From What-Ever!!!
To Understanding and Applying Basic Statistics

By
Garnett Lee Henley

Contents

Part 1
Chapters 1 - 7

Chapter 1 Introduction
Chapter 2 Common Types of Variables

Chapter 3 Two Types of Problem Statements
- Research Question
- Hypotheses Statements
  - Null Hypothesis
  - Alternative Hypothesis
  - Directional Hypothesis
  - Non-Directional Hypothesis

Chapter 4 Data Scales and Data Types
- Nominal and Ordinal
- Interval and Ratio

Chapter 5 Common Research Designs and Models

Chapter 6 Introduction to Probability Theory
- Single Event Probabilities
- Mathematical Rules and Multiple Event Probabilities

Chapter 7 Populations, Samples and Sampling
- Population vs Sample
- Bias, Randomness and Independence
- Two Sampling Approaches and Models
  - Non-Probability
  - Probability
- Sampling Terminology
- Use of Random Numbers Table
- Example of Sample Size Calculations for Means and Proportions

Chapter 8 Descriptive Statistics
- Measures of Central Tendency (For Interval and Ratio Data)
- Measures of Dispersion (For Interval and Ratio Data)
  - Residuals and Variance
  - Standard Deviation
  - Coefficient of Variation
- Frequencies and Cross-Tabulations (Nominal and Ordinal Data)

Chapter 9 The Normal Distribution and Hypothesis Testing
- Chebychev's Rule
- The Empirical Rule
- Z-Scores
- Confidence Interval of the Mean
- Standard Error of the Mean
- Meaning of "P", Statistical Significance, Power, α and β

PART 2
CHAPTERS 8 - 14

Chapter 10 Inferential Statistics
- Parametric vs Non-Parametric Methods
- Measures for Nominal and Ordinal Data
  - Chi-Square Distribution and Test
  - Odds Ratios and Relative Risk
  - Rates
  - Sensitivity and Specificity
  - Incidence and Prevalence

Chapter 11 Measures for Interval and Ratio Data
- Concept of Homogeneity of Variance
- The T-Distribution and T-test
- The F-Distribution and 1 & 2 Way ANOVAs
- Correlation Analysis
- Introduction to Linear Regression
- Introduction to Binary Logistic Regression

Chapter 12 Glossary of Statistical Terms

Chapter 13 Glossary of Statistical Formulae

Chapter 14 Statistical Tables

Introduction

Statistics employ mathematical algorithms that assess data with a degree of confidence and assurance for differences and relationships. Much of the analysis involves accounting for variance (how observations disperse about the mean) and determining the amount of error (systemic and non-systemic) that is present in the data. An acceptable definition of statistics would be to conclude that it is, in part, *the accounting for the presence and effects of variance in a dataset.* The outcome of statistical analysis is a value that is interpreted as the *likelihood or probability* that the event occurred by chance and a degree of assurance that the event will re-occur at an predictable frequency. We must be careful, therefore, to limit the interpretation of any statistical analysis to the *measurement of the probability* that an event (given the same circumstances) repeats with a given frequency.

There are actually two types of statistics: Descriptive and Inferential. Descriptive statistics provide the means by which data can be organized and presented. The approach here is mostly with analyzing individual variables in a dataset. Descriptive statistics has two analysis components: *Measures of Central Tendency* – which looks at the location of observations in a dataset, and includes assessing for a mean, mode and median; and *Measures of Dispersion* – which looks at how observations disperse about a mean, and assesses for variance, standard deviation, standard error of the mean and coefficient of variation. With both components, low power assumptions can be made about the findings, but care must be taken not to infer that the findings truly represent the population from which the data were taken.

Inferential statistics are based on complex mathematical algorithms that are probability based. Almost every inferential test has as part of its equation a formula to correct for, account for, or control for the effects of variance. Inferentials are powerful enough to allow samples of data to represent the population from which they were taken. There are two types of analysis approaches in inferential statistics: Parametric and Non-Parametric. Parametric approaches use tests that assume that the data are normally distributed (more on this later in chapter ___), while non-parametric tests do not assume that they are processing normally distributed data.
.

Variables

Variables are terms that have been assigned a numeric meaning. The numeric meaning suggests that they are "code words" that have numeric form so that they can be counted and measured in other ways. They are nominal in nature because each has a unique name. Variables are used in research questions and hypotheses as objects that identify the focus of the research. Age, height, weight, and hair color are examples of variables. Each term is sufficiently different from the other, and each can be measured by various means we will discuss in detail in descriptive and inferential statistics. Therefore, uniqueness and measurability are key characteristics of variables. Uniqueness is assured because each variable is defined and assigned an *"operational definition"*. For example, a variable named "good oral health" might have been defined as "having no existing or recurring caries, no pockets > 3mm, and no calculus on the anterior teeth". Patients who meet those conditions could be scored as a "1" and those who do not could be scored as a "2". Someone else who conducts a different study might define "good oral health" as being something entirely different, depending upon the nature of their research questions and hypotheses. Although operational definitions are specific to a particular study, they sometimes become "standardized" and are used across various studies. A good example here is in the definition that is used to classify people as having HIV mediated AIDS. The Centers for Disease Control definition regards "those infected with HIV who have <200 CD4+ and an opportunistic infection as having AIDS". This criteria (definition) is exported and used for classification in almost all National (American) HIV/AIDS studies.

Generally speaking, there are two types of variables: Independent and dependent.

a. Independent variables – These variables, known as "X", by their nature *cause an effect* that is seen in the dependent variable that is known as "y". Independent variables are used to predict the value of dependent variables. The easiest way to determine which variable in a hypothesis is independent is to determine which variable changes the least (i.e. is being controlled), or can be changed least easily. In the hypothesis "increased *sugar intake increases numbers of carious teeth*". The variable "numbers of carious teeth" is dependent. The change in numbers of carious teeth is the effect that is seen when too much sugar is eaten. In the hypothesis "breast cancer rates are higher in females than in males", gender is the independent variable and breast cancer rates is dependent. The hypothesis says that being female (the cause) predisposes one to breast cancer (the effect).

b. Dependent variables – are variables that show the effect caused by independent variables. They are known as the "y" variable. While the independent variable is "controlled" and changes very little, the dependent variable will in comparison, change a lot- usually to the maximum limits allowed by the study design. In the hypothesis "breast cancer rates are higher in females than in males", having breast cancer depends on whether gender is female.

One way to determine which term is independent or dependent, is to use what I have coined the "dependent" statement. Here is an example using the hypothesis – "flossing eliminates sub-gingival plaque". The dependent statement would be "Having sub-gingival plaque is dependent on whether one flosses". That sentence makes sense. Let's say it the other way and see if it makes sense when the variables are flipped. "Flossing is dependent on whether one has sub-gingival plaque". That makes no sense at all. So it is easy to see that sub-gingival plaque is the dependent variable.

Exercises: Which of the two terms are dependent and independent?

1. IQ and GPA
2. Drill speed and drill type
3. Height and weight
4. Age and Gender

Types of Problem Statements.

Organized research investigations follow the scientific method and are based on defined criteria that allow for objective and definitive outcomes. Sometimes the greatest effort is spent specifying the intent of the research and in ensuring that processes are appropriate to provide unbiased and accurate data. Therefore, all formal research begins with a statement of the problem to be resolved. There are two types of problem statements: *The Research Question and the Hypothesis.*

A. Research Question – Sometimes there is not enough information to define relationships between variables, or sometimes the problem is very complex and cannot be summarized into a simple hypothesis. When this occurs, the research question is the problem statement of choice. Major studies such as NHANES III begin with study groups developing questions to be answered by the research. Each question is narrowly constructed to address only core factors that contribute to part of the problem. Several hypotheses are then drafted to fully answer each research question.

A typical research question might ask: *"What factors contribute to tooth loss in children and are those factors influenced by socioeconomics, ethnicity, and nutrition?"* This is really a complex question because hypotheses would have to be developed determine which "factors" contribute to tooth loss in children. Hypotheses would also be developed to characterize tooth loss factors, in relation to the effects of being in a particular ethnic group, consuming or not consuming proper nutrition, in being in a particular socioeconomic group.

B. Hypothesis – A hypothesis is a statement of the relationship between independent and dependent variables. The statement must also be structured to imply how the relationship should be tested. This will become evident when we get to inferential statistics. There are two types of hypotheses:

1). Null Hypothesis – The null hypothesis is a statement that no difference exists between the independent and dependent variable. It is a statement that no effect-relationship exists between x and y. Null is the German word for zero (0). The symbol for the null hypothesis is H_0 . A typical null hypothesis might state that "prolonged dipping of snuff does not increase the risk of getting oral cancer". Or, we might hear "aspirin therapy does not interfere with Sular absorption in

patients with high blood pressure". Mathematically, the null says that: H_0=0. The scientific method requires that investigators reduce the risk of injecting bias into a study by seeking to "accept" or "reject" the null hypothesis as the outcome of the research.

2). Alternative Hypothesis – The alternative hypothesis is a statement that a difference exists between independent and dependent variables. Sometimes, the alternative hypothesis is called the research hypothesis. It is an observational hypothesis, in that it is a situation that is different from the outcome one would normally expect. The symbol for the research hypothesis is HA, and it is mathematically stated as HA $\neq$0. The relationship between the Null hypothesis and the Alternative hypothesis can be expressed as: H0=0 $<$ HA. A typical alternative hypothesis might state that: "The addition of Lipitor therapy with a diet low in saturated fat reduces cholesterol level".

Students see the alternative hypothesis in clinic when they examine a patient's mouth and find rampant caries. This is clinically abnormal. However, the examination does not reveal the degree to which the finding differs from the rest of the population to which the patient belongs (statistical abnormality). The question to ask then is, "Is this patient unique or do many more in the population have rampant caries?" The difference is determined statistically by sampling the population and testing the null hypothesis. The idea is to either accept the null of no difference (that is, to find that the patient is not unique to the population), or to reject the null (i.e. find that this patient has more or fewer caries than the population). To resolve the issue otherwise, by testing the research hypothesis, would result in the injection of bias into the results.

Scientists who attempt to prove that the research hypothesis is true usually succeed, and discredit themselves when others repeat the study and report different findings.

Alternative Hypotheses can be directional or non-directional.
A *Directional hypothesis* confers magnitude on the effect of the independent variable. A typical directional hypothesis might state: "Potassium increases membrane permeability". *The non-directional* version of this same hypothesis might state: "Potassium effects membrane permeability". Both say that potassium has an effect, but the directional hypothesis makes it clear that the effect is to "increase" membrane permeability.

The Three Forms of Data.

All data can be classified into 3 forms: 1) Nominal; 2) Discrete or Discontinuous; and 3) Continuous.

Nominal data is data in the form of counts. It is often recorded in non-numeric form, where yes/no or +/- or good/bad answers are given. Your name, gender, student ID number are classic examples of nominal data. In its purest sense, nominal data is a label, a unique descriptor. Each category in a nominal variable is unique. In gender, one is male or female, each being mutually exclusive of the other.

Gender Breakout of Database

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	16	45.7	45.7	45.7
	Female	19	54.3	54.3	100.0
	Total	35	100.0	100.0	

Discrete or discontinuous data is numerical data that is in the form of integers and has no intermediate points. It is usually generated by averaging without decimals. It is sometimes generated by counting items.

Descriptive Statistics: Example of Discrete Data

	N	Mean	Std. Deviation
Pre-Treatment AlPase (mg/l)	35	1	0
Post-Treatment AlPase (mg/l)	35	1	0
Socket Perforin (ng/dl)	35	25	15
Socket Granzyme A (ng/dl)	35	25	15
Age in Years	35	30	7
Valid N (listwise)	35		

Continuous data is numerical data that has intermediate points. It can be generated when exact measurements are made. Examples of continuous data include measurements of molar root lengths, determinations of average brushing times among elementary school children, and the amount of anesthesia or anesthetic to administer to patients.

Descriptive Statistics: Example of Continuous Data

	N	Mean	Std. Deviation
Pre-Treatment AlPase (mg/l)	35	.528	.279
Post-Treatment AlPase (mg/l)	35	.610	.268
Socket Perforin (ng/dl)	35	25.393	14.877
Socket Granzyme A (ng/dl)	35	24.552	14.537
Age in Years	35	29.886	7.459
Valid N (listwise)	35		

2. The Ordering of Numerical Data.

All data can be fit to one of four primary numerical scales: 1) **Nominal Scale**; 2) **Ordinal Scale**; 3) **Interval Scale**; and 4) **Ratio Scale**. Numerical scales are rank ordered by hierarchical prominence and are easily remembered as an acronym of the French word for "black", which is "NOIR".

The Nominal Scale. Nominal is the lowest data scale and refers to data that is used mostly as an identifier. Nominal data has no numerical significance other than its value as a descriptive label. A patient's jacket number is nominal data; so is a student's ID number. The fact that Mr. Jone's ID # is 91107 and Miss Thomas' ID# is 91105 does not imply that Mr. Jones is a better student than Miss Thomas.

Patient Has High Blood Pressure: Example of Nominal Data

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	29	82.9	82.9	82.9
	No	6	17.1	17.1	100.0
	Total	35	100.0	100.0	

The Ordinal Scale. Ordinal data is of a higher order than nominal data and refers to data that is ranked by significance. The ranking is qualitative and not quantitative. For example, Kareem graduated Suma Cum Laude and LaShawn graduated Magna Cum Laude. We cannot tell the magnitude of the difference between Kareem and LaShawn. We can only say that one out-performed the other during their academic experience. Data that is scaled on a Likert scale like the one below is said to be ordinal. For example, answers to the question "Were the burrs sharp?" might be:

_____	_____	_____	_____	_____
Very Dull	Dull	Not Dull or Sharp	Sharp	Very Sharp

The fact that "Very Sharp" was checked does not infer a quantifiable degree of sharpness, but only that the burr was qualitatively sharper than burrs that were thought to be "Sharp". Although it is the position of the observation on the scale that is important, means are often calculated for Ordinal data. If "Very Sharp" is coded as a 5 and "Very Dull" is a 1, an average score of 4 would suggest that most subjects thought that the burr was "Sharp".

Periodontal Index: Example of Ordinal Data

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Normal	10	28.6	28.6	28.6
	Mild	6	17.1	17.1	45.7
	Moderate	14	40.0	40.0	85.7
	Severe	5	14.3	14.3	100.0
	Total	35	100.0	100.0	

The Interval Scale. Interval data *has an arbitrary "0"* as a starting point, and is subdivided into divisions that are calibrated in equal units or parts. A Fahrenheit thermometer is an example of an interval scaled device. Because of the arbitrary "0" starting point (water freezes at 32° and not at 0°), there is not a "doubling" effect relationship associated with the units. Therefore, a temperature of 100 is not twice as hot as a temperature of 50 on a Fahrenheit thermometer. A study that evaluates the efficacy of a drug in preventing strokes might record diastolic pressures of patients (since diastolic is more related to strokes and systolic is more related to heart attacks). Diastolic would be on an interval scale, since its arbitrary "0" would be 80 and not absolute "0". Interval data is usually treated as ratio scaled data during statistical analyses, because the units analyzed are continuous or discrete forms.

The Ratio Scale. The ratio scale *has a definite "0"* as a starting point and is subdivided into divisions that are calibrated in equal units or parts.. Ratio scaled data is continuous or discrete. Height and weight are ratio scaled measurements. A person that weighs 250 lbs is twice as heavy as a person weighing 125 lbs, because of the absolute "0" starting point.

Descriptive Statistics: Example of Ratio Data

	N	Mean	Std. Deviation
Pre-Treatment AlPase (mg/l)	35	.53	.28
Post-Treatment AlPase (mg/l)	35	.61	.27
Socket Perforin (ng/dl)	35	25.39	14.88
Socket Granzyme A (ng/dl)	35	24.55	14.54
Age in Years	35	29.89	7.46
Valid N (listwise)	35		

Briefly, Back to Variables. It is interesting to note that variables are sometimes "called" by the type of data or data scale that they represent. A variable such as "age" might be referred to as being a "continuous variable" or a "ratio variable", instead of being an independent or dependent variable. Likewise, gender is often referred to as being a "nominal variable". All are considered correct.

Common Research (Study or Design) Models

1. Quantitative – Analysis of numerical data
2. Qualitative – Analysis of narrative data
3. Interpretive – Analysis of communicative meanings
4. Experimental – Analysis of relationship between x & y variables
5. Naturalistic – Study of normal life behavior
6. Laboratory – Observations in a controlled environment
7. Field Research – Observations in natural social settings.
8. Participant – Researcher contributes a portion of collected data.
9. Non-participant – Researcher is transparent observer.
10. Cross-Sectional – Observations at a single point in time.
11. Longitudinal – Same subjects over a period of time
12. Basic – Analysis of theoretical relationships without concern for practical applications.
13. Applied Research – Analysis of theoretical relationships for the purpose of practical applications and problem solving.
14. Meta-Analysis – Abstraction and analysis of data from records and reports.
15. Cohort Study – Subjects have a common factor
16. Prospective Study – Measuring that which is effective in the future.

Introduction to Probability

The term probability means to prove or to test. It tests the "chance" that an event can occur. The chance of an event occurring is calculated on a ***continuous scale***, with "0" implying that the event will not occur, and with "1" implying that the event will occur. Intermediate decimal points between "0" and "1" give the "odds" or probability that the event will occur. For instance, a probability of .5 means that there is a 50:50 chance that the event will occur. The probability can never be >1 or <0. The probability that an event will occur plus the probability that it cannot equals 1, and can be stated as:

$P\{E_1\} + P\{E_0\} = 1$

1. Single Event Probabilities

Event probability can be calculated as: **P{E}={E}/{TO}**. This equation says that the probability that an event "P(E)" will occur equals the event occurring divided by the total number of outcomes that are possible {TO}. For example, a nickel is a 2-sided coin with one side being a head and the other a tail. There are only 2 possible outcomes when a coin is tossed: a head or a tail. Therefore, the probability of getting a head is (p=.5) or 1 out of 2 tosses. Likewise, the probability of getting a tail is also .5. The probability of getting a head or a tail are "***mutually exclusive***" because you can only get one or the other, never both on a single toss. If you toss a fair coin several times, the probability of getting a head on each toss is still .5. This is because the probability for each toss is "***independent***" of the results of previous tosses.

These concepts also apply when tossing a fair die. A die is a six-sided cube, with each side uniquely numbered between 1 and 6. The probability of obtaining any number between 1 and 6 on a single roll of a die is 1/6 (p=.167).

Probabilities of drawing specific cards from a fair deck of cards is also calculated the same way. There are 52 cards in a fair deck of cards. There are 4 suits (diamonds, hearts, clubs, and spades) of 13 cards (1 ace, 1 jack, 1 queen, 1 king, and 9 cards numbered from 2 to 10). The probability of drawing an ace from a well-shuffled deck is 4/52 (p=.08). The probability of drawing the ace of spades is 1/52 (p=.019), and the probability of drawing any spade is 13/52 (p=.25).

All of the above probability calculations were based on the occurrence of a single event that was independent of any other events that could occur. The deck of cards probabilities as given are true only if drawings are made with "***replacement***". This means that all cards must be returned to the deck before it is shuffled and sampled

again. Drawings that are done "***without replacement***" must allow for the change in the numbers of cards that remain in the deck (which is to say the change in total number of outcomes). For example, the probability of drawing an ace from a deck that has a missing 10 of spades is 4/51 (P=.019), and the probability of drawing a spade from that same deck is 12/51 (p=.231).

Thus far, we have only talked about single event probabilities. These same rules and others apply for multiple events probabilities.

2. Multiple Event Probabilities. The probability that multiple events will occur follow simple rules of addition and multiplication.

A. **The Multiplication Rule**. The multiplication rule says that the probability that two ***independent*** events will occur equals the product of their separate probabilities. Here, the conjunctions ***"and"*** is usually used as part of the syntax that forms the question. The equation is stated as:

*P{(E1)(E2)} = P{E1}*P{E2}*.

For example, what is the probability of throwing a 5 *and* a 6 on a roll of dice.

P{5 and 6} = P{5}*P{6} = 1/6 * 1/6 = 0.028 or 1/36

Here one can expect to roll a 5 followed by a 6 on 1 of 36 attempts using a die, which should happen about 2.8% of the time.

We also could ask, What is the probability of getting 2 heads on two tosses of a fair coin ?

The probabilities of obtaining a head on both tosses is 1/2, since each event will occur independent of previous toss results. Therefore,

P{H and H} = P{H}*P{H} = P{1/2}*P{1/2}=1/4 (p=.25)

So 1 out of 4 times you can expect to toss consecutive heads or consecutive tails.

B. **The Addition Rule**. The addition rule applies to multiple probability events where the conjunctions "either", or "or", are usually used as part of the syntax that forms the question.

1). **Mutually Exclusive Events**. Mutually exclusive events are conditions where only one outcome can occur. The probability of rolling a 4 on a fair die is mutually exclusive of rolling a 3 on the same toss. Another way of saying that events are mutually exclusive is:

$$P\{E1E2\} = 0$$

Therefore, the probability of occurrence equals the sum of the separate probabilities. The equation is stated as:

$$P\{E1+E2\} = P\{E1\} + P\{E2\}$$

For example: What is the probability of rolling a 3 **or** a 6 on a single roll of a die ?

$$P\{E1+E2\} = P\{E1\} + P\{E2\} = 1/6+1/6 = 2/6 = 1/3$$

This should occur about a third of the time.

2). **Non-Mutually Exclusive Events**. Here the probability of occurrence equals the sum of the separate probabilities minus the product of the separate probabilities. Non-mutually exclusive events involve the double counting of possible total outcomes, and is corrected by subtracting the product of independent probabilities (P{E1*E2}) from the mutually exclusive probability (P{E1+E2}). This can be stated as:

Mutually Exclusive Product of Prob.

$$P\{E1+E2\} = (P\{E1\} + P\{E2\}) - P\{E1*E2\}$$

For example, we could ask the question: What is the probability of drawing a 10 **or** a diamond on a single draw ? In this instance, a 10 of diamonds could be drawn. Therefore, the events are not mutually exclusive. Therefore, the probability is:

$$P\{4/52 + 13/52\}- (P\{4/52\} * P\{13/52\}) = 17/52 - 1/52 = 16/52 = .30769$$

C. **Conditional Probability**. Conditional probabilities are probabilities that are calculated from intermediate points in the data. These events occur in a specific order. For example, if we asked: "What is the probability that the 1st person walking through the door will be a male and the 2nd person a female?" The event is conditional at the start only if the 1st person through the door is a male. Conditional probability is stated as:

$$P\{B \mid A\} = P\{A \text{ and } B\}/P\{A\}$$

This equation says *the probability of "B" occurring, given that "A" has occurred, equals the product of the probabilities of "A" and "B", divided by the probability of "A".*

Conditional probability events have occurred independently of each other when:

$$P\{A|B\} = P\{A\} \text{ and } P\{B|A\} = P\{B\}$$

Example
Now, let's consider examples of how this can be applied using the data below.:

Gender vs Medical Complications Crosstabulation

Count

		Medical Complications		
		Yes	No	Total
Gender	Male	7	9	16
	Female	12	7	19
Total		19	16	35

1. The probability of a randomly selected individual being a male with complications = 7/35 or .20
2. The probability of randomly selecting a male = 16/35 or .457
3. The probability of randomly selecting an individual with complications = 19/35 or .5429
4. The probability of a selected male having complications is 7/16 or .4375
5. The probability that an individual selected is female and has no complications = 7/16=.4375

Let's solve problem 5 using the conditional probability equation stated above:
$P\{B \mid A\} = P\{A \text{ and } B\}/P\{A\}$
$P(7/16 \mid 9/35) = P(19/35)*P(7/16)/P(19/35)$
$P(.4375 \mid .5429)=P(.5429*.4375)/.5429$
$= .4375$

Populations and Samples

A perfect statistical study would be one that includes every item in existence that is relevant to the object of the study. However, this can usually be done only when the total study ***population or universe*** is very small. For example, if salivary calcium levels were studied in children in the Howard University Day Care Center, and a saliva sample was collected from every child in the center, the entire ***population or universe*** of the center would have been sampled. In contrast, a study on salivary calcium levels in American pre-school children would find it impossible to obtain a specimen from every pre-school child in the country. In this case, a ***sample*** of pre-school children from across the United States would be selected to study. It is important that the composition of the sample reflect the make-up (demographics, socioeconomics, cultural, racial, etc..) of the population from which it is taken. In each of the above saliva studies, each child in the sample would be an ***Observation or Case***. The data that will be collected, such as the age of the child, calcium:phosphate ratios, unbound calcium levels, alkaline phosphatase levels, tartrate-resistant acid phosphatase levels, calcitonin levels, and unbound phosphate levels, are ***variables*** that will be analyzed for significant relationships, similarities, and differences.

It is crucial that ***bias*** (*subjective influences that skew data to a desired outcome)* be controlled or excluded from the study so that the ***sample*** is a true representation of its ***universe***. An **Unbiased** situation is one in which an observation is selected through a **random and independent process**. **Random** selection means that every observation in the population has an equal chance of being selected for the study. **Independent** means that the selection of a particular observation will not influence the possibility of some other observation being selected. The how and where data are collected, therefore, greatly effects whether sampling will be biased or not.

A classic example of statistical bias would be to sample only geriatric diabetics for a study on serum glucose levels in the elderly. One would falsely conclude that all elderly are diabetics.

Why Sample

Several factors make it impossible to include entire large populations in some studies:

1. The investigator might not have sufficient funding to include everyone.
2. Time restraints, such as funding cycles for persons conducting the study might limit them to a few months in which to conduct the study.
3. All subjects might not be available or willing to participate in the study.

4. Populations of people are never static. They come and go, removing themselves from the sampling frame. They get older and no longer meet the inclusion criteria for studies

Two Sampling Approaches

There are two common approaches to sampling: non-probability and probability.

*In the **Non-probability** approach*, where there is less concern about fitting data to the population. I call this approach the "Let's get an idea what's out there" approach. A sample of "X" size is drawn and the data are explored. Since this approach is a bit non-scientific, truly accurate conclusions cannot be drawn, although the conclusions could suggest the need for a more scientific study.

Models Associated with Non-Probability Sampling –

1. Purposive Sampling – This is usually a case study of a target group that has a particular trait. It could be a study on the quality of care received by patients on medicaid, or perhaps it could be caries in geriatric African American men. The sample is limited to those with the particular trait of interest, irrespective of the total number that might be in the population.

2. Convenience Sampling – We all have experienced this kind of sampling. Remember walking through the mall and having someone shove a survey in your face and asking you to please fill it out? You were conveniently there and they conveniently sampled you. This type of sampling is usually performed during the busiest times of day and seasons of the year. The research group administering the survey needs numbers - that is lots of people to fill out the survey. Later, the data will be correlated with demographics and preferences on the survey. They will also assign quasi-probability validity to the survey by estimating the approximate foot traffic through the mall during sampling times, and then calculate the amount of error associated with the responses.

3. Quota Sampling – This is a non-scientific approach to stratification and clustering. All subclasses in the population are sampled at ratios equivalent to their census in the population. For example, a convenience survey is designed to look at what people buy from a particular store during the Christmas holidays. The store estimates that it's 5,000 clients are primarily 40% White, 30% Black, 20% Hispanic, and 10% Asian (Please no offense to anyone else). Surveyers will survey specific ethnic groups until those percentages are reached in the sample. Have you ever been in a mall and had a someone walk, rush, or run past 20 people to hand you a survey? Chances are that you were being quota sampled.

*The other sampling approach is the **probability** approach,* which requires that a mathematical algorithm be used to determine the number of subjects to be studied, if accurate conclusions are be made from the data. The cornerstone of the probability approach is a regard to have randomness and independence in selecting the observations to be studied.

Probability Models –

1. Simple Random Sampling (SRS) – This is the simplest approach to probability sampling. The minimum sample needed is calculated and a random numbers table can be used to decide who gets selected (e.g. using jacket number, SSN, Date of Birth, etc.).

2. Stratified Sampling – This method ensures accurate representation within a stratifying variable. The investigator needs forehand knowledge about the population to be sampled. To sample a stratified variable such as ethnicity, all ethnic groups in the population would have to be included. *Proportionate allocation* is the term that describes a sample that is stratified according the percentage make-up of the variable in the population.

3. Cluster Sampling – This approach is a form of stratified sampling, but it concerned with specific strata within a population. A study using this method might only look at one gender, or patients who use a specific plan to pay for their dental treatment.

Note: Sometimes clustering and stratification were not included in the original study design, but occurs after the data have been collected. This method of stratification is called *"inverted sampling"*.

Terms that are Associated with Probability and Non-Probability Sampling

1. Sampling Unit – Definitive object being studied, e.g. Teenagers.
2. Observation Unit – Object or source (Experimental object) to be studied (e.g. teen drinkers vs non-drinking controls)
3. Target Population – All available sampling units that can be included in the study.
4. Sampling Frame – List (names, phone #, etc..) of all sampling units in the population.
5. Sampling error – The extent to which the sample does not represent the population. Usually given as a range, e.g. ±N.

Calculating Sample size

Deciding how many subjects to sample from a population can be quite complex, especially if the sampling algorithm requires stratified or multistage sampling. The approach is different if the sample is based on mean-difference as opposed to selecting a proportion-based sample. However, there are basics that should be discussed. First of all, there are three things that commonly need to be known to calculate a sample size:

1. *The amount of error that is acceptable in the sample*. Error in the form of Margin of error is the difference between what was sampled and what can be expected upon repeated sampling of the population
2. *The size and characteristics of the population*
3. *The magnitude of difference that is to be detected.* Sometimes, the desire is to know if a change is of a desired magnitude. If a drug is needed to increase red blood cell numbers by 800 in sickle cell transgenic mice, how many mice would be needed to detect the increase.

Polling, Sampling and Calculations

The media always gives poll results with the error that is associated with the poll numbers. The error associated with sampling can be calculated using a simple equation:

Margin of Error = $1/\sqrt{N}$. If 1000 people are sampled, the Margin of Error = $1/\sqrt{1000} = .0316$ That is 3%. The size of the population does not matter. *If you randomly and independently select 1000 subjects for the survey, the error upon repeated re-sampling will be 3%*. The equation can be refined by replacing the reciprocal with the value for the confidence level. 99% = 1.29, 95% = .98, and 90% = .82.

Simplified Approaches to Calculating Sample Size

There are numerous equations to calculate sample size, such as the one below.

$$D = \frac{b^2}{4} = \frac{.05^2}{4} = .000625 \qquad n = \frac{Npq}{(N-1)D + pq} = 369.80 = 370 \; records$$

This common approach was used to calculate sample size for records review in the College of Dentistry. There were at the time, 4887 active patient charts in the College, and they were located in the clinics below. Proportional Allocation was used to determine how many records from each section should be reviewed:

Table Sampling Algorithm for Records Review

Clinic	# Stored Records	% of All Records	**# Records to Sample**
AGD	143	2.93%	**11**
CIAP/DAU	1305	26.70%	**99**
RECORDS ROOM	2089	42.75%	**158**
DENTAL HYGIENE	1350	27.62%	**102**
Total:	4887	100.00%	**370**

If the population is small and known, the sample size can be reduced with precision using the Finite Population Correction Factor. The equation for the FPC is: $fpc = \sqrt{N - n / N - 1}$ Where N = size of population, and n = size of sample.

Therefore, the final sample size for the records review was 343 records.

Detecting Magnitude of Difference

If a change of a certain magnitude needs to be detected, the equation $n = \frac{16\delta^2}{\Delta^2} + 1$ can be used. Suppose that you need to detect if drug a reduces systolic blood pressure in cats by 10mm/hg with a standard deviation of 3mm/hg. How many animals would you need to detect the 10mm/hg change?

$$n = \frac{16(3)^2}{10^2} + 1 = 2.44 \; or \; 3 \; cats$$

The exercises above are simplified but accurate ways sample sizes can be calculated. It becomes quite complex when multi-stage, repeated sampling is performed.

