Advanced Functions – Statistics Vocabulary Reading Sheet Name: _____

Directions – There are A LOT of vocabulary in statistics and we will NOT have time to cover them all in our notes. There are also many famous statisticians that have contributed to statistics. Read the following passage about statistics and people who contributed to this subject.

Data is the currency of the information age. The real value of Facebook is not in the degree of interaction members have, but in the amount of information the company collects — information which can, in turn, be sold to give merchants insights into what consumers want. The same is true of Google and Yahoo — while they provide a service to those who wish to search the internet, their value is the information they provide to those who have products and services to sell. **Statistics** is the art and science of extracting meaning from data.

Accurate assessments of consumer trends do not require a ***census*** — contacting an entire population. Attempting to contact every member of a particular group (a population) is often impractical. Instead, it can be shown mathematically that careful *sampling* can produce just as accurate results. ***Sampling*** is the process of collecting data from smaller groups within the population to predict how the entire population feels, reacts, etc.

What type of data are we talking about?

Age, gender, geographic location, products of interest, subjects of interest, politics, what you "like," +1, or dislike.

In the language of statistics, Facebook users are called ***subjects***. The data collected on each subject is referred to as a ***variable***. All variables fall into one of two types: categorical and quantitative.

> ***Categorical variables*** assign characteristics of each subject to a group or category: eye color, gender, favorite genre of music, etc. Typically, this is non–numerical
>
> ***Quantitative variables*** are numerical in nature and will most likely be mathematically manipulated in some way: height, weight, salary, etc. In each case, we would probably want a measure of center and spread for the data.

Note: not all numbers are quantitative variables. For example, phone numbers and zip codes are numerical, but it's highly unlikely that we'll want an average zip code! Also note: some quantitative variables are frequently converted to categories. For example, ages may be broken down into categories: 15 – 19, 20 – 24, 25 – 29, etc.

There are three ways to collect data: observational studies, experiments, and simulations. The choice and specifics of how a researcher will collect data is called the ***sample design***.

> In an ***observational study***, the researcher simply observes phenomena or responses, striving to avoid influencing the subjects in any way. Although it involves direct communication, a survey is classified as an observation study.
>
> In an ***experiment***, the researcher intentionally imposes a "treatment" upon the subjects. Drug companies typically use experiments to collect data on the effectiveness and side–effects of new medications. Experiments usually employ a ***placebo***, a "dummy" treatment that is known to have absolutely no effect. Those that receive the placebo are called the ***control group***. When the experiment is completed, the effects of the treatment are compared with the control group to determine what effects the treatment had. [Sometimes the control groups actually shows a response to the treatment! This is known as the ***placebo effect***.] The best experiments are **double–blind**: that is, neither the person receiving, nor the person giving the treatment know whether the actual product or a placebo is being administered.
>
> A ***simulation*** is used when an observational study or experiment is impractical or unethical. Simulations use mathematical models and probabilities to examine possible outcomes.

In all sample designs, ***randomization*** is crucial. Selecting subjects completely by random is a sure way to avoid ***bias***, the systematic preference for one outcome over another.

There are some bad sample designs, which are quite common.

> A ***voluntary response sample*** allows subjects to choose whether or not they wish to participate. A common example: many radio and TV talk shows ask their viewers to respond to a survey question by calling or texting. Typically, only those with strong (and usually negative) opinions are the most likely to respond.
>
> ***Convenience sampling*** takes place when the researcher uses subjects who are the easiest to reach. A common example is the "man or woman on the street" interview. But consider this: people who are out on the street during a typical day most likely do not reflect the population as a whole. Many people have jobs that do not allow them to be out and about.

While these poor designs can be avoided, some problems are just inherent and must simply be acknowledged.

> ***Undercoverage*** refers to the fact that some elements of a population are bound to be overlooked despite our best efforts. The homeless and those without phones are typically left out of national polls; it is very difficult to contact such people.
>
> ***Non response*** refers to those who are chosen, but who (a) can't be contacted or (b) refuse to participate. Folks who work night shifts typically sleep during the day when many researchers are collecting data, and some folks just do not want to be bothered.

Surveys are observational studies in which researchers pose questions (in person or by other means) and record the subjects' responses. Surveys have their own special problems: those who use surveys to collect information must be sure that that those who conduct the survey have been carefully selected and trained. In addition, careful attention must be given to the wording of the questions. Common problems include:

> ***Asking biased or leading questions***. By asking questions in a certain way, the researcher can lead the respondents to answer in the way he or she wants them to. For example, asking a question such as “Are you going to vote for candidate Jones even though the latest survey indicates that he will lose the election?” instead of “Are you going to vote for candidate Jones?” may dissuade some people from answering in the affirmative.
>
> ***Using confusing words***. In this case, the participant misinterprets the meaning of the words and answers the questions in a biased way. For example: "Despite the climatic vicissitudes of Middle East politics, would you espouse a solution that circumvents diplomacy vis–á–vis military intervention?"
>
> ***Asking double–barreled questions***. Sometimes questions contain compound sentences that require the participant to respond to two questions at the same time. For example, the question “Are you in favor of a special tax to provide national health care for the citizens of the United States?” asks two questions: “Are you in favor of a national health care program?” and “Do you favor a tax to support it?”
>
> ***Using double negatives in questions***. Questions with double negatives can be confusing to the respondents. For example, the question “Do you feel that it is not appropriate to have areas where people cannot smoke?” is very confusing since not is used twice in the sentence.
>
> ***Ordering questions improperly***. By arranging the questions in a certain order, the researcher can lead the participant to respond in a way that he or she may otherwise not have done. For example, a question might ask the respondent, “At what age should an elderly person not be permitted to drive?” A later question might ask the respondent to list some problems of elderly people. The respondent may indicate that transportation is a problem based on reading the previous question.

Carefully planned surveys often take a considerable amount of time to prepare, test, refine, and execute.

FAMOUS STATISTICIANS

Sir Ronald Fisher (1890 – 1962) was an English geneticist and statistician, who almost single–handedly created the foundations for modern statistical science. His contributions have led some to call him "The Father of Statistics."

John W. Tukey (1915 – 2000) was a professor of statistics at Princeton University. He made many contributions to the field of statistics, but most middle and high school students know of only one of his inventions: the boxplot, or "box–and–whiskers" plot, used and described by Tukey in 1977.

Gertrude Mary Cox (1900 – 1978) was founder of the Department of Experimental Statistics at North Carolina State University (NCSU). In 1949 Cox became the first female elected into the International Statistical Institute, and in 1956, she was elected president of the American Statistical Association.

David Blackwell (1919 – 2010) originally taught mathematics at the prestigious Howard University in Washington, DC, before becoming professor at the University of California (Berkeley) in 1954. Through his interest in the behavior of statistical procedures as more and more data was added, he gained much insight into what was needed to reach firm conclusions. His work with probability (gaming) theory is considered foundational for statistical inference.

Carl Friedrich Gauss (1777 - 1875) was a German mathematician who made many contributions to the field. Through his attempts to mathematically predict the movement of stellar bodies, he developed the most commonly used method to determine a line of best fit: the least squares regression method. In addition, he developed much of the foundational theory for working with normal distributions, often referred to as Gaussian distributions.