

What you need to know (Ch.1):

- **Data**
- **Population**
- **Census**
- **Sample**
- **Parameter**
- **Statistic**
- **Quantitative data**
 - **Discrete data**
 - **Continuous data**
- **Categorical data**
- **Four levels of measurement**
 - The nominal level
 - The ordinal level
 - The interval level
 - The ratio level
- **Observational study**
 - Cross-sectional study
 - Retrospective study
 - Prospective study
- **Experiment**
 - **Randomization**
 - **Replication**
 - **Blinding**
 - **Placebo effect**
 - **Placebo/control group**
 - **Double-blind experiment**
 - **Controlling effects of variables**
 - **Confounding**
 - **Completely randomized experimental design**
 - **Randomized block design**
 - **Rigorously controlled design**
 - **Matched pair design**
 - **Outline the design of an experiment**
- **The pitfalls in experimentation**
 - Lack of realism
 - Non-compliance
 - Impractical and unethical experiments
- **Sampling error**
- **Nonsampling error**
- **Systematic sample**
- **Convenience sample**
- **Probability sample**
 - **Simple random sample (SRS)**
 - **Cluster sample**
 - **Stratified sample**
 - **Multistage sample**
- **Non-response**
- **Sample surveys**
 - **Wording of question**
 - **Ordering of questions**
 - **Type of questions asked**

What you need to know (Ch.2):

- **How to create and interpret graphical displays for one categorical variable**
 - Pie chart (you don't need to make them, just "read" them)
 - Bar graph
- **How to create and interpret graphical displays for one quantitative variable**
 - Histogram (frequency, and relative frequency)
 - Dotplot
 - Stemplot
- **You need to be able to summarize the most important features of statistical graphs**
 - Shape (symmetric or skewed, uni-, or bimodal, uniform)
 - Center
 - Spread
 - Outliers

What you need to know (Ch.3):

- **The four measures of center**
 - Mean
 - Median
 - Mode
 - (Midrange)
 - **You need to be able to calculate these above**
 - **You need to be able to compare the two main measures for the center: the mean and the median**
 - **You need to know when you can use the mean, and when you can use the median to describe the center of a distribution**
 - **You need to know the relative relationship between the mean and the median when the distribution is symmetric, skewed to the right, and skewed to the left**
 - **The three measures of spread**
 - Range
 - IQR
 - Standard deviation
 - **You need to know how to calculate the first two of these**
 - **You need to know which measure of center is paired up with which measure of spread, and which pair to use to describe symmetric and skewed distributions**
 - **You need to understand standard deviation (for example to guess which distribution has larger standard deviation)**
 - **You need to know how to find the five-number summary**
 - **You need to know how to construct a boxplot**
 - **You need to be able to identify outliers using the 1.5(IQR) rule**
 - **You need to know how to construct a modified boxplot (when you mark the outliers with asterisks)**
 - **You need to be able to use the Empirical Rule**
- You need to be able to convert raw score (x) to z-scores, and decide if the value usual or unusual**

What you need to know (Ch.10):

- **You need to know how to create a scatterplot**
- **You need to know what types of variables can be graphically represented in a scatterplot**
- **You need to be able to describe a scatterplot**
 - **Direction**
 - **Form**
 - **Strength**
 - **Deviations from the pattern**
- **You need to understand what the correlation coefficient measures, and its properties**
- **You need to know what the least squares regression line is**
- **You need to know how to find the equation of the least squares regression line**
- **You need to be able to use the equation of the LSRL to predict some values of the response variable**
- **You need to know how to interpret in context the slope and the y-intercept of the regression line**
- **You need to know what a residual for a given point is**
- **You need to know what extrapolation is**
- **You need to understand what a lurking variable is**
- **You need to understand and apply the concept that association does not imply causation**

1. In statistics, what is meant by a **variable**?
 2. What is the difference between a **categorical variable** and a **quantitative variable**? (Also give an example of each.)
 3. What is meant by **exploratory data analysis**?
 4. What is meant by the **distribution** of a variable?
 5. What two types of charts/graphs are usually most appropriate for **categorical** data?
 6. Is there always space between the bars of a **bar graph**?
 7. What is an important “must” for **pie charts**?
 8. When describing the overall pattern of a distribution of a quantitative variable, what **three features** should you mention?
 9. What is a simple way to describe the **center** of a distribution of a quantitative variable?
 10. How do you describe the **spread** of a distribution of a quantitative variable?
 11. Informally define an **outlier**.
 12. List four graphs that are used for **quantitative** data.
 13. When is it better to use a **stemplot** rather than a **dotplot**?
 14. List the steps for constructing a **stemplot**.
 15. What information is lost when you choose a **histogram** over a **dotplot** or a **stemplot**?
 16. Are there spaces between the bars of a **histogram**?
 17. List the steps for constructing a **histogram**.
 18. If a distribution is **skewed right**, what does its shape look like?
 19. If a distribution is **skewed left**, what does its shape look like?
 20. In statistics, what are the most common **measures of center**?
 21. Explain how to calculate the **mean**, \bar{x} .
 22. Explain how to find the **median**, M .
 23. Explain why the median is **resistant** to extreme observations, but the mean is **nonresistant**.
 24. The **mean** and **median** are close together if the distribution is what?
 25. In a skewed distribution, which will be farther towards the long tail—the mean, or the median?
 26. Which measure is most appropriate for a highly skewed distribution—the mean, or the median?
 27. In statistics, what is meant by **range**?
 28. Explain how to calculate the **first quartile** Q_1 and the **third quartile** Q_3 .
 29. What is the **inter-quartile range (IQR)**?
 30. Explain why it might be better to use the **IQR** instead of the **range** to describe the spread of the distribution.
 31. What is the IQR based “rule of thumb” for defining **outliers**?
 32. What is the **five-number summary**?
 33. What type of graph gives a picture of the **five-number summary**?
 34. The “box” in a **boxplot** represents what percentage of the data?
 35. The middle line of a **boxplot** represents the _____.
 36. Can the value of the **mean** be identified from a **boxplot**?
 37. What does **standard deviation** measure?
 38. The sum of all the **deviations** of the observations from their mean will always be _____. Explain why.
 39. When does the **standard deviation** equal zero?
 40. Can the **inter-quartile range** or the **standard deviation** ever be negative?
 41. Is the **standard deviation** resistant or nonresistant to extreme observations? Explain.
 42. When is it better to use the **five-number summary** versus the **mean and standard deviation**?
-
1. What is the difference between a **response variable** and an **explanatory variable**?
 2. What is another set of terms for “response and explanatory variables”?
 3. A **scatterplot** shows the relationship between two _____ variables.

4. Which variable always appears on the *horizontal axis* of a scatterplot?
5. When describing a *scatterplot*, to what three aspects of the pattern should you refer?
6. **True or false:** In describing the form of a scatterplot, it is important to say whether the graph appears to be linear or not.
7. In describing the *direction* of a scatterplot, when there is a positive or negative slope, we say that the variables are positively or negatively _____.
8. **True or false:** In describing the *strength* of a scatterplot, we look at the amount of “scatter” in the data points—how close the points lie to a simple form such as a line.
9. Suppose that you want your scatterplot to reflect the relationship of a third, categorical variable, in addition to the relationship of the two quantitative variables that are plotted. For example, suppose you want to use one graph to show the relationship between percent of body fat and waist size, in such a way that also shows the relationship separately for males and for females. What should you do?
10. What is the best method for judging the strength of a linear relationship: simply to look at the graph, or to use a calculated numerical statistic that summarizes the strength of the linear relationship, or something else?
11. What does *correlation* measure?
12. What is the meaning of a *positive/negative sign* associated with the correlation coefficient?
13. **True or false:** Correlation makes a distinction between the explanatory and response variables.
14. **True or false:** A correlation coefficient has units.
15. What is true about the *relationship* between two variables if the *r-value* is:
 - a. Near 0?
 - b. Near 1?
 - c. Near -1?
 - d. Exactly 1?
 - e. Exactly -1?
16. What sort of correlation coefficient do you find when two variables have a very strong linear relationship, and when the *first gets greater, the second gets smaller*?
17. Suppose that for each of the days of 2006 we know the number of words Mrs. O. spoke in that day (variable 1), and the peak barometric pressure for that day in Caracas, Venezuela (variable 2). About what would you guess the correlation between these two variables to be? Why?
18. Suppose there are two variables which, when graphed in a scatterplot, form an almost perfect U-shaped parabola. Would the strong relationship between these variables imply a high correlation coefficient (meaning close to 1 or -1)? Why or why not?
19. Does the correlation coefficient resemble the median and IQR in being fairly *resistant* to outliers, or does it resemble the mean and standard deviation in being heavily influenced by outliers (i.e. *non-resistant*)?
20. Finish this statement: A *regression line* is a straight line that is used to _____.
21. The *equation* for a LSRL is $Y =$ _____
22. The slope of a LSRL is $b =$ _____.
23. The intercept of a LSRL can be found by $a =$ _____.
24. Interpreting the *slope* is important—think of it as a “*rate of change*”. That is the amount of change in _____ when _____ increases by one unit.
25. Once you have a LSRL, how do you find a predicted value of y for a given x -value?
26. Suppose that someone measures height and weight for a bunch of human adults, and gets a regression equation predicting height from weight. Why is the y -intercept of the equation not as meaningful or important as the slope, or as the equation as a whole?
27. **True or false:** In a *regression line*, like a correlation coefficient, you get the same numbers (slopes and intercepts) no matter which variable is considered the explanatory variable and which is considered the response variable.
28. Every LSRL passes through the point (_____ , _____)

29. What is *extrapolation* and what is the problem with it?
 30. Define *lurking variable*. Why is it such an important concept?
 31. If two variables have a strong positive association, then the larger the value of one variable, the larger the value of the other variable. Is it fair to say that an increase in one variable *causes* an increase in the other variable? Explain.
1. Why are *voluntary response samples* unreliable?
 2. Why is *convenience sampling* be unreliable?
 3. What is a *biased study*?
 4. Define *simple random sample*.
 5. What is a *stratified random sample*?
 6. Give an example of *undercoverage* in a sample.
 7. Give an example of *response bias* in a sample.
 8. How can the wording of questions cause *bias* in a sample survey?
 9. Explain the difference between an *observational study* and an *experiment*. **Important!**
 10. Describe the *placebo effect*.
 11. What is the significance of using a *control group*?
 12. What do we mean by *random assignment*?
 13. What are the advantages of a *double-blind study*?
 14. Describe a *block design*.
 15. Describe a *matched pairs design*.
16. How would you describe the shape of a *normal curve*? Draw several examples.
 17. Draw and describe the Standard Normal Distribution.
 18. Where on the normal curve are the *inflection points* located?
 19. What is the Empirical Rule?
 20. Explain how to *standardize* a variable.
 21. What do z-scores mean? When is a value considered to be usual/unusual?
22. Explain the difference between a *population* and a *sample*. **Important!**
 23. Explain the difference between a *parameter* and a *statistic*. **Important!**
 24. Explain the difference between the meaning of the symbols p and \hat{p} .
 25. Explain the difference between the meaning of the symbols \bar{x} and μ .