

# Chapter 1

## The Language of Statistics

### CONTENTS

1.1 Introduction .....	1
1.2 How Statistics Can Work For You: A Brief Example .....	2
1.3 Structure of Statistical Analysis .....	5
1.4 Understanding Variability .....	6
1.5 Population Versus Sample: The Difference Is Huge! .....	7
1.6 How to Sample: Avoiding Bias and Mitigating Confounding .....	8
1.6.1 Sampling to Avoid Issues of Bias.....	9
1.6.2 Issues of Confounding.....	10
1.6.3 Self-Selected Samples and Medicine .....	12
1.6.4 The Role of Sample Size .....	13
1.6.5 Inclusion and Exclusion Criteria.....	13
1.7 Levels of Measurement .....	14
1.7.1 Standard Levels of Measurement .....	15
1.7.2 Likert Scales: A Special Type of Data.....	17
1.8 Example .....	18
1.9 Terminology in Clinical Trials.....	20
1.10 Case Study.....	22
1.10.1 Primary Research Goals .....	23
1.10.2 Study Variables.....	24
1.10.3 Confounding Variables .....	24
1.11 Chapter Summary.....	25
1.12 Exercises .....	26

## 1.1 INTRODUCTION

**Evidence-based practice**<sup>1</sup> is a common buzz-word in healthcare research. McKibbin, in 1998, defined evidence-based practice as “an approach to healthcare wherein health professionals use the best evidence possible...to make clinical decisions for individual patients.<sup>2</sup> At the heart of evidence-based practice is the field of **statistical science**. Why? Because people vary (or differ) in many ways. In the practice of medicine, variability among patients means that one generally cannot know for certain whether (or how well) a given treatment will work for a specific person. While a selected treatment may not fully cure a disease, at the very least it should be chosen based on the likelihood for improving the quality of life for the patient while also minimizing side effects.

This idea of **variability** is the foundation for the field of statistics. Perhaps the best medical definition of variability lies in the idea that different patients will react differently to a given treatment. Outcomes may range from complete cure to no change to side effects that could result in great harm to the patient. Understanding this idea of variability is key to recognizing and endorsing the need for proper statistical methods and experimental design when evaluating treatments. Because one can expect that patients will vary in their response to a treatment, it is also possible to design methodology to analyze variation in ways that will separate scientific evidence of a treatment effect from the expected random noise.

**Statistics** as a scientific field leverages an understanding of variability to make reasonable guesses about the impact of treatment at two different levels. Ideally, impact will be assessed for a **population** (a group much larger than the **sample** of individuals for whom you actually have data). The result is a generalization about the benefits (and side effects) a treatment would be likely to have on a population as a whole. Second, and perhaps of greater importance in medicine, impact may be assessed at the individual patient level. **Models** derived from data may help us understand what to expect when a treatment is applied to an individual having certain characteristics. Individual impact is often expressed in terms of probability, or chance, that a treatment will be effective based on an individual’s characteristics. Statistical models can help determine which individuals are best suited to receive a given treatment. They can also aid in choosing appropriate amounts of treatment.

There are generally two areas of published healthcare research. **Quantitative research** involves the collection of measurements, usually for several different variables of interest. The search for relationships among these variables should almost always involve **statistical inference**. On the other hand, **qualitative research** involves the collection of written responses to various questions. Generally, statistical methods are not all that applicable to qualitative research. **But ideas related to bias, as well as an understanding of sampling variability, absolutely apply even to qualitative data.** It is imperative to recognize that responses to a given question will vary, and therefore no single individual response should receive too great a focus. Conversely, if many

---

<sup>1</sup> Nolan, P., & Bradley, E. (2008). Evidence-based practice: implications and concerns. *Journal of Nursing Management*, 16(4), 388-393.

<sup>2</sup> McKibbin, K.A. (1998). Evidence-based practice. *Bulletin of the Medical Library Association*, 86(3), 396-401.

participants in a qualitative study give similar responses, then greater focus on those responses as a potentially important result is likely appropriate.

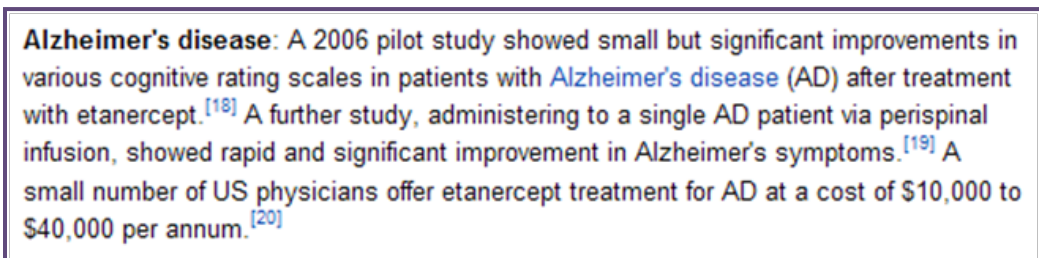
This book will focus primarily on quantitative research. **Statistical inference** relies on a probabilistic assessment of variability and proper incorporation of that assessment into the scientific decisionmaking process. Ideally at this point you are convinced that a basic understanding of statistical inference is necessary in order for healthcare professionals to be able to read, understand, and critique available research with the ultimate goal of choosing the most effective (or most likely to be effective) treatment for their patients. It is also of even greater importance for those who plan to be directly involved in research and ultimately add their own publications to the available body of evidence on a particular medical topic.

## 1.2 HOW STATISTICS CAN WORK FOR YOU: A BRIEF EXAMPLE

Let's begin with an example to illustrate several of these ideas. Alzheimer's disease is common in the elderly population and its symptoms typically become debilitating. Etanercept, an anti-inflammatory drug, has been proposed as an effective treatment for this disease. Imagine that your hospital is considering whether to participate in ongoing clinical trials related to this treatment. You are part of a research group that will be making a decision on this matter.

Where should you start? Continued growth of the internet makes access to information easier than ever before. In 2012, a Google search of "Etanercept AND Alzheimer's Disease" produced around 125,000 results. This may be helpful but there are two important issues to consider. One problem here is that there are too many results to effectively look at them all. Perhaps more importantly one must also remember that people can literally post *anything* on the internet. Even if all 125,000 results could be easily examined, it could be difficult to determine which are most reliable.

However, examining a Wikipedia page concerning Etanercept (found within the search) may prove useful: <http://en.wikipedia.org/wiki/Etanercept>. Of course the Wiki may be edited by anyone – hence the content of the Wiki (excerpt related to Alzheimer's shown in the box below) may not necessarily be trustworthy. But it can (and in 2012 did) provide a reference list that contained several items that may prove to be useful.



**Alzheimer's disease:** A 2006 pilot study showed small but significant improvements in various cognitive rating scales in patients with Alzheimer's disease (AD) after treatment with etanercept.<sup>[18]</sup> A further study, administering to a single AD patient via perispinal infusion, showed rapid and significant improvement in Alzheimer's symptoms.<sup>[19]</sup> A small number of US physicians offer etanercept treatment for AD at a cost of \$10,000 to \$40,000 per annum.<sup>[20]</sup>

Figure 1.1. Screenshot excerpt from <http://en.wikipedia.org/wiki/Etanercept> taken June 2012.

The references point us to a team of researchers led by Edward Tobinick as well as an expository article by Daniel Elkan. It is now time to for a more rigorous check of journal sources using a

library. A search of the *Academic Search Premier*<sup>3</sup> database provides four articles from Tobinick's team (2007, 2008, 2009, and 2012).<sup>4,5,6,7</sup> Also available is Elkan's 2008 article which presents the ideas from both sides but is substantially lacking in any scientific application.<sup>8</sup>

Consider some of the findings in detail. Elkan's article presents the story of Walter Skotchdopole – a man who showed what Elkan terms a “miraculous recovery” after receiving the treatment developed by Tobinick. How should you interpret this knowledge in line with the proper use of statistics and evidence-based practice? Let's start with what is known. A 79-year old male had been diagnosed with Alzheimer's and was unable to maintain his own care. After treatment with Etanercept, he substantially regained the ability to live independently. Is this proof that the Etanercept is an effective treatment for Alzheimer's disease? **Absolutely not!**

### IMPORTANT TRUTHS ABOUT SAMPLE SIZE

**Key point #1: In the above example, Skotchdopole represents a sample of size ONE. No statistical inference about the rest of the population (in this case those people with Alzheimer's disease) should ever be drawn on the basis of only one data point.**

**Key point #2: This line of thinking leads to the question “How many data values *are* needed? The answer to that question is quite complicated and depends on several aspects of statistical design. But the required sample size is almost always more than 10 and quite often the requirements may be in the 100s.**

**Key point #3: A common misconception is that the required sample size relates to the size of the population and that for larger populations, more sampling is required. That simply isn't true, and seldom is there truly a statistical need for sample sizes in the tens of thousands.**

Certainly it is unacceptable to imply that, because the treatment worked in one person that it would be as effective in every Alzheimer's patient. Remember, patients will vary! Of scientific importance is whether a trend exists. One patient alone can never be used as evidence of a statistical trend. Further, it is also not appropriate to imply that Etanercept caused improvement

<sup>3</sup> EBSCO Industries, Inc. Academic Search Premier. [Software] Online Reference: <http://www.ebscohost.com/academic/academic-search-premier>.

<sup>4</sup> Tobinick, E. (2007). Perispinal etanercept for treatment of Alzheimer's disease. *Current Alzheimer Research*, 4(5), 550-552.

<sup>5</sup> Tobinick, E. L., & Gross, H. (2008). Rapid improvement in verbal fluency and aphasia following perispinal etanercept in Alzheimer's disease. *BMC Neurology*, 8, 1-9.

<sup>6</sup> Tobinick, E. (2009). Tumour necrosis factor modulation for treatment of Alzheimer's disease. *CNS Drugs*, 23(9), 713-725.

<sup>7</sup> Tobinick, E. (2012). Deciphering the physiology underlying the rapid clinical effects of perispinal etanercept in Alzheimer's disease. *Current Alzheimer Research*, 9(1), 99-109.

<sup>8</sup> Elkan, D. (2008). Awakenings. *New Scientist*, 198(2668), 32-35.

in this particular patient. There may have been something else – something completely unrelated to the medication – that created the improvement in Skotchdopole’s condition. For these reasons, Elkan’s article provides no real convincing evidence. But what about Tobinick’s published research? Does it provide evidence that Etanercept is an effective treatment for Alzheimer’s disease?

**\*\*\*\*Before reading on, consider checking out the referenced articles for yourself!\*\*\*\***

The answer: while one might certainly want to believe that Etanercept is a cure, the evidence at best only mildly supports this possibility. There are several facts that should create skepticism:

- All current research appears to be from the same author. While specialization can be a good thing – research is more convincing when multiple sources produce similar results.
- Sample sizes are small and the design was not **randomized**. Further, while in some cases the author finds **statistical significance**, he fails to address **clinical relevance**. Not only is it important to assess whether or not the treatment results in symptomatic improvement; but the magnitude of improvement must be great enough to outweigh the costs (including side-effects) involved in treatment.

The second point here is of great importance. To put it into perspective, if the statistically significant improvement were that patients would be cognizant for an additional average of around two minutes per day for the next year, most reasonable people probably would conclude that the outcome would not justify the multi-thousand dollar yearly cost.

**Key point: When presented with “statistically significant” improvement, one must always still ask the question – is the amount of improvement large enough to justify the costs? In statistics, when possible the idea of “amount” is best investigated using a **confidence interval**. For this reason, confidence intervals are a primary focus of this text.**

**Some Updates:** Amgen, the company that produces Enbrel® (etanercept), posted a statement regarding the use of this medication in Alzheimer’s patients. An archived version of that statement can be found here: <https://www.alzconnected.org/archive.aspx?g=posts&t=18669>. It confirms that our concerns are well founded, particularly in that it comes from a company that has a financial interest in promotion, not opposition, of this medication. Since that statement, there has been further study including at least one clinical trial.<sup>9</sup> Additional information about that can be found at <http://www.alzforum.org/therapeutics/etanercept>.

---

<sup>9</sup> Butchart, J., Brook, L., Hopkins, V., Teeling, J., Puntener, U., Culliford, D., Sharples, R., Sharif, S., McFarlane, B., Raybould, R., Thomas, R., Passmore, P., Perry, V., & Holmes, C. (2015). Etanercept in Alzheimer disease: A randomized, placebo-controlled, double-blind, phase 2 trial. *Neurology*, 84(21):2161-8.

### 1.3 STRUCTURE OF STATISTICAL ANALYSIS

Our example involves numerous statistical ideas that must be considered if we are to fully understand the results. In this first chapter, we will investigate several of these concepts, all of which are part of the following general structure of statistical analysis illustrated in Table 1.1. Most if not all of these ideas you may have encountered in an undergraduate statistics course but perhaps do not yet fully understand. A deeper understanding of these concepts will enable you to develop your “Statistical BS” radar as you examine research conducted in the field of healthcare.

Table 1.1 Basic structure of a quantitative research study.

Stage	Item	Description
1	Develop Research Questions	In general, a statistical analysis begins with the construct of one or more <b>research questions</b> . This also should involve careful definition of the <b>population</b> , <b>variables</b> , and <b>parameters</b> of interest to the researcher.
2	Determine an Appropriate Sample	The <b>sample</b> will produce the <b>data</b> to answer the research questions. The key word here is “appropriate”. In collecting the sample we must work to avoid <b>sampling bias</b> as well as to minimize issues with <b>confounding</b> variables which may mask true relationships.
3	Collect Data	Variables involved in research questions must be measured for each <b>experimental unit</b> (person) in the sample. There are three standard <b>levels of measurement</b> : <b>nominal</b> , <b>ordinal</b> , and <b>interval/ratio</b> . Additionally, <b>Likert scales</b> <sup>10</sup> represent a fourth type of measurement that is important in healthcare research. Understanding and identifying the <b>datatype</b> is paramount to the selection of the appropriate statistical method.
4	Perform Analyses	Analysis of data generally involves the use of statistical software. The key job of the analyst is to ensure that correct methodologies are used to produce the results. Much of this textbook is devoted to choosing the proper methods; the choice of method is generally based on datatype.
5	Draw Conclusions	Conclusions may be statistical or practical in nature. It is important to understand the difference between <b>statistical significance</b> and <b>clinical importance</b> . The strongest, most useful conclusions will be those that are both statistically significant <u>and</u> clinically important.

<sup>10</sup> Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140: 1-55.



By the end of the first two chapters, your goal should be to have attained a solid working understanding of the terminology and concepts important to a statistical investigation. You'll also note that very little is said of the mathematics in this textbook. This should not be taken to imply that the mathematics underlying statistical methods are unimportant; it is simply the case that for our purposes the computer will handle their application. Our job as consumers of statistics is to ensure that correct procedures are applied, reasonable interpretations are given, and reasonable decisions are made based on those interpretations. The remainder of Chapter 1 will review important terminology; Chapter 2 will review the use of descriptive methods in assessing the viability of a sample. We will also consider ways in which descriptive statistics should not be used (and are too often abused within the literature). The remainder of the text is devoted to the application of correct inferential methods (confidence intervals and hypothesis testing) to yield valid statistical conclusions that form the basis for sound practical decisions.

## 1.4 UNDERSTANDING VARIABILITY

Before continuing with some of the key terminology involved in data collection, we will first address three important questions that provide the foundation for statistical understanding:

1. What is variability?
2. Where is variability found?
3. How does variability help us make decisions?

The word variability stems from **variable**. In scientific practice, often we collect data for several variables. At the base level, these **data** are collected on **experimental units** (defined to be the people or objects on which we might collect data). In medical practice, the experimental units available for data collection are most often patients. The data we collect for different patients almost certainly will vary. For example, some patients will be men and others women. Some will have higher blood pressure than others. Breathing rates will differ. Daily food intake will differ. Daily exercise will differ. All of these differences between experimental units represent **variability**. Variability is found everywhere!

**Consider the following challenge:** Find an attribute of human beings for which everyone is the same. To do this it would seem necessary to zoom all the way out to "All human beings are born on the earth." Even this is not necessarily a scientifically proven fact.

Variability is everywhere. But how does it help us make decisions? As an example, let's take a relatively simple research question: suppose that we wish to check a certain brand of aspirin to determine whether the tablets actually contain 200mg as indicated on the bottle. What are the expectations going into the study? In particular:

1. Do you expect every aspirin tablet in every bottle to contain exactly 200mg?
2. Do you expect any aspirin tablets from the bottle to contain 100mg? 300mg?

If you answered “no” to both of these questions – then you have already developed a reasonable intuitive understanding of variability. First, aspirin will be mixed in large batches before it is formed into tablets. There will be batch-to-batch variability as well as tablet-to-tablet variability within each batch. Mixing machines do not exist that would completely eliminate these two components of variability. But they should minimize them to a large extent, because dosages as low as 100 might not be enough to relieve symptoms while dosages as high as 300 might lead to unwanted side effects. Thus such a large amount of variability cannot be accepted. Here is where statistics and medicine mix. Based on knowledge of what we want to happen medically, a statistician would ask the following two questions about aspirin:

1. Is the **average** amount of aspirin per tablet different from 200mg?
2. Is the variability in amount reasonably small? We might look at this in terms of **standard deviation**. Perhaps we would like to determine whether the standard deviation is less than 2 mg. Or we might simply decide we want nearly all tablets to be between 194 mg and 206 mg (you might recall that nearly all data will fall within three standard deviations of the mean for any specified population).

These questions would be answered using statistical methods. We would use a one-sample T-confidence interval to get an idea of the average amount of aspirin (alternatively we can use a hypothesis test, but a confidence interval will provide more detailed information). Chi-square methods might be used to estimate the variability. In later chapters, we will learn more about how to perform the actual analyses using a computer.

## 1.5 POPULATION VERSUS SAMPLE: THE DIFFERENCE IS HUGE!

When we conduct a statistical analysis we are attempting to make inference related to a **population**. The simplest definition for population is that it consists of all people (or objects) that would be of interest. In the two examples discussed so far, the populations are:

- All persons having Alzheimer’s disease.
- All aspirin tablets of the brand in question.

Our goal in an analysis is to draw scientific conclusions that apply to the entire population. But in most cases (including those above), it will not be feasible to collect data for every person (or item) in the population. In the first example, simply finding everyone who had Alzheimer’s would be an impossible task. In the second example, measuring the amount of aspirin in every tablet would be an arduous process that ultimately would not leave any tablets remaining for people to use. So while we might like to take a **census** (collecting data for every experimental unit in our population) we will generally be forced to settle for a sample. A **sample** is quite simply a subset of the population. There are many different types of samples; but the most important thing to understand about samples is that they will exhibit variability. While the general procedure will be to collect and analyze a single sample, if we did collect several samples (each consisting of a different subset of the population), we would find that they differ in regard to the collected data and in various **estimates** calculated from that data.



**Key point: Different samples will result in different data; different estimates; and in some cases different statistical conclusions!**

That's right! If you and I both collect samples of aspirin, your sample might estimate that the average amount of aspirin is between 199 mg and 203 mg (showing no evidence that it is different from 200 mg). At the same time my sample might estimate that the average amount of aspirin is between 201 mg and 204 mg (evidence that the tablets contain a slightly higher average than expected). But wait – can it be both? Certainly not! There is only one true population average. The fact is that while we both used correct statistical methods, one of us was unlucky in our sampling and therefore did not attain a correct conclusion. Why does this happen? Back to our key word: variability! Variability is inherent to sampling – there is no way to get rid of it. So right from the start it is important to understand:

**Key point: Statistical methods by which we use samples to draw inference about populations are by their nature imperfect. Using them will draw correct conclusions a large percentage of the time – but an error is always possible. Since we cannot eliminate it, our goal will be to minimize the chance of error (discussed further in Chapter 4).**

## 1.6 HOW TO SAMPLE: AVOIDING BIAS AND MITIGATING CONFOUNDING

Most statistics textbooks will tell you that the best type of sample is a **random sample**. They give many reasons for this (including that random samples will be unbiased), however none of those reasons matter too much because if you think about random samples for just a moment – you should recognize a fact that most textbooks won't discuss: random samples are generally impossible to obtain!

**Key point: The goal in practice is to get a sample that is as reasonably likely to be representative of the population as possible.**

It turns out that there are several things that we can do regarding representation. Let's begin thinking about this by considering two key sampling mechanisms that we will encounter quite often in practice:

1. **Convenience samples** are samples in which the experimenter picks or decides who/what experimental units will participate in the sample. Perhaps the best type of convenience sample is **systematic** (e.g. taking every 5<sup>th</sup> patient for the sample).
2. **Self-selected samples** are samples in which the experimenter may advertise for participants, but ultimately the participants themselves must volunteer.

At this point you should understand that we have to collect a sample which will provide a basis for drawing conclusions about our population. Furthermore, we have no choice but to use imperfect methods in collecting that sample. Since samples will generally be one of the two types described above, the important question we must consider is how we might ensure that a sample is collected in such a way that it is reasonably likely to provide fair representation for the population. For research conclusions to be reliable, the chosen sample should avoid two key issues: **bias** and **confounding**. Another way to look at this is from the perspective of the following question: **How broadly can we generalize the results observed in our sample?**

### 1.6.1 SAMPLING TO AVOID ISSUES OF BIAS

As mentioned in Section 1.5, all samples involve some amount of random variability. **Sampling bias** occurs when a sample, because of how it was chosen, contains systematic variation that makes it more likely (as compared to what might normally happen by chance) to misrepresent the population with respect to the variables under investigation. As an example, suppose one wanted to look at development for children under the age of one. To select a sample that consisted of children who were all two to three months premature at birth would surely result in a sampling bias. Because premature birthed children take up to two years to “catch up”, this sample would surely underestimate development for all children in the population of interest. If it seemed obvious to you that this was an example of a bad sample, that’s good!

Another type of sampling bias may occur due to **non-response** (and/or self-selection). Consider an exercise and weight loss study in which patients must come in to be weighed at completion of the study. Patients who have not lost much weight may be embarrassed by that fact and fail to show up to complete the study. If this occurs, the exclusion of such data would surely lead to over-estimation of average weight loss in the population.

**Key point: Sampling bias affects accuracy. If sampling bias is present, the sample is less likely to accurately reflect the truth about the population – both in estimation for and assessment of association between study variables.**

Results from biased samples are often misleading and generally should not be used. Hence the question becomes: how do we avoid bias? Fortunately, there are techniques that may help:

- **Convenience samples** are preferred (but not always ethically possible) because they allow the experimenter some latitude in attempting to avoid known sources of bias. Typically, **inclusion/exclusion criteria** are designed for this purpose.
- **Systematic samples** can be helpful in cases where the researcher choice of which patients to ask for participation in a study might result in selection bias.
- **Non-response rates** may be minimized by providing a reward for participation. One must be careful, however, that the promise of reward would not affect measurement.

As an example, suppose the variable of interest is time spent exercising. An experimenter might select a convenience sample from among the co-workers at the hospital at which she works. This would be much less likely to result in bias than if she chose her sample from the local health-club she visits twice a week; the latter sample would surely result in overestimating the average time spent exercising. Self-selection would be a poor choice here as well, since it would be possible – perhaps even probable – that interested participants in the survey would only represent those who appreciate the value of exercise. If non-response is deemed an issue in this example, the researcher might choose to offer a reward for completion of the survey. She must be careful that the reward would not have any relationship to the variables under study. For example, food would be an inappropriate reward for this study – can you see why?

---

### 1.6.2 ISSUES OF CONFOUNDING

**Confounding** is quite similar to bias and in fact the two concepts are often confused. While bias involves systematic errors that are a direct result of the method of the data collection plan, confounding involves related, uncontrolled (and perhaps uncontrollable) **nuisance variables** in relation to **group assignment**. In particular, confounding occurs when groups to be compared are by their nature dissimilar with respect to characteristics that are also related to the primary outcome under study.

**Key point: To help distinguish bias from confounding, remember that bias is a global issue of the entire sample while confounding is an issue specific to group assignment.**

As an example, perhaps we want to compare skills development for children at two different locations: St. Louis, MO and Cité Soleil, Haiti. In both places, children are selected by sampling birth records from various countries and skills development is measured 3 years after birth. Perhaps data are collected and appropriate statistical methods indicate numerous differences in development which researchers then attribute to location. The problem with this conclusion is that they have ignored the fact St. Louis and Cité Soleil differ greatly in terms of economic conditions. In fact Cité Soleil is described by the International Committee of the Red Cross as a place of extreme poverty.<sup>11</sup> Economics is almost certainly a confounder to location in this case, as development would be logically expected to have far greater connection to poverty level than location. Even a similar study to compare two locations within the United States could have problems. Suppose that the locations are Detroit, MI and Portland, ME. Could differences in skills development be reasonably attributed to location? Surely any such differences are more likely attributable to pollution levels (or some other specific characteristics of the two regions) rather than actual location on a map.

---

<sup>11</sup> Revol, D. (2006). Hoping for change in Haiti's Cité Soleil. International Red Cross. Online Reference: [http://www.redcross.int/EN/mag/magazine2006\\_2/10-11.html](http://www.redcross.int/EN/mag/magazine2006_2/10-11.html).

Known confounders can sometimes be mitigated using a technique called **stratification**. Stratification in sampling ensures that a known confounder will be equally distributed across the study groups. In the case of the baby development example above, suppose that we want to remove economics as a potential confounder from the study in comparing the cities of Detroit and Portland. To do this, we might ensure that our samples each contain the same number of children at each of five different household income levels (where these levels are probably also adjusted for cost-of-living differences). This similarity of samples would ensure that income level did not have the ability to affect comparative estimates for development. As an additional benefit, stratification can also be used to attain a **balanced design** (a desired property any time several groups are to be compared).

Of course there is always the risk that a comparison will be affected by a confounding variable that is not easily foreseeable. For **observational studies** in which participants fall into groups based on observed data, there is little to be done about this issue. Sometimes it is possible to develop a statistical model (e.g. Analysis of Covariance) that may take possible confounders into account. For studies having careful **experimental design**, in which participants are assigned to their treatment groups by the experimenter, risk of issues related to unknown confounders are best minimized (but not necessarily eliminated) using **randomization**.

**A Note on Causation:** Experimental designs are absolutely necessary if inferential results are to assess **causation**. For example, it may be possible to show that aspirin reduces the risk of heart attacks; but it is not possible to show that increased risk of heart attacks is triggered by being overweight. The latter study would be observational in nature, since a person's weight cannot be randomly assigned.

Randomization means that the assignment of participants to treatments is random, and is a typical component of most medical studies. To understand how it works, consider a researcher working to test a new program to see if the program increased time spent exercising. To do this well, she now needs two groups of participants (one group will participate in the new program while the other will not). The assignment of subjects to groups should be randomized. To understand why, consider that bias is quite possible here if the experimenter selects the groups herself. As a proponent of exercise, she might (even subconsciously) consider the "build" of her subjects when assigning them to groups. This may result in the assignment of the most healthy individuals to the control group (perhaps she views them as likely already exercising well and not in need of treatment); while likewise the least healthy individuals are assigned to the treatment (because they are viewed as needing it). These assignments would almost certainly render the results of the study to be useless as the more healthy individuals likely have substantially less room for improvement.

**Blinding** is another aspect of experimental design that can help to avoid investigators or participants becoming confounders in their own study. Most medical studies will be **double blind**, meaning that neither the investigator nor the participants know whether a specific participant is getting the study drug or a placebo. It isn't difficult to see how such knowledge

might itself impact a study. From the patient perspective, knowing they are receiving study drug might create a **placebo effect** in which they report their health concerns from a more positive viewpoint because they believe that the drug is supposed to be helping them. Medical professionals running the study may have a similar but magnified experience, since they may feel a desire for the study to result in a positive outcome. Of course while blinding is ideal, whether a study is blinded or not must be balanced with other concerns such as subject safety.

---

### 1.6.3 SELF-SELECTED SAMPLES AND MEDICINE

While stratified convenience samples give the experimenter the greatest amount of control over bias, in medical practice a standard convenience sample won't always be an option. Why? Simply put, clinical trials involve a certain amount of risk. Therefore one must usually obtain the permission of the patient. This will result in either a self-selected sample (if you advertised for participants and they have to contact you to sign up) or a convenience sample with substantial **non-response** (if you asked specific patients of your choosing but they must have the option to say no). In the first case, sampling bias is a risk if participants' reasons for signing up might somehow impact results. For example, in a study of Alzheimer's disease, the most symptomatic patients might not experience the recruiting materials (advertisements) in a way that would allow them to consider signing up. In the second case, bias is a risk if the population of non-respondents would for some reason have different results. For our study of Alzheimer's disease – perhaps patients who are not as symptomatic refuse to participate because they don't want to admit they are diseased. In either case, you can see that a potentially important sector of the population is being excluded from the sample in such a way that the **response variable** (in this case the degree of symptoms) could certainly be impacted.

Notably some statistics texts (and some statistics instructors) lead one to believe that self-selected samples by their nature must be biased. **This is not at all the case!** Whether or not systematic exclusion from the sample truly leads to bias is a matter for careful consideration; ultimately one must assess whether or not a relationship exists between the sampling mechanism and the variables under study.

Self-selected samples may benefit from advanced strategies as was the case for convenience samples. Consider the study of the Alzheimer's patients in which we would want to ensure that each arm of the study has a proportionate number of participants at each stage of the disease. A **stratified randomization** could be implemented in the following manner. First, group available participants into several "stages" of disease. If desired, recruiting could also be structured so as to enroll participants from each group until a certain number is reached (thus ensuring that the study avoids missing any particular group). Then randomly assign half of the participants within each stage to the treatment group and the other half to the control group. This will result in the two groups being of similar **distribution** in terms of disease stage and thus mitigating the impact of the variable "disease stage" as a confounder. Additionally, inference will be possible at all stages of the disease since enrollment was structured to ensure that all were represented ("stage" would probably be included as a predictor variable in any statistical model associated to such a study).

---

#### 1.6.4 THE ROLE OF SAMPLE SIZE

It is important to note one strategy in particular that is not found in the above advice related to avoidance of bias and confounding: taking a larger sample. Larger samples improve precision but they have nothing to do with accuracy or avoidance of bias. If a sampling method results in bias, it is important to understand that increasing the size of the sample will do absolutely nothing to change that. Likewise large comparison samples in the presence of a confounding effect will not be any more useful smaller samples.

**Key point:** Increasing **sample size** improves **precision**, but **does not** improve **accuracy** in any way.

The “dart-board” analogy provides a relatively simple way to understand the ideas of precision and accuracy. A large number of darts thrown into the 1-point region shows strong precision (repeatability) but still very poor accuracy (the throws are not bulls-eyes).

Perhaps one of the worst statistical misconceptions is the prevalent belief in literature (among authors who lack training in statistics) that larger sample size is an effective solution to remediate bias and/or confounding. The recognition that sample size doesn't solve these particular problems can be helpful to you in identifying research that may have issues of bias or confounding that have gone unrecognized and unexplored.

---

#### 1.6.5 INCLUSION AND EXCLUSION CRITERIA

In most published medical studies, authors will describe inclusion and exclusion criteria related to their sample. It is important to understand the definitions of these terms because they inform us about the population and sample, respectively. First, some basic definitions:

- **Inclusion criteria** describe requirements for inclusion in the population and participation in the study.
- **Exclusion criteria** describe conditions under which a person who meets the criteria for inclusion in the population would still be disallowed from participation in the sample.

At first these may seem to be almost identical, but they should differ greatly in application. The inclusion criteria fully describe the population of interest. The inclusion criteria combine with the exclusion criteria to define the sample. While often the primary reason for exclusion is safety, people may be excluded from the participation in the sample may be removed to avoid a wide variety of other issues as well (including statistical reasons such as to avoid a potential confounder). Those excluded would still be considered part of the studied population in the sense that it would be hoped that conclusions drawn from the study might still apply to them.



As an example, consider a study of the impact of bariatric surgery on pregnancy published by Stone et. al.<sup>12</sup> The goal of this study was to compare infant characteristics for obese vs. non-obese mothers. Study participants are pregnant women having a history of this surgery. These are inclusion criteria (i.e. they define the population of interest). The study results will apply only to women who are pregnant and have had bariatric surgery prior to their pregnancy. Excluded from the sample are women having pre-gestational diabetes or chronic hypertension. We hope that the study results still apply to such women; however we must exclude them from the sample because such infants would have different characteristics (i.e. weigh more, etc.) based on these confounding variables. We would not want this to impact our conclusions about weight; however such infants and their mothers remain part of the population of interest in that we have reason to believe that even with these conditions any relationship between obese and non-obese would remain similar.

## 1.7 LEVELS OF MEASUREMENT

So we have one or more research questions and we collect data on many variables for a reasonably chosen sample. What next? Of course we want to analyze that data in an attempt to answer our questions. The problem: the list of potential statistical procedures we might apply is quite lengthy. Arbitrarily choosing a procedure and blindly applying it is unlikely to produce appropriate results. We must select appropriate statistical methods – but how?

It turns out that appropriate methods will be determined almost entirely by the **level of measurement** (or **datatype**) for the **response variable(s)** to be studied. For this reason, identifying the level of measurement is extremely important. Textbooks vary as to how many “different” levels of measurement exist. Some classify data as either **categorical** or **quantitative** (two datatypes) while others propose as many as four different datatypes. In order to select correct methods, however, it is appropriate to consider the three standard datatypes: **nominal, ordinal, and interval-ratio**. These form something of a hierarchy, with interval-ratio data at the top and representing the ideal level of measurement to support the strongest applications of statistical methods. Descriptions of these levels, together with some general guidance toward statistical methodology, are found in the Table 1.2.

There is also a fourth data-type often used in healthcare: **Likert scales**. Variables of this data-type are generally built from ordinally structured **Likert items** using interval-ratio methods (e.g. by summing or averaging variables measured using Likert items). As we might expect, there are unique challenges involved in their analyses. Though interval-ratio methods are often applied to Likert Scales, they are neither truly interval-ratio nor ordinal). A more comprehensive discussion of these issues is found in Section 1.7.2.

---

<sup>12</sup> Stone, R.A., Huffman, J., Istwan, N., Desch, C., Rhea, D., Stanziano, G., & Joy, S. Pregnancy outcomes following bariatric surgery. *Journal of Women's Health*, 20(9), 1363-1366.

Table 1.2 Standard Levels of Measurement (Datatypes)

Datatype	Definition	Methodology for Comparing Groups <sup>‡</sup>
Interval/Ratio (Quantitative)	Measurements must be numerical (quantitative) in nature; a standard metric for distance must make sense.	Histograms and Boxplots Mean (Average) / Standard Deviation Parametric Statistical Methods
Ordinal (Ranked)	Measurements represent classifications that have natural rank-order. <b>Likert items</b> are a classic example.	Bar Graphs Median / Range / Quartiles Non-Parametric Statistical Methods
Nominal	Measurements represent classifications (categories) for which there exists no natural ranking system.	Bar Graphs and Pie Charts Percentages / Proportions Non-Parametric Statistical Methods

<sup>‡</sup> Methods indicated include graphical (line 1), descriptive (line 2), and inferential (line 3). Note: direct comparisons of graphics or descriptive statistics do not constitute inferential statistical analysis.

## 1.7.1 STANDARD LEVELS OF MEASUREMENT

**Interval/ratio measurements** are generally considered the highest level of measurement and are perhaps the most powerful because they allow for the application of **parametric statistical methods**. In the definition of interval/ratio data we indicate the necessity of a distance metric. What does this mean? *Ultimately – it means that addition, subtraction, and averaging must all make sense when considering the data.* For example the difference between measurements of 1 and 2 must be exactly the same as the difference between measurements of 5 and 6. This is true for variables such as temperature, pressure, length, time etc. One key to help with identification is that interval/ratio variables will typically have physical measurement units (e.g. degrees, mm Hg, inches, hours).

As a side note, some authors distinguish between interval and ratio variables (the only difference is that for ratio variables, “zero” has natural meaning). If for example we are measuring the time (in minutes) since dosing, zero certainly has natural meaning and the variable would be ratio. On the other hand, temperature as measured on either the Celsius or Fahrenheit scales can be negative – so zero temperature doesn’t mean “none” and that variable would be considered to be interval. Fortunately, while this aspect may play a role in interpretation, it does not affect the choice of statistical method.

**Ordinal measurements** are often also presented as numbers; the difference is that the numbers are not as meaningful. They may be viewed as rankings, but mathematical operations such as addition and subtraction are generally inappropriate. One of the most common ordinal

measurements is the **Likert item**<sup>13</sup> (note that Likert items and Likert scales are different). Likert items are often developed using whole numbers. For example one might indicate the amount of pain they experience on a scale from 1 (no pain) to 10 (extreme pain). Higher values are certainly intended to be more severe, but a few moments of critical thinking should illuminate several potentially serious drawbacks to these types of scales.

1. **Distances aren't as meaningful!** Suppose that on three consecutive days I report pain levels of 2, 5, and 8. Is it reasonable to suggest that my pain is increasing in increments of 3? Absolutely not! While it is believable to simply say that my pain levels have increased, the numbers 2, 5, and 8 represent concepts (perhaps "some", "moderate", and "quite a lot"). Mathematically, the concept of  $8 - 2 = 6$  simply doesn't make sense here. It would be akin to stating that "quite a lot" minus "some" equals "moderate".
2. **Discussing an average also doesn't make sense!** Imagine reporting that the average pain was 6.32. What does this mean? It doesn't really mean anything, because this scale is categorizing pain, not truly quantifying it. The average is a concept based on distance, but distance in this case isn't meaningfully defined. So with this type of data, it makes more sense to report that the **median** level of pain reported was 6 (moderate).
3. **Different people will "measure" differently.** Presented with the same "scale", people will interpret it in different ways. For example, if you and I experience the same amount of pain, I may think the pain is a 10 while you feel it is a 7. Or we may both indicate the same value while truly experiencing different levels of pain. Note: This does **not** present an issue of bias, but rather simply *additional variability* which hurts precision. This is also an issue that can be mitigated by **pairing** (using pre/post data).

We will discuss ordinal data in substantially more detail as we approach non-parametric methods; for now however it is enough to simply understand that ordinal data should not be treated as interval/ratio data; it will require statistical methods that are different from the parametric methods used to analyze interval/ratio data.

**Nominal measurements** are typically considered the lowest level of measurements – not only does the idea of "distance" between levels make no sense – but in fact nominal classifications cannot be rank-ordered. An example of nominal measurement is marital status. Possibilities include "single", "married", "divorced", and "widowed". One could argue for a long time (and never come to an agreement) as to which of these classifications should be considered the best. For nominal measurements the idea of a median does not make sense; in comparing nominal variables we would instead consider **percentages** of the population that fall into each category. This works especially well for **binary variables** which have only two possible responses.

**Key point:** Binary variables can be quite important in study design because of the statistical methods available to analyze them.

---

<sup>13</sup> Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140: 1-55.

Finally, when considering the three variable types it is important to recognize the **hierarchy of datatypes**. When necessary, interval-ratio variables may occasionally be converted to ordinal (or nominal) variables. Likewise ordinal variables could be converted to nominal if a researcher so desired. However it is not possible to convert in the opposite direction. Nominal variables may never be converted to ordinal; neither may nominal or ordinal variables be converted to interval-ratio. Thus when recording data for variables, it is best to use the highest possible level (interval-ratio) whenever possible.

**Key point:** The hierarchy is particularly important when a study response variable is of the ordinal or nominal data type. Converting such variables to binary (yes/no or equivalent) can lead to some very useful applications of statistical methodology.

### 1.7.2 LIKERT SCALES: A SPECIAL TYPE OF DATA

A common practice in healthcare-related fields is the creation of **Likert Scales**<sup>14</sup> which are sometimes used to “measure” a concept that cannot be easily quantified using interval-ratio data. Likert scales consist of Likert items that have been “summed” (or sometimes “averaged”). Of course because Likert Items are ordinal measurements that lack clearly defined distance, adding them together likewise does not make perfect sense. Despite this, a Likert scale may be the best available measure and so we should try to make reasonable use of the data while at the same time recognizing its limitations.

As an example, in medical studies it is not uncommon to see published research utilizing some measure of “overall health”. The first thing to note is that this is not a concept that really could not be measured quantitatively. To see this, just ask yourself the following question: what are the units? That question cannot be reasonably answered, so the variable isn’t being measured at the interval-ratio level. We also recognize that “overall health” is probably a combination of several different measures – making the combination of Likert Items into a Likert Scale a reasonable and convenient assessment of “overall health”.

We can see by their definition that Likert scale data are clearly not interval-ratio. Are they ordinal? Continue to think of how they are constructed. Generally, Likert scales are constructed as the sum of several Likert items. That is to say that Likert scales are sums of ordinal data. The process of addition implies that distance matters. Of course having just read about ordinal data, you know that for ordinal data, distance is not meaningful. So what is going on here? Likert scale data are not ordinal either – they represent a specialized type of data that do not fit directly within the three standard categories.

You will generally find in the literature that Likert scale data will be analyzed using interval-ratio methods (we’ll discuss those later). That application invokes two major assumptions:

<sup>14</sup> Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology* 140: 1-55.

- The Likert items being added to create the scale are of **equal** importance.
- The “distance” between responses for each Likert item (both within and among items) is **equal**. That is to say for example that, on your fairly standard “agreement” item, “strongly agree” is equally far from “agree” as “neutral” is from “disagree”. Furthermore “strongly agree” is equally far from “agree” for each item in the scale.

Of course we know these assumptions cannot be met – since the Likert items that construct these scales represent ordinal data. Perhaps things can be “close enough”. It turns out that for many such scales, interval-ratio methods for averages can be used to differentiate between groups. *But one must be very cautious with distance; the use of confidence intervals or really any discussion of the size of any difference found is not truly valid.*

**Key point:** Confidence intervals and Likert Scales generally do not mix. This makes the clinical application of statistical conclusions related to Likert Scales very subjective.

## 1.8 EXAMPLE

Let’s consider an example study that incorporates many of the ideas already covered within this chapter. Suppose that a researcher wishes to evaluate the claim that large daily doses of Vitamin C help prevent the common cold in adults. As results may differ by race, three races will be assessed: African-American, Caucasian and Hispanic. As a general plan, some participants will receive a daily dose of Vitamin C while others do not; the incidence of colds for each individual during the study period as the response variable. Finances allow sampling near hospitals in five major U.S. cities. Study designers solicit participants from these cities until they have a group of 400 for each race; the randomization scheme is structured such that half of each group receives a daily dose of Vitamin C while the other half receive **placebo** (pills that look the same but contain no treatment). They collect data during the winter months of December, January, and February. Before reading on, try to answer the following questions about this study:

1. What is the target population for the study?
2. Identify the variables to be studied (and their types).
3. Assess the study design. Are there likely to be issues of bias? Confounding? What suggestions would you make to improve the sampling and/or randomization schemes?

What is our desired target population for this study? Ideally one would like the population to be as large as possible. For example, the results of a study applying to all people are stronger than the results of a study applying to all residents of a specific city. In this case, our goal should be to apply the results of the study to all adults of the three races to be considered. We also must consider time – it would be best if our results apply at all different times of the year, even though the common cold may be more likely during certain seasons. Notably this study as currently designed only considers winter months.

The variables under study include:

- **Response Variable:** the incidence of colds could be measured in a few different ways. Perhaps we simply record the number of colds experienced by each person during the study. Alternatively, we might record the amount of time that our participants spend under the effect of a cold. In either of these cases, the level of measurement will be interval-ratio. An alternative nominal response variable might simply be whether or not the person contracts a cold during the study. It is important to recognize that all of these probably would be affected by seasonal variation.
- **Predictor Variable / Factor:** Type of tablet given; this is a nominal variable with two levels in this case. While all participants will be given a daily tablet, some tablets will contain Vitamin C while others will be **placebo**.
- **Predictor Variable / Factor:** Race is also a nominal variable having three levels: African-American, Caucasian and Hispanic.

When assessing a study for bias or confounding, an important point is that one must at the same time consider the variables under study. What outside factors are being ignored that might affect the study? Can we in some way compensate for them?

**Important question: How might our variables be affected by extraneous aspects of the sampling?**

In the case of this example, we may fairly quickly recognize three potential issues:

1. Only residents of five major U.S. cities were sampled. Do our results still apply to people living in all parts of the world?
2. Samples were taken only during the winter months. Will our results apply to all other times of year as well?
3. Are there potential confounding variables that should be considered as **covariates**?

The answers to such questions are generally not easy; and typically there is substantial room for debate. Ultimately, we probably cannot know precisely how these things may affect the study. But we might, after due consideration, argue the following:

1. It seems reasonable that the number of colds could be affected by location (at least through some confounding variable such as climate). This could be addressed by using a more advanced model in which location is a covariate. In terms of the sampling design, we would want to stratify our sample by, for example, taking 100 people of each race at each location and randomly assigning half to the treatment (Vitamin C) and the other half to



placebo. Even doing this, we should recognize that results might not apply well to other locations not involved in the sample.

2. It may be the case that the number of colds differs by season. However, as long as we've chosen the same season for all participants, there may be less of a problem. All participants in the study are evaluated at the same time of the year; therefore unless the *magnitude* of effect would differ by season, the playing field for comparison of the Vitamin C group to placebo will be level.
3. There may well be other variables that should be considered. One that comes to mind is gender. Perhaps there are gender differences in the number of colds people have, and perhaps the effect of Vitamin C may even differ by gender. We could easily test this by taking 50 people of each gender, for each race, at each location (and then randomly assigning half to treatment or placebo as described above).

Something you might have noticed in our assessment of best design is that we ultimately arrived at the suggestion of a **balanced design**. Each combination of gender, race, and location would have the same number of observations in the above design. One last question to consider: Would this design allow for causal inference? The answer is yes. Since the treatments (Vitamin C vs. Placebo) are randomized to stratified groups of participants (i.e. that part of the design is experimental), a conclusion that Vitamin C were somehow responsible for a reduction in the incidence of colds could indeed be possible. The main concern in drawing such a conclusion would be the potential for bias in applications to participate in the study.

## 1.9 TERMINOLOGY IN CLINICAL TRIALS

This text wouldn't live up to its title if it didn't reserve at least a small amount of space to discuss clinical trials. While statisticians have a set of general terminology that is important to communication, those involved in clinical trials use several terms that have statistical meaning but are not generally seen in other environments. An excellent glossary of terms is available from Wiley:

<http://onlinelibrary.wiley.com/doi/10.1002/9780470475911.gloss/pdf>

Some particularly important definitions are those of safety and efficacy:

- **Safety:** Generally refers to the ability of humans to tolerate a drug or treatment. Safety concerns may range from minor **side effects** which may be a nuisance to **severe adverse events** which may be life-threatening. From the statistical perspective, safety information is often presented using percentages.
- **Efficacy:** Is the drug or treatment effective in treating the condition it is intended to treat? How effective? The use of statistics is always needed to make this assessment. Ideally, it is possible to use confidence intervals to evaluate the magnitude of effectiveness. A medication that has only a very mild effect while resulting in several side effects may not be worth producing.

These are in fact the primary targets in the drug development process, which following initial research in animals typically goes through up to four phases of human-subject research as shown in Table 1.3.

Table 1.3 Phases for Clinical Trials<sup>15</sup>

Phase	General Purpose	Statistical Implications
Phase 1	Evaluating Safety and Maximum Tolerable Dose	Small samples sizes will be used, often in conjunction with <b>sentinel dosing</b> in healthy volunteers; stopping criteria are developed using <b>probability theory</b> to maximize safety. Inferential statistics are generally not relevant to Phase 1 studies.
Phase 2	Efficacy and Safety	Much larger sample sizes are used with the main goal of learning whether the drug has the intended effect. <b>Power analyses</b> are important prior to dosing and inferential statistics are heavily involved in the analysis of data. A statistician should be heavily involved beginning with the design of the protocol.
Phase 3	Efficacy and Safety	Even larger sample sizes are used to gain further information about the magnitude of effect, as well as adverse reactions to the drug. Again a statistician should be heavily involved starting with the protocol design.
Phase 4	Long Term Efficacy and Safety	Phase 4 studies are for drugs that are on the market – generally the purpose is to evaluate public use of the medication and carefully evaluate long-term effects.

Some additional terms of importance that are key to the intersection of statistics and medicine include<sup>16</sup>:

- **Clinically Significant:** This is a term that we must be very careful with. For an effect to be clinically significant, it has to first be established as an effect (i.e. it has to be statistically significant). Beyond this, clinical significance also implies that the effect has been shown to be large enough to be of practical importance to those being treated. Because it is easy

<sup>15</sup> U.S. Food and Drug Administration. *Step 3: Clinical Research*. Online Reference: <https://www.fda.gov/forpatients/approvals/drugs/ucm405622.htm>

<sup>16</sup> Abdel-aleem, S. (2009). *Design, Execution, and Management of Medical Device Clinical Trials*. John Wiley & Sons, Inc. Retrieved from: <http://onlinelibrary.wiley.com/doi/10.1002/9780470475911.gloss/pdf>.

to get these terms confused, this text will use the term clinically relevant (or clinically important) when discussing this idea.

- **Bioequivalence (or equivalence) Study:** The goal is to show that treatments (e.g. two formulations of a drug) are equivalent. For example a study looking to establish a generic version of a drug would be a bioequivalence study. As we've already learned, random variability will ensure that two groups are likely to differ at least slightly. In such a case the statistician might develop a decision rule based on confidence intervals – one that concludes equivalence as long as the true treatment difference is likely to lie in a small range of clinically acceptable differences.
- **Arm (of a study):** This just refers to groups that are being evaluated (for example, the control group, treatment group, etc.)
- **Comorbidity:** Diseases or conditions other than the one for which we are attempting to treat. Comorbidities can easily be **confounders**.
- **Standard of Care:** Refers to the (medical) care that a participant would normally receive for a particular condition. Standard of care is quite often used as the control in cases where a placebo might be deemed unethical.

For human subject research, generally an **Institutional Review Board** must determine that the potential benefits from a study are balanced with its risk. Other regulatory agencies such as the FDA<sup>17</sup> and OHRP<sup>18</sup> may also have input into the approval of a research protocol. All review should incorporate a study's **Statistical Analysis Plan** to ensure that it is sufficient to ensure a reasonable probability of individual and/or societal benefit from the study design.

## 1.10 CASE STUDY

In this textbook, case studies will be generally comprised of article critiques in which we apply knowledge gained from the chapter. Before reading this section, you will wish to obtain from a library the associated original research manuscript entitled “Measuring and modelling body mass index among a cohort of urban children living with disadvantage” and published by Hollywood et. al. in the *Journal of Advanced Nursing*.<sup>19</sup> Also before continuing, read the research article and try to answer the following questions on your own:

1. What are the authors' goals? Identify the primary *research question(s)* for this study. In whom are we interested? Identify both the *population* and the *sample* in this study.
2. What are the variables and how do they relate to the goals and participants? Identify the *response variable(s)* and *predictor variables* measured as part of the study. In addition, classify each variable by its measurement type.

<sup>17</sup> U.S. Food & Drug Administration. <https://www.fda.gov/>

<sup>18</sup> Office for Human Research Protections. <https://www.hhs.gov/ohrp/>

<sup>19</sup> Hollywood, E., Comiskey, C., Begley, T., Snel, A., O'Sullivan, K., Quirke, M., & Wynne, C. (2013). Measuring and modelling body mass index among a cohort of urban children living with a disadvantage. *Journal of Advanced Nursing*, 69(4), 851-861.

3. What factors may the authors be ignoring? Consider *exclusion criteria* for the study and assess whether these would be a source of any bias. Consider whether there are any outside variables that might *confound* results. Why would they be expected to do so?

The next three subsections will attempt to answer many of these questions, based on the manuscript the authors have published. Notably, without direct questions of the authors, a substantial part of this is guesswork. This guesswork is very important, though, if one wishes to truly make some assessment about the level of faith to place in their results.

---

### 1.10.1 PRIMARY RESEARCH GOALS

In this example there are two places we might look for the goals of the study. First, there will usually be some indication of the overarching purpose within the abstract. There will also often be a short section of the manuscript dedicated to explicitly outlining the goals. In this case, the authors have a section devoted to examining the study “aim” on page 852 of the manuscript. From this section we see that, generally, the authors wish to evaluate the effectiveness of a program on the physical health of children. In particular, they are interested in physical aspects of development, fitness, and healthy eating.

One thing that we should note immediately is that these are intangible concepts. It will be important to try to understand how the authors intend to measure them. Because of the abstractness of these three things, we should expect them to be extremely difficult (if not impossible) to measure using interval-ratio data.

As to “who”, the section describing the sample (page 853) provides a detailed list of inclusion and exclusion criteria. Based on the first bullet point together with the preceding paragraphs, we may surmise that the authors are interested in all children between the ages of 4 and 12 years. This is our population. Notably, their second and third listed inclusion criteria pertain to consent and willingness to participate. Lacking either of these would result in non-response. But we would arguably still be interested in the participant as part of the population, so despite where they have located it, the non-consent really falls into the category of an exclusion criterion.

The development of their sample is somewhat complex. Seven schools were selected (or we might assume these agreed to participate). From these schools, all willing participants who satisfied the inclusion criteria were used. This **subsampling** structure is somewhat important when we consider that we would like results to apply to all schools and all students within the appropriate age range. In particular, we should consider:

- The seven schools selected are all in Ireland. Would Irish culture play a role in fitness and health? If so, we must be careful applying results of the study to other parts of the world. (In other words, consider restricting the *population* to Ireland).
- Would reasons for non-participation and/or lack of consent have any relationship to the fitness and health? If so, non-response bias becomes possible.

As a reminder, these questions likely do not have clear-cut answers. In fact, a lively debate might be developed as to the effect of these issues on the authors’ published results.

---

### 1.10.2 STUDY VARIABLES

The authors of this study are interested in measuring three concepts: development, fitness, and healthy eating. How will they do this? The data collection section beginning on page 853 of their publication provides the details. Table 1.4 summarizes their primary response variables and identifies the level of measurement for each.

Table 1.4 Response Variables from the Hollywood et. al. Study

Variable	Level of Measurement	Comments
Body Mass Index	Interval/Ratio (kg/m <sup>2</sup> )	BMI is likely being used as a stand in for “fitness”. It may also in some way measure “development”. Height and weight are actually measured here.
Waist Circumference	Interval/Ratio (cm)	The authors briefly discuss this variable at one point but it isn’t clear that they ever actually use it in analysis.

In addition to the response variables, the authors have collected demographic data including gender (nominal) and age (interval-ratio). They also have data from a referenced questionnaire that, based on their second and third tables, seems to cover two aspects:

- Several binary (nominal data) questions seem to be related to the “healthy eating” component of the study
- Ordinal level questions about the frequency of participation in various activities outside of school. It is not immediately clear which, if any, of the three constructs this would “measure”.

Though not addressed in their manuscript, it is possible that the questionnaire data could be combined into one or more construct variables that would be similar to Likert scales.

At this point, it is appropriate to consider whether the variables they collect are appropriate to their goals. Many clinicians might agree that BMI could serve as a measure of “fitness”. And it seems clear that the binary (yes/no) questions about breakfast could at least in part assess “healthy eating”. But what about “physical development”? Though the authors mention this as one of their primary outcomes, after a full read of their manuscript it remains unclear how they intended to assess this concept. This of course means that any claims made with regard to “physical development” should be treated with an appropriate amount of caution.

---

### 1.10.3 CONFOUNDING VARIABLES

After identification of their research goals, their variables, and their sample, we are now well equipped to consider whether issues of bias and/or confounding may affect their results. In particular, their goal is to compare their three primary responses across the two groups. Notably

in their sample, they had five intervention schools and two comparison schools that did not have the intervention. There are several possible issues that may result from this sample:

- *Irish Bias?* If we want to apply the results of this study outside of Ireland, can we reasonably argue that nothing about Irish culture would have an impact on BMI? Can we make the same argument for cultural patterns related to eating breakfast?
- *Similar School Cohorts?* They didn't provide a lot of information about the schools. But it is easy to imagine that there could be school to school differences that have an impact on the outcome variables. As an example, suppose it were found that the two comparison schools placed a much greater emphasis on sports? If that were true, this would be an example where the emphasis on sports would be confounded with the intervention they are interested in (i.e. one could not tell whether any differences found were related to the intervention or simply present due to school-to-school variability connected to sports. Similar consideration should be given to demographics as well as any other notable differences between the schools that could impact outcomes.

Another possible issue is noted in the exclusion criteria. They appear to want to apply the results to 5<sup>th</sup> and 6<sup>th</sup> graders, but because of the school structure they are unable to follow up on these students. In other words, their longitudinal data really only extends through the 4<sup>th</sup> grade class. Hence it would be acceptable to extend results to 5<sup>th</sup> and 6<sup>th</sup> graders only if we have reason to assume that BMI and breakfast habits would not in any way be related to grade level.

**Key Point: The case study illustrates many statistical issues that should be considered prior to data collection. Likewise when one examines a research manuscript, it is equally important to consider these issues prior to examining study results.**

## 1.11 CHAPTER SUMMARY

You should note at this point in the text, no statistical methods have yet been covered. The importance of the material in this chapter, however, is paramount. Whether you are using statistical methods to interpret your own data or reading a journal article that interprets someone else's data, an understanding of the statistical concepts discussed in this chapter can help you to separate the wheat from the chaff.

Understanding variability is the key to understanding statistical significance. Sampling is especially important; poor sampling methods aren't always easy to identify, but must be a consideration as any bias introduced by their presence can potentially make an analysis worthless. Using procedures that match the data type is vital; since successful procedure selection stems from correctly identifying data types. This topic will be a common discussion item throughout the remainder of this text. Lastly, failure to adequately consider clinical relevance and/or causation is perhaps the most distasteful transgression that occurs frequently in



medical literature. Countless authors report their results as statistically significant (or not) and incorrectly presume that statistical significance implies relevance (or conversely that lack of significance implies unimportance). Others likewise assume incorrectly that causation automatically follows from a statistically significant result.

The simple recognition of all these issues can help you to determine just how much credence should be given to the results of a research study. It can also be useful should you ever find yourself in the position of designing a study yourself.

## 1.12 EXERCISES

1. For each of the following variables, determine whether the variable is nominal, ordinal, interval/ratio, or Likert scale:
  - a. Admitting diagnosis of patients admitted to a mental health clinic.
  - b. Weights of babies born in a hospital during a given year.
  - c. Pain scores provided by patients who indicate an integer between 0 = no pain at all and 10 = worst imaginable pain.
  - d. Gender of patients visiting an eye clinic.
  - e. Score on the *Psychology Today* depression test (see website).<sup>20</sup>
  - f. Range of motion for the elbow joints of students enrolled in a university health services curriculum.
  - g. Temperatures for day-old infants who remain at least 24 hours in a hospital.
  - h. Disease-stage for patients having Parkinson's disease as defined using a 4-point Likert item.
  - i. Satisfaction scores based on a five-question survey taken by patients upon release from a hospital stay (scores are the sum of the questions).
  - j. Amounts of blood transfusions given to patients who have experienced trauma.
2. For each item in question #1, identify the population of interest to the researcher.
3. You have three available potential treatments for a newly identified infection that is prevalent only in women (the presence of infection is easily identified based on a somewhat painful rash). For a variety of reasons it is not feasible to treat patients with more than one of the three treatments. It is also of note that, typically, the body would eventually fight off the infection on its own. Identify the research questions of interest here and design an experiment that should help to answer them. (You may assume that reasonable funding is available.)

---

<sup>20</sup> Depression Test. *Psychology Today* (website). Retrieved from [http://psychologytoday.tests.psychtests.com/take\\_test.php?idRegTest=1308](http://psychologytoday.tests.psychtests.com/take_test.php?idRegTest=1308).

4. Consider the updates on Etanercept found in the 2015 publication by Butchart, et. al.<sup>21</sup> (you will want to obtain a copy of the manuscript).
  - a. In this study, 41 patients were randomized into the two groups (20 to Etanercept and 21 to Placebo). This may not have been the best methodology. What might they have done instead, and why?
  - b. In the “tolerability and safety” section, it is noted that there were 8 non-completers. They reference “no significant difference” in the second sentence of that paragraph. Explain why we wouldn’t care about a test for statistical significance here, and why the non-completers could in fact represent a limitation when it comes to examining efficacy.
  - c. Consider the final paragraph of the manuscript, which is really their overall conclusion. Do you think the authors got it right? Explain.
5. Suppose that the hospital at which you work wishes to estimate the average age of their patients. For each of the following sampling schemes, assess the described sample for bias and confounding, explaining why certain samples are poor and ultimately making some argument for which sampling scheme you believe is best.
  - a. Data are collected for every 25<sup>th</sup> patient admitted to the hospital during the next three weeks.
  - b. Dr. Adams specializes in treating cancer. Data are collected for all of her patients from the next two months.
  - c. Data are collected for all patients who are referred (for any reason) to Children’s Hospital during the next year.
  - d. Data are collected for all patients who come through the entrance over the next week and stop at a marked table to fill out the form, after which they receive a coupon for \$5 off a meal in the hospital cafeteria.
  - e. Data are collected for all patients who visit the emergency room between 1:00 p.m. December 31 and 11:00 a.m. January 1.

---

<sup>21</sup> Butchart, J., Brook, L., Hopkins, V., Teeling, J., Puntener, U., Culliford, D., Sharples, R., Sharif, S., McFarlane, B., Raybould, R., Thomas, R., Passmore, P., Perry, V., & Holmes, C. (2015). Etanercept in Alzheimer disease: A randomized, placebo-controlled, double-blind, phase 2 trial. *Neurology*, 84(21):2161-8.

6. Consider the Hrisanfow and Hagglund article published in the Journal of Clinical Nursing (you'll need to use a library to obtain the article).<sup>22</sup>
  - a. Identify the population of interest to the researchers.
  - b. Identify the primary response variable(s) and associated data type(s).
  - c. Only 66% of study participants completed the survey. Discuss whether this is likely to result in bias.
  - d. Would results of the study apply to people who are 35 years of age?
  - e. Consider all variables listed in the first column of Table 1 (page 100 of the manuscript). Identify the level of measurement for each.
  - f. The researchers intend to evaluate the impact of urinary incontinence (UI) on quality of life. Average BMI for male study participants with UI is nearly two units higher than for those without UI. Explain why this could be a problem.
7. Find a research article in your discipline that involves sampling. Identify population and variables and assess the sampling design used by the researchers for potential effects of bias and confounding. Make suggestions for the improvement of their design.

---

<sup>22</sup> Hrisanfow, E. and Hagglund, D. (2013). Impact of cough and urinary incontinence on quality of life in women and men with chronic obstructive pulmonary disease. *Journal of Clinical Nursing*, 22(2), 97-105.