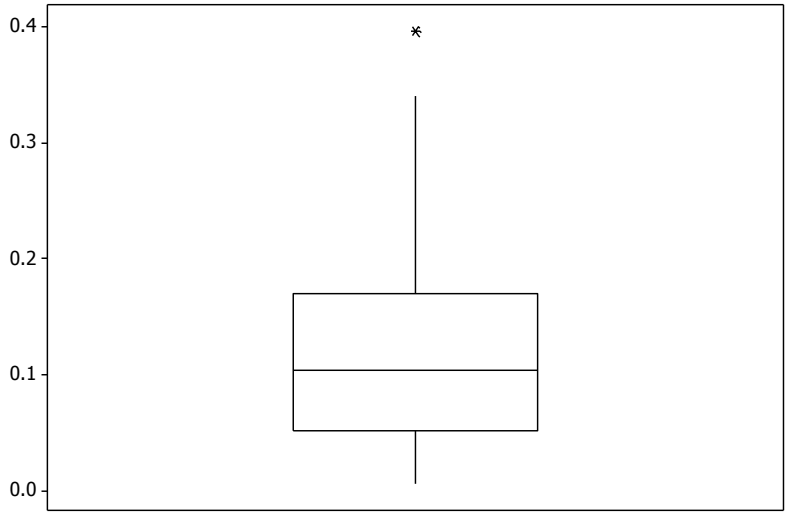

- $z_{0.90} = 1.28$
- $\frac{119.2-\mu}{15} = 1.28$
- $\mu = 100$

# Answer (4)



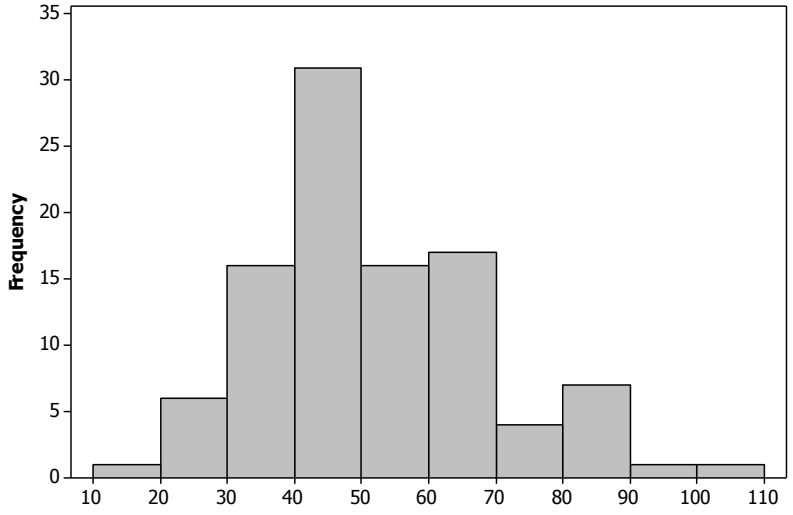

- Using the 68-95-99.7 rule 68% of the curve is within 1 standard deviation of the mean.
- From the numbers given we have 68% of the curve between 100 plus or minus 15.
- Therefore $\sigma = 15$.

# Know how to interpret or manipulate



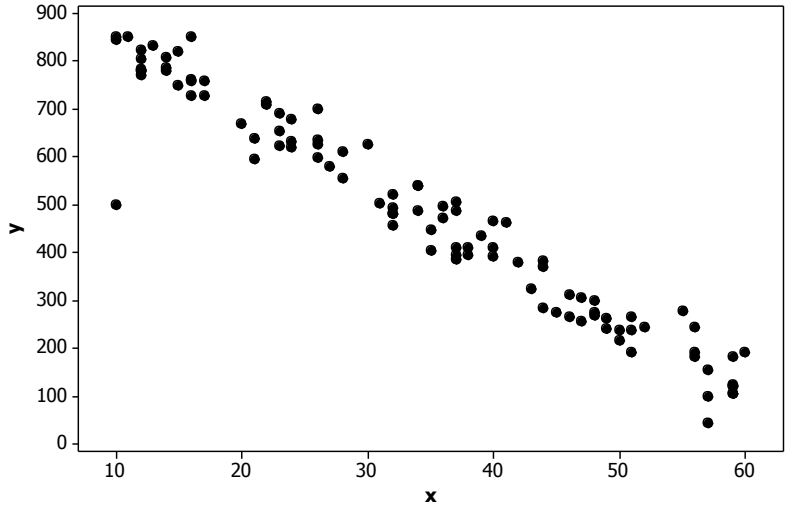

- What are the values of the five numbers in the five-number summary?
- Is the distribution symmetric or skewed?
- Is there an outlier?
  - How do you know it is an outlier?

# Know how to interpret or manipulate




- Is the distribution symmetric or skewed (in what direction)?
- What bin is the first quartile (or median or upper quartile) in?
- What fraction of the observations are greater than 80 (or less than 40)/

# Know how to interpret



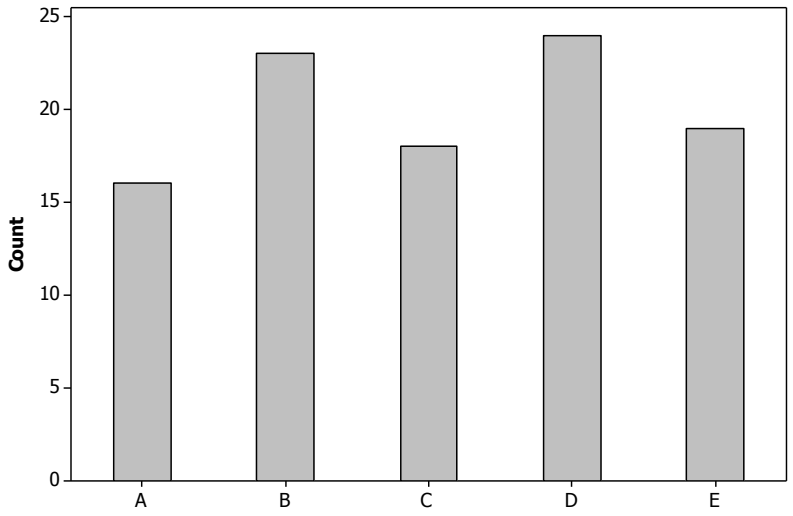

- What is the strength and direction of the trend?
- Is it strong or weak?
- What is a good guess for *r*, the correlation coefficient?
- Are there any outliers?
  - If so, what is the effect if removed?

# Know how to interpret or manipulate

| Stem | Leaf |
|---|---|
| 20 | 6 |
| 21 | 89 |
| 22 | 69 |
| 23 | 348 |
| 24 | 4455689 |
| 25 | 048 |
| 26 | 37 |
| 27 | 144 |
| 28 | 44 |

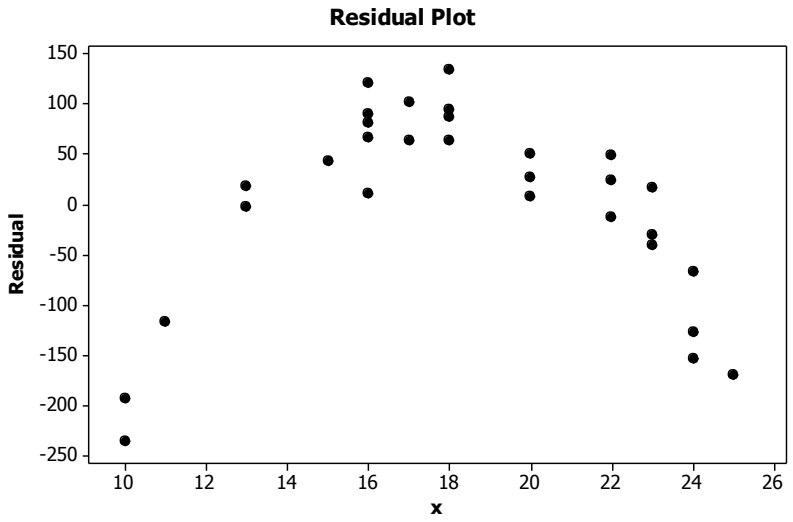
- Find the mean, median and upper and lower quartiles
- Is the distribution symmetric skewed?
- If skewed, in what direction?

# Know how to interpret or manipulate




- What fraction of the total is taken up by A’s and B’s?
- Would a pie chart be as useful?
  - If not, why not?

# Know how to interpret or manipulate




- Is there a pattern to the residuals?
  - What does it mean?
- Does the residual plot show any outliers?

# Know how to interpret or manipulate

| | A | B | C | D | Totals |
|---|---|---|---|---|---|
| X | 42 | 26 | 31 | 22 | 121 |
| Y | 28 | 28 | 22 | 50 | 128 |
| Z | 22 | 47 | 28 | 35 | 132 |
| Totals | 92 | 101 | 81 | 107 | 381 |

- Marginal totals (rows or columns)
- Marginal distributions (based on proportions)
- Conditional distributions (based on proportions)

# Know Concepts and Definitions with Any Associated Calculations

- Chapter 1
- Chapter 2
- Chapter 3
- Chapter 4
- Chapter 5
- Chapter 6
- Chapter 8

# STATISTICS IN SUMMARY

## Analyzing Data for One Variable

Plot your data:
Stemplot, histogram

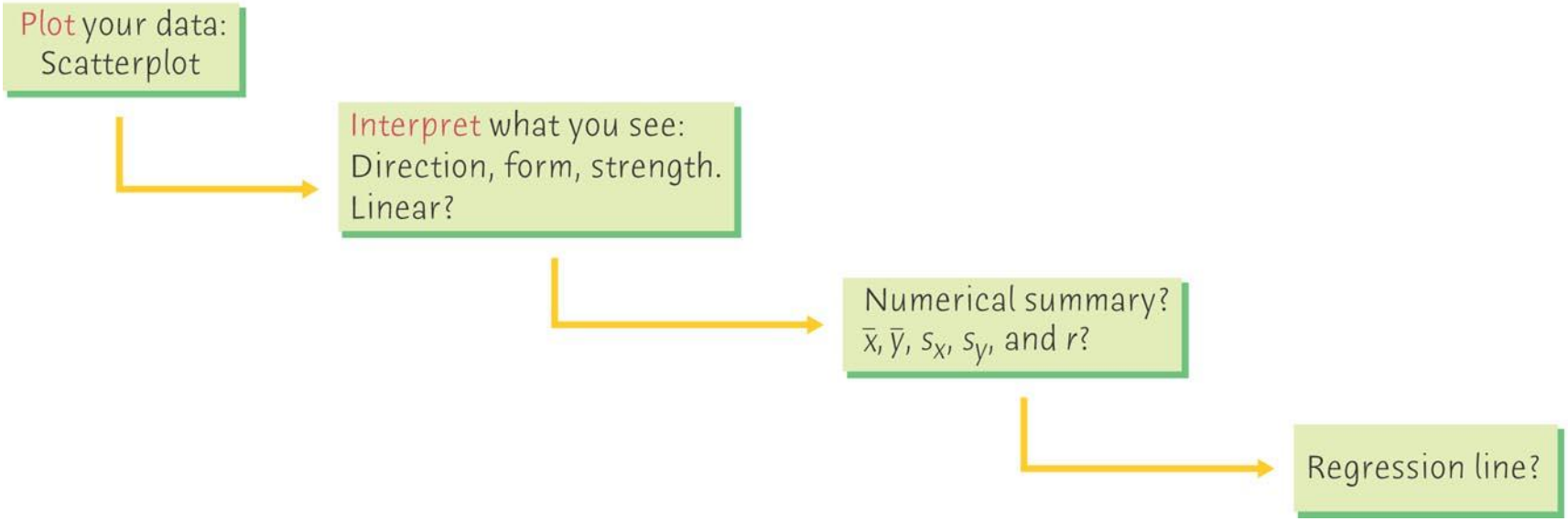
→ Interpret what you see:
Shape, center, spread, outliers

→ Numerical summary?
$\bar{x}$ and $s$, five-number summary?

→ Density curve?
Normal distribution?

## STATISTICS IN SUMMARY

### Analyzing Data for Two Variables



## A. DATA

1. Identify the individuals and variables in a set of data.
2. Identify each variable as categorical or quantitative. Identify the units in which each quantitative variable is measured.
3. Identify the explanatory and response variables in situations where one variable explains or influences another.

## B. DISPLAYING DISTRIBUTIONS

1. Recognize when a pie chart can and cannot be used.
2. Make a bar graph of the distribution of a categorical variable, or in general to compare related quantities.
3. Interpret pie charts and bar graphs.
4. Make a time plot of a quantitative variable over time. Recognize patterns such as trends and cycles in time plots.
5. Make a histogram of the distribution of a quantitative variable.
6. Make a stemplot of the distribution of a small set of observations. Round leaves or split stems as needed to make an effective stemplot.

## C. DESCRIBING DISTRIBUTIONS (QUANTITATIVE VARIABLE)

1. Look for the overall pattern and for major deviations from the pattern.
2. Assess from a histogram or stemplot whether the shape of a distribution is roughly symmetric, distinctly skewed, or neither. Assess whether the distribution has one or more major peaks.
3. Describe the overall pattern by giving numerical measures of center and spread in addition to a verbal description of shape.
4. Decide which measures of center and spread are more appropriate: the mean and standard deviation (especially for symmetric distributions) or the five-number summary (especially for skewed distributions).
5. Recognize outliers and give plausible explanations for them.

## D. NUMERICAL SUMMARIES OF DISTRIBUTIONS

1. Find the median M and the quartiles $Q_1$ and $Q_3$ for a set of observations.
2. Find the five-number summary and draw a boxplot; assess center, spread, symmetry, and skewness from a boxplot.
3. Find the mean $\bar{x}$ and the standard deviation $s$ for a set of observations.
4. Understand that the median is more resistant than the mean. Recognize that skewness in a distribution moves the mean away from the median toward the long tail.

5. Know the basic properties of the standard deviation: $s \geq 0$ always; $s = 0$ only when all observations are identical and increases as the spread increases; $s$ has the same units as the original measurements; $s$ is pulled strongly up by outliers or skewness.

## E. DENSITY CURVES AND NORMAL DISTRIBUTIONS

1. Know that areas under a density curve represent proportions of all observations and that the total area under a density curve is 1.
2. Approximately locate the median (equal-areas point) and the mean (balance point) on a density curve.
3. Know that the mean and median both lie at the center of a symmetric density curve and that the mean moves farther toward the long tail of a skewed curve.
4. Recognize the shape of Normal curves and estimate by eye both the mean and standard deviation from such a curve.
5. Use the 68–95–99.7 rule and symmetry to state what percent of the observations from a Normal distribution fall between two points when both points lie at the mean or one, two, or three standard deviations on either side of the mean.
6. Find the standardized value ($z$-score) of an observation. Interpret $z$-scores and understand that any Normal distribution becomes standard Normal $N(0, 1)$ when standardized.
7. Given that a variable has a Normal distribution with a stated mean $\mu$ and standard deviation $\sigma$, calculate the proportion of values above a stated number, below a stated number, or between two stated numbers.
8. Given that a variable has a Normal distribution with a stated mean $\mu$ and standard deviation $\sigma$, calculate the point having a stated proportion of all values above it or below it.

## F. SCATTERPLOTS AND CORRELATION

1. Make a scatterplot to display the relationship between two quantitative variables measured on the same subjects. Place the explanatory variable (if any) on the horizontal scale of the plot.
2. Add a categorical variable to a scatterplot by using a different plotting symbol or color.
3. Describe the direction, form, and strength of the overall pattern of a scatterplot. In particular, recognize positive or negative association and linear (straight-line) patterns. Recognize outliers in a scatterplot.
4. Judge whether it is appropriate to use correlation to describe the relationship between two quantitative variables. Find the correlation $r$.
5. Know the basic properties of correlation: $r$ measures the direction and strength of only straight-line relationships; $r$ is always a number between $-1$ and 1; $r = \pm 1$ only for perfect straight-line relationships; $r$ moves away from 0 toward $\pm 1$ as the straight-line relationship gets stronger.

## G. REGRESSION LINES

1. Understand that regression requires an explanatory variable and a response variable. Use a calculator or software to find the least-squares regression line of a response variable $y$ on an explanatory variable $x$ from data.
2. Explain what the slope $b$ and the intercept $a$ mean in the equation $\hat{y} = a + bx$ of a regression line.
3. Draw a graph of a regression line when you are given its equation.
4. Use a regression line to predict $y$ for a given $x$. Recognize extrapolation and be aware of its dangers.
5. Find the slope and intercept of the least-squares regression line from the means and standard deviations of $x$ and $y$ and their correlation.
6. Use $r^2$, the square of the correlation, to describe how much of the variation in one variable can be accounted for by a straight-line relationship with another variable.
7. Recognize outliers and potentially influential observations from a scatterplot with the regression line drawn on it.
8. Calculate the residuals and plot them against the explanatory variable $x$. Recognize that a residual plot magnifies the pattern of the scatterplot of $y$ versus $x$.

## H. CAUTIONS ABOUT CORRELATION AND REGRESSION

1. Understand that both $r$ and the least-squares regression line can be strongly influenced by a few extreme observations.
2. Recognize possible lurking variables that may explain the observed association between two variables $x$ and $y$.
3. Understand that even a strong correlation does not mean that there is a cause-and-effect relationship between $x$ and $y$.
4. Give plausible explanations for an observed association between two variables: direct cause and effect, the influence of lurking variables, or both.

## I. CATEGORICAL DATA (Optional)

1. From a two-way table of counts, find the marginal distributions of both variables by obtaining the row sums and column sums.
2. Express any distribution in percents by dividing the category counts by their total.
3. Describe the relationship between two categorical variables by computing and comparing percents. Often this involves comparing the conditional distributions of one variable for the different categories of the other variable.
4. Recognize Simpson's paradox and be able to explain it.