On Statistical Inference

Neil J. Hatfield
Department of Statistics,
Pennsylvania State University

Revised January 2021

This reading is meant as an opportunity for you to examine the endeavor of statistical inference from several perspectives. I start off by discussing the goals of statistical inference (i.e., the big picture) before moving into key components of how we actually carry out the process of making a statistical inference.

1 The Big Picture

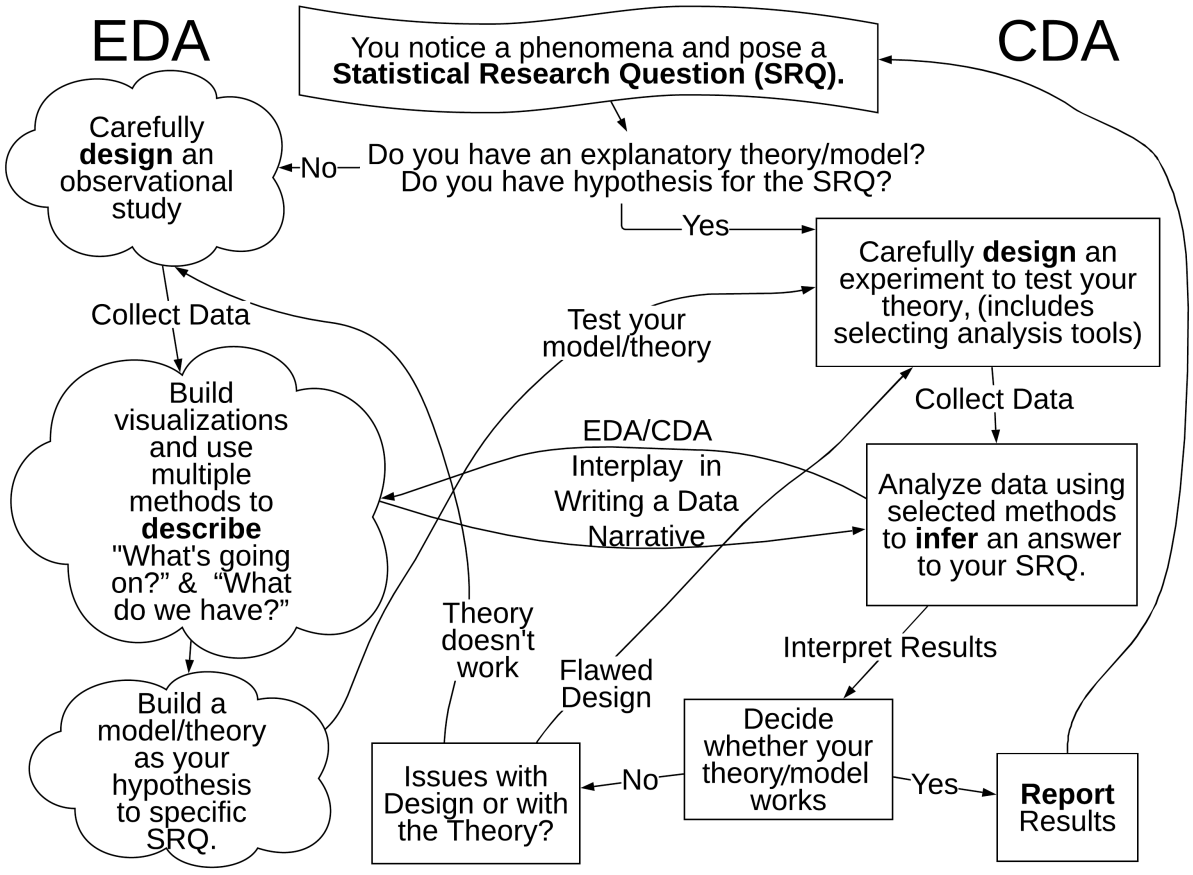


Figure 1: The Statistical Workflow

Within the workflow of Statistics (see Fig. 1), Statistical Inference occurs as part of the right-hand side. While the cloud shapes indicate using the Exploratory Data Analysis (EDA) approach and the rectangular shapes indicate Confirmatory Data Analysis (CDA) approaches, keep in mind that we often move back and forth between these approaches. This is to say that while Statistical Inference lives within CDA, the tools of Statistical Inference can be used within EDA to build models. The key is to ensure you have thought carefully about your Statistical Research Question.

At the heart of Statistical Inference is the guiding question "How Do We Make Decisions Based on Our Data?" To answer this question, we must engage in inference making. Our inferences often fall in line with one of two goals:

1) We want to say something substantial/useful about the broader population that our sample came from (i.e., population features).

2) We want to say something substantial/useful about how a particular attribute behaves in the long-run (i.e., distribution features).

In both cases we want to use the data that we have from the sample as evidence to support our reasoning (from statistical tools and methods) to make a conclusion and answer our research question. The conclusion we reach is what we refer to as the **statistical inference**. We use these inferences to make decisions such as whether a particular treatment improves patient outcomes, a teaching method improves student understandings of a particular topic, how many individuals of a particular species might live in an area, or what proportion of the population might hold a particular belief.

The two goals of making statistical inferences serve as our initial prompts for a statistical research question. Below I give examples of research questions that fall under both goals. While there are questions that fit squarely under one goal or the other, there are research questions that fit both goals. Think about the examples in both categories and their commonalities/differences. You should then think up of your own research questions and see which goal(s) your questions fit.

Population Features
Some times the goal of our inquiry is to learn something about the population from which we drew our sample. Research questions that fall into these kinds of situations include:

- What is the ethnic breakdown of people living in the United States of America?
- What proportion of individuals prefer Coke to Pepsi?
- How many Amur leopards (*Panthera pardus orientalis*) are there?
- How many Western Prairie Fringed Orchids (*Platanthera praeclara*) are there?
- What are college students' top ways to procrastinate?

Aspects of Distributions
Other times our goal is learn something about the long-run behavior of an attribute or the interaction of attributes. Research questions in this category often look like the following:

- Does Drug A reduce the length of hospital stay for individuals who suffered a heart attack when compared to a placebo as well as standard care?
- Does this new substrate increase the efficiency of the microchip?
- Is there a relationship between a person's ethnicity and political affiliation?
- What will be the expected global temperature in 2050?
- Do teachers' meanings have an impact on what students learn?

2 The Key Components to Statistical Inference

When we want to make an inference, we need to ensure that we work in a methodical way to ensure that we have valid reasoning to underpin our conclusion. To this end, we can imagine the following sequence of steps:

1) The Statistical Research Question (SRQ)
    a. Classify the SRQ by the Type of Problem
    b. Identify the necessary parameter(s)
2) Form Hypotheses
3) Select an Estimator
    a. Point vs. Interval (or both)
    b. Select a function(s)
4) Design the Observational Study or the Experiment
    a. Develop a First-order Stochastic Process to collect data
    b. Develop a Second-order Stochastic Process to get the necessary sampling distribution for your estimator (Are you going to do replication, use simulations, or employ a shortcut method?)
5) Set our Significance Level (Threshold of Unusualness)
    a. Pick a Type I Error Rate to Control
    b. Set your Error Rate
    c. Use an appropriate method to set your Threshold of Unusualness (Significance Level)
6) Carry out the study/experiment
7) Analyze the data
    a. Clean your data
    b. Describe what you have (EDA)
    c. Take the Inference Step via the Sampling Distribution
8) Report

Be aware that there is a continual back-and-forth relationship between the first four steps (i.e., the Statistical Research Question, Hypotheses, Estimator Selection, and Designing the Experiment). As you work through each of these components you'll find yourself revisiting and revising the others. However, once you begin Step 6 (Carrying out the study/experiment) everything becomes "locked in". Making any changes at this point can jeopardize the validity and generalizability of your conclusions.

2.1 The Statistical Research Question

Recall that a Statistical Research Question (SRQ) is any question that can only be answered by collecting and analyzing data using statistical methods. The SRQ is the "starting" place for all statistical inference[1]. From our noticing and curiosity, we begin to shape the SRQ into something that is reasonable and feasible. Research questions often go through dozens of revisions before we reach the point where we can move on from the question to the next stage. Even after we've moved to a subsequent step, we can and will continue to refine the SRQ; all the way up to the point of collecting data.

2.1.1 Types of Problems

One of the most challenging parts of doing statistical inference is not computing the values of your statistics/estimators, but trying to decide what statistical method(s) you need to use. Over the decades, statisticians have developed theories that support different inference methods. We've categorized these methods into overarching problem types much like Tukey proposed Reference Situations to help us recognize when certain named distributions are in play. By seeing whch of these problem types our SRQ is most compatible with, we can then leverage statistical theory to narrow our options. Below are some of the most common types of problems.

2.1.1.1 The Dichotomous Data Problem

The Dichotomous Data Problems are related to Bernoulli Trials. In fact, we're interested in estimating the probability of "success" in these problems. Alternatively, we could also think about these problems as being about what proportion of a population fits into a particular group. These types of SRQs can refer to a single population ("one sample") or can compare two different populations ("two sample").

Examples:

- Do the highest grade earning Swedish students know important information about Global Health? (I encourage you to go watch the following TED Talk: https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen).
- What proportion of USA citizens are left-handed?
- Do left- and right-handed university students have the same 4-year graduation rate?

[1] In all actuality, the starting place is when someone notices something and asks "Why is that happening?" That initial curiosity is what will drive the entire Statistical Workflow through both the EDA and then CDA sides.

2.1.1.2 The Location Problems

Location Problems come in a variety of formats. What links them all together is that we're primarily interested in learning something about the location parameter(s) for the distribution(s). The two most commonly used distribution features here are the Expected Value and the Distribution Median. In essence, we're trying to find the value of the attribute to use a central/key reference point. There are four common types of Location Problems:

- **One Sample Location Problem:** When we want to infer the value of the location parameter for an attribute based upon our single sample.
  - What would be the typical height of an adult man (woman)?
  - What is the typical score on SAT for high school seniors in Colorado?
- **Matched Pairs or Paired Replicates (Pre/Post) Problem:** When we want to infer the shift in the location parameter's value when we compare the sample in the pre-treatment condition to the same sample post-treatment. This type of problem is sometimes referred as having "dependent samples".
  - Do white leghorn chicken embryos make more beak claps in the dark or in the light?
  - Are federal pay scales commensurate with private sector salaries?
  - Do students correctly answer basic multiplication facts faster after receiving instruction on fact families?
- **Two Sample Location Problem:** When we want to infer the effect of some treatment by looking at the shift in the location parameter's value using two independent samples which had different levels of the treatment applied (e.g., control vs. treatment).
  - Does a particular social skills training program for alcoholics reduce alcohol intake?
  - Are children who watched TV or film violence more tolerant of "real-life" violent behavior than children who instead watched a nonviolent TV show or film?
  - Does using Inquiry Based Learning lead to an increase in student test scores in a Calculus I course?
- **The k Sample Location Problem:** When we want to infer the effect for multiple (3+) levels of treatment by looking at shifts in the value of the location parameter. See the Oneway Layout Problem.

2.1.1.3 The Independence Problem

The Independence Problem deals with SRQs that mimic the generic forms "Does [Attribute A] have a relationship with [Attribute B] ?" or "Are the values of [Attribute A] independent from the value [Attribute B] ?" To answer these questions, we must conceive of each member of the population as having values for both attributes. We are then trying to investigate the statistical relationship between these two attributes. If there is no relationship between the attributes we say that they are

independent. However, if there is a **dependency** between the attributes, we then investigate the nature/type as well as the degree/strength of the dependency. The Independence Problem can be extended to three attributes.

Examples:

- Is blood type independent from ethnicity?
- Is there a relationship between a person's choice of favorite color and their self-identification of gender?
- Does a person's biological sex impact political affiliation?
- Is there a relationship between a person's ethnicity and political affiliation when we account for gender?

2.1.1.4 Correlation/Association Problems

Correlation/Association Problems are intimately tied up with the Independence Problem. Here, we move beyond whether there is a dependency into to trying to measure the strength and direction of the relationship.

Examples:

- What is the strength of the relationship between blood type and ethnicity?
- What is the relation between blood pressure and age in human subjects?
- What is the relation between students' attendance and the morning temperature?
- What is the strength and direction of the relationship between biological sex and political affiliation?

2.1.1.5 The Dispersion Problems

Dispersion Problems are a lot like location problems except that instead of the location parameter being the target of interest, we're interested in the scale parameter. The most common scale parameter is to focus on the distribution Variance (or Standard Deviation).

Examples:

- Is there more variability in the time spent in the room after witnessing "real-life" violent behavior by children who watched TV/film violence than those who did not?
- Is there more variability in the amount of plant matter consumed by larval southern armyworms for Kentucky and Florida pokeweed plants?

2.1.1.6 The Oneway Layout

The k Sample Location Problem is often handled by a technique known as the Oneway Layout. The Oneway Layout not only allows us to look at $k \geq 3$ different groups at a time but allows us to get a handle on the Variation Between Collections, the Variation Between Individuals/Variation Within Collections, and the Variation Within

Individuals. This family of methods also inherently controls Type I Errors at the Experimentwise Error Rate (EER) as these methods are *omnibus tests*. An omnibus test is any inference method/hypothesis test where the alternative hypothesis is not a single statement but is a collection of statements. Omnibus tests do require doing additional testing, referred to as *Post Hoc Analysis*, to determine which of the individual statements within the alternative hypothesis collection are supported by the data.

The Oneway Layout requires us to build models for our data by envisioning each data value as being made up components tied to specific causes/factors. Each term in our model then represents a factor that contributes to the data value. The Oneway Layout can be extend out through the ideas of Blocking as well as the addition of more terms (e.g., Twoway and Threeway Layouts).

Examples:

- What is the rate of dust removal in normal people, people who have obstructive airway disease, and people who have asbestos?
- Is there a difference in the number of glucocorticoid receptor sites per leukocyte cell when looking at normal people, those who have hairy-cell leukemia, those with chronic lymphatic leukemia, those with chronic myelocytic leukemia, and those with acute leukemia?
- Is there a difference in lifetime salary based upon college major?

2.1.1.7 Twoway+ Layouts

As mentioned, the Oneway Layout Problems can be extended to what are called Twoway Layouts or even further through Factorial Designs. The underlying commonality is that we're attempting to partition a particular attribute's (i.e., the response) value into components to make up a factor model. Oneway implies that we have one factor, twoway implies two, and factorial implies that we have more than two. When we have two (or more) factors in our model, we can start adding in interaction terms. We can also use one of our factors as a "Block". A Block refers to an attribute that we don't necessarily care about in terms of explaining the response but we do want to account for in our experiment.

Examples:

- Is there evidence to suggest that a particular variety of corn produces higher yields when we account for any effects from the fields used? (The fields are Blocks.)
- Does a student's mother's level of education and mother tongue play a role in the student's achievement on a standardized test? (Twoway)
- Do students' ethnicity, letter grade in prior class, socioeconomic status, and self-reported gender impact their ACT scores?

2.1.1.8 Regression Problems

Regression Problems often deal with prediction. Specifically, we want to be able to

use the known value(s) of a certain attribute(s) to estimate the value of another attribute. Regression is like the Oneway/Twoway+ Layouts in that we are building mathematical models. In fact, the mathematical tools that we use with the former (i.e., ANVOA) are the same as here in regression; the difference is how we conceptualize the attributes we use. In ANOVA, we call them "factors" and they tend to be character-nominal or character-ordinal. In Regression, we call them "predictors" and they tend to be mostly numeric-ordinal or numeric-continuous.

Examples:

- We want to understand how the peak plasma growth hormone level behaves given what we know about short children's ages, sexes, heights, weights, and various skin fold measurements.
- We want to study the relationship between the number of bids requested by construction contractors for basic lighting equipment during a week and the time required to prepare the bids.
- We want to predict whether a student will successfully complete a class given information about his/her attendance, past academic success, and interest in the course.

2.1.1.9 Distribution/Density Problems

Distribution or Density Problems refers to questions where we want to know one of two things:

1) Does our sample support believing that the long-run behavior of the attribute(s) of interest follow the particular named distribution we've theorized?
2) Do our groups have the same long-run behavior?

Remember that when we speak of "long-run behavior" are referring to the distribution of the attribute/stochastic process that underpins our data. Such questions can allow us to make predictions about the occurrence of particular values for our attribute or even allow us to see whether or not a grouping attribute creates changes in the distributions (along the lines of independence questions).

Examples:

- Do US adult men's heights follow a normal distribution with an Expected Value of 70 inches and Variance of 12.35 inches-squared?
- Is there a difference in the distribution of student scores due to gender?

2.1.1.10 Other Types of Problems

There are many other kinds of problems that will not be dealt with here. For the most part, these additional problems can be thought of as extensions or novel applications of the previous problems designed to handle new contexts/situations. Some examples would be looking at Survival Analysis, Quality Control, Principal Components, [Latent] Factor Analysis, Discrimination/Classification Problems, Homogeneity, and Clustering, just to name a few.

2.1.2 Identifying the Parameter

Once you've identified what type of problem the SRQ is, we can identify what parameter we need to use. The parameter we select will play a role in writing our hypotheses as mathematical statements, picking our estimator, and carrying out our analyses. The following table can serve as a way to identify what parameters to use. By convention, I'm using numbered subscripts to denote different groups/samples. The above table is not exhaustive by any stretch of the imagination. Rather, think

SRQ Problem Type	Parameter
One Sample Dichotomous	p or π
Two Sample Dichotomous	Common: $p_1 - p_2$ or $\pi_1 - \pi_2$ Rare: $\frac{p_1}{p_2}$ or $\frac{\pi_1}{\pi_2}$
One Sample Location	μ or θ
Two Sample Location	Common: $\mu_1 - \mu_2$ or $\theta_1 - \theta_2$ Rare: $\frac{\mu_1}{\mu_2}$ or $\frac{\theta_1}{\theta_2}$
Dispersion	Same as Location but using σ^2 for Variance and σ for Standard Deviation
k-Sample/Oneway Layout	α_i or τ_i where i denotes group
Independence	"Independence"

about the table as providing a starting point for some of the most common types of statistical inference research questions.

2.2 Writing Hypotheses

Once you have formed an SRQ, you then start turning your attention to the possible answers. We will always form a pair of statements: one question reflects what we believe is happening and the other reflects the "status quo". The two statements are complements and together must cover the entire space of possible answers. Since the statements are complements, they are mutually exclusive. The answer to the question that you believe explains what is going on (before you collect any data) is what we call the **alternative hypothesis, H_1**. The answer reflecting the "status quo" is what we call the **null hypothesis, H_0**.

The null hypothesis often reflects the simplest model that answers the SRQ. Often this simplest model is the model that reflects that there is "no statistically significant difference" between our data and theory (or between the groups we're studying). However, we can have null hypotheses where the model does entail an actual difference (e.g., are double stuf oreos really double the stuf?), in these cases the phrase "no statistically significant difference" really becomes "no statistically significant difference beyond what we already expect".

While we typically write the pair of statements with the Null Hypothesis being listed

first, keep in mind that we often form the Alternative Hypothesis first. This reflects the Alternative Hypothesis's status as what we believe is going on. You can write both hypotheses as labeled statements or as labeled mathematical statements. I often find myself using full sentences when I'm writing and using labeled statements when I have a bullet list. My usage of mathematical statements tends to occur most within my work notes/scratch work. However, the exception is when I am expressing particular models; then the mathematical forms of those models will be included in the written report.

Examples:

- SRQ: Are adult men typically taller than 70 inches?

  H_0 : Adult men are typically 70 inches tall or shorter.
  H_1 : Adult men are typically taller than 70 inches.

  Alternate Form

  $H_0 : \ \mu \leq 70$
  $H_1 : \ \mu > 70$

- SRQ: What is the rate of dust removal in normal people, people who have obstructive airway disease, and people who have asbestos? Given $X_{ij} = \theta + \tau_j + \epsilon_{ij}, \quad i = 1, \ldots, n, \quad j = 1, 2, 3$

  $H_0 : \ \tau_1 = \tau_2 = \tau_3$
  H_1 : Not H_0 [i.e., at least one τ_j is different from the others]

  Alternate Form

  $H_0 : \ \tau_1 = \tau_2 = \tau_3$
  $$\begin{cases} H_1 : \ \tau_1 \neq \tau_2 = \tau_3 \\ H_2 : \ \tau_2 \neq \tau_1 = \tau_3 \\ H_3 : \ \tau_1 = \tau_2 \neq \tau_3 \\ H_4 : \ \tau_1 \neq \tau_2 \neq \tau_3 \end{cases}$$

- SRQ: Is the proportion of left-handed (lh) citizens of the USA 0.25?

  $H_0 : \ \pi_{\text{lh}} = 0.25$
  $H_1 : \ \pi_{\text{lh}} \neq 0.25$

- SRQ: Is there a statistically significant association between a student's level of income and academic success (GPA)?

  $H_0 :$ Student's level of income and academic success (GPA) are statistically independent of each other

  $H1 :$ Student's level of income and academic success (GPA) are statistically dependent of each other

Before going any further, look back at the examples of hypotheses. Notice that for the ones that contain mathematical statements, there is a particular format: mathematical symbol, relation symbol, and then a specific value. While not every hypothesis will be in this format, the vast majority follow this format and highlights the interdependence between the hypotheses and the parameter. Generally speaking, you want your parameter(s) all on one side of the relation symbol and the "value under the null" on the other.

2.3 Estimator Selection

As you continue on with the process of statistical inference, you must select an estimator to use. This is the function of data whose long-run behavior you'll investigate under the null hypothesis. Ultimately, you'll use the sampling distribution for this estimator to make a decision between the two hypotheses. Your choice in estimator is important but is also fairly flexible. To make your decision you can turn back to your SRQ. The type of problem that your SRQ fits under can provide you guidance as to what estimator to use. Additionally, you might end up revising your choice of estimator based upon your analysis plan that you develop later on[2]. In some cases, you might use both kinds of estimators and report both a point estimate and an interval estimate.

Recall that we have two types of estimators in Statistics: point estimators which are a single function that returns an estimate for the parameter and interval estimators which are a pair of functions whose estimate is an interval.

2.3.1 Point Estimators

When working with Point Estimators, we will often make use descriptive statistics and then make a inference leap. For example, if we want to understand whether adult mean are typically taller than 70 inches in height, we will often use the *Sample Arithmetic Mean* on our data collection. While we will get a measure of how well our collection performs the accumulation of height, when we look at the long-run behavior of the *SAM*, we know that the Expected Value of the sampling distribution will be equivalent to the Expected Value of adult men's height[3]. The following table offers a few suggestions for point estimators for some of the most common types of problems:
Again, this table only provides a small sampling of the point estimators that you could choose to use.

[2] As you gain more experience you'll quickly realize that writing hypotheses, picking estimators, and designing the experiment really merge into a single thought process that circles back to the SRQ.

[3] "Equals" and "Equivalent" aren't the same concept. When we say that two things are equal, we saying that the two things really are the same entity. Equivalent, on the other hand, says that effect is the same. This is how the Expected Value for the *SAM* can still be a rate but the Expected Value for adult height can be an amount. We know this through carefully constructed proofs in the theory side of Statistics.

SRQ Problem Type	Parametric	Nonparametric
Dichotomous Data Problems	relative frequency	
Location Problems	*SAM* Z Student's Pooled t Welch's t Matched Pairs t Post Hoc Tukey's HSD	*Sample Median* Wilcoxon Signed Rank Wilcoxon Rank Sum Wilcoxon Signed Rank Post Hoc Steel-Dwass
Independence Problems	χ^2 statistic	Kruskal-Wallis H Jonckheere-Terpstra J
k Sample Problem/Oneway Layout	ANOVA F statistics	Kruskal-Wallis H statistics

2.3.2 Interval Estimators

While you may construct interval estimators out of any of the point estimators you chose, we typically do not use interval estimators for the Independence Problems or the k Sample Problem/Oneway Layouts. Most of the time, we use interval estimators for the Dichotomous Data Problems, the Location Problems and the Post Hoc analyses.

There are two dominate types of interval estimators: point estimator based (common to parametric shortcuts) and percentile based (common to permutation and bootstrap simulations). Regardless of which style of interval estimator you use, the interpretation of the interval is the most difficult part. Suppose that you have found that the 95% confidence interval for expected value of adult men's height is $(69.3, 72.1)$. How might we interpret this interval?

The key here is to remember that being *confident* that something is true and *knowing* something is true are not the same thing. When we say that we're 95% confident we are not talking about the interval we found; rather, we're talking about the ***method*** that we used to construct the interval. Confidence in Statistics refers back to the Sampling Distribution for the Interval Estimator: 95% confidence means that 95% of the time (i.e., probability) we carry out the second-order stochastic process to get an interval estimate, the interval will contain the true value. This is the proper interpretation of a confidence interval (the outcome of an interval estimator). We have no way of knowing which of the intervals we form will actually contain the true value, thus to say that the true value is between 69.3 inches and 72.1 inches is incorrect.

This is one of the most commonly messed up portions of statistical inference, even by those who should know better. Hence the mistaking of claiming that the true value of the parameter is between the lower and upper bounds that we calculated is known as the **Fundamental Confidence Fallacy**. This most common **MIS-**interpretation of a confidence interval is as follows:

Incorrect Phrase Below*Incorrect Phrase Below*Incorrect Phrase Below

We're 95% confident that the true value of adult men's height is between 69.3 and 72.1 inches.

Incorrect Phrase Above*Incorrect Phrase Above*Incorrect Phrase Above

This erroneous phrase 1) confuses being confident with knowing, 2) makes our confidence about the particular estimate (which we already know is wrong) rather than the method of construction, and 3) supports readers in making the same faulty inference that underpins the phrase to start.

If you encounter someone who makes this incorrect interpretation you have some choices on what to do: if you feel confident enough in explaining the issue, then help them improve their understanding. In the event that you don't, then just smile, nod, and silently judge/have pity on them.

At their core, confidence intervals cannot be used to verify possible values for the parameter. Nor can confidence intervals be used to say that any value is more probable or plausible than any other value (this is known as the **Plausibility Fallacy**). The way in which confidence intervals are constructed does not allow for this. However, confidence intervals can be used to *falsify* values for the parameter. Recall that we have set up our null hypothesis to be associated with a particular value for the parameter. We then used the null hypothesis to get to the sampling distribution. The confidence interval we find contains all of the values for *non-rejected* null hypotheses based upon our collected data. Thus, if we do not find the value we specified in the null hypothesis in the interval, we have evidence to say that the parameter isn't that value.

2.4 Designing the Experiment/Study

When you reach this phase of statistical inference, you will find yourself returning to your SRQ, your hypotheses, and your choice of estimators to do constant tweaking. This is a natural part of designing any experiment or observational study. In an experiment you are going to actively control at least one attribute (referred to as a **factor**) to see what the effect is on another attribute (called the **response**). In an observational study, you'll have a much less active role in the study, merely observing and recording the co-existing values of the attributes you're looking at. Experiments allow you to build cause-and-effect models while observational studies allow you build hypotheses for later testing. Whether you are planning an experiment or an observational study, there are two distinct phases that you need to go plan as you design your experiment/study: data collection and data analysis.

2.4.1 Data Collection

The Data Collection phase of any experiment/study is where you will collect and record the data you believe is necessary to answer your SRQ. In other words, this is where you'll carry out your first-order stochastic process a certain number of times. Thus, this part of planning entails designing your first-order stochastic process.

You will need to think through what will count as the necessary data for your SRQ as well as potential sources for variation. These two aspects could cause you to refine your SRQ and/or hypotheses. For example, if you come up with a previously un-thought of source of variation that you want to track, you'll have to adjust your models to now account for this source.

2.4.2 Data Analysis

This phase is where you will think through what you'll do with your data after collection. Again, based upon your SRQ, hypotheses, and estimator, you'll be planning out your analysis approach. As you think through this component of design, you may refine the prior pieces. In essence, you will be designing the second-order stochastic process in this phase.

As you plan what analysis you're going to do, you'll need to decide on what method you want to take to get to the **sampling distribution of size n for statistic/estimator J** (where J is your chosen estimator). Remember, that there are a variety of ways in which you can get to the sampling distribution including:

- **Replication:** while the gold standard, this method is the most resource intensive; works for all SRQs and estimators provided you've designed your first- and second-order stochastic processes well.
- **Permutation:** this simulation method does require a computer to do analysis on any data sets that aren't small. This can be used to get p-values and construct confidence intervals (percentile based).
- **Bootstrapping:** this simulation method does require a computer for analysis on data sets that aren't small. This method can be used with just about any estimator and will provide confidence intervals (percentile based). This method is particularly good when we don't have a good understanding of an estimator's standard error.
- **Monte Carlo Type A:** needs a computer; can be used with most estimators; not as commonly used as the other methods. Another name for this is parametric bootstrapping.
- **Monte Carlo Type B:** needs a computer; can be used with most estimators; mainly used as the idea behind the Shortcut methods.
- **Parametric Shortcuts:** the most common approach to doing statistical inference. Many of these methods can be done by hand as well as through software. We are required to make the most assumptions with these methods. These

methods will generate p-values as well as confidence intervals (point-estimator based).

- **Nonparametric Shortcuts:** the second-most common approach to doing statistical inference. Many of these methods can only be done by hand for small data sets; computers do most of the work for us. We are required to make many assumptions for these methods, but not as strong of assumptions as in Parametric Shortcuts. These methods will generate p-values and confidence intervals (point-estimator based).

2.5 Setting the Signficance/Confidence Level

Before you claim that you've finished designing your experiment/study, you should stop and pick which Type I Error Rate you're going to control for and at what level. This is also the time for you state what you're going to use as your cutoff for unusual events under the null hypotheses (i.e., your threshold of unusualness or level of significance).

2.5.1 Picking the Type I Error Rate

Recall that a Type I Error is when we reject the null hypothesis (i.e., conclude that the null is not consistent with our data) even when the null hypotheses does a better job at describing our data than the alternative. There are five Type I Error Rates that we can choose to control. Listed from least to most conservative (i.e., amount and strength of evidence required to reject the null):

- **Comparisonwise Error Rate (CER):** This rate controls the probability of making a Type I error on a single hypothesis test. Control this error rate when you are only going to do ***ONE*** hypothesis test. The more hypothesis tests you conduct, the faster the probability of making at least one Type I error reaches 1. For example, conducting 10 independent tests while controlling CER at 0.1 will see your probability of making at least one Type I error at 0.65.

- **Experimentwise Error Rate (EER):** This rate controls the probability making a Type I error for a set/collection of hypothesis tests. If you have one set of data and you're planning on conducting multiple independent hypothesis tests on that data, this is an error rate you should look at controlling rather than CER.

- **False Discovery Rate (FDR):** This rate controls the rate at which you make Type I errors relative to the number of discoveries you make. The FDR only focuses on *statistical discoveries* (rejections of the null hypothesis) whereas the EER looks at the number of *hypothesis tests*. You can control this rate when you want a more fine grain control on Type I Errors for a set of hypothesis tests.

- **Strong Familywise Error Rate/Maximum Experimentwise Error Rate (MEER):** The FDR allows for more Type I Errors with an increasing number of statistical discoveries; to create a limit on the inflation of the upper limit to FDR, the MEER caps the probability of making a false discovery regardless of the number of tests or discoveries made.

- **Simultaneous Confidence Intervals:** This is the most conservative method and accounts for trying to fit multiple confidence intervals at the same time. In essence, this method caps the probability of making a Type I error for all of the methods of confidence interval construction occurring within the experiment/study. While designed primarily for making confidence intervals, you can also use this error rate for point estimation and getting p-values.

For most introductory courses, you'll operate mostly at the CER and EER levels. Any problem where we need to use the Oneway layout (i.e, ANOVA), the methods that we use will automatically control the EER. When you start conducting your own research, you will need to think carefully about which of these five you want to pick.

2.5.2 Setting Your Error Rate

Once you have picked which Type I Error Rate you're going to control, you need to pick the level at which you're going to control the error rate. We get to pick this value for ourselves. However, there are certain guidelines/conventions. Keep in mind that we interpret this value as the probability of making at least one Type I error. Thus, we want to keep this in mind when picking a value. The general guidance is to pick a value in the interval $(0, 0.15]$. When you go to pick your value, there are few guidelines you can follow:

- Does your lab/division/company have a set standard which you need to use?
- Does the journal you're submitting to for publication have a set standard?
- What are the standards of your field?
- Are you dealing with something that has a critical impact on people (think life/death, medical research)?
- What is your gut telling you?

These questions can help you make a decision as to the initial value. You can change this value at any time EXCEPT AFTER YOU COLLECT AND BEGIN ANALYZING YOUR DATA. The moment you have your data and you start analysis, changing your threshold of unusualness value puts you into ethically ambiguous territory. (There are a few cases where you will have to make a change; document those cases.)

2.5.3 Setting Your Threshold of Unusualness, α_{UT}

Depending on which Type I Error Rate you've selected and your initial value, you may need to make adjustments for individual tests. For what is listed below we'll use α_{UT} to represent your Threshold of Unusualness.

Controlling CER: If you only plan to control CER, then set $CER = \alpha_{UT}$.

Controlling EER: If you plan to control EER AND you're using the methods we've covered in class, then set $EER = CER = \alpha_{UT}$. Otherwise, you'll use an additional adjustment.

Everything Else: If you plan to control one of the other error rates, then you will need to adjust α_{UT} for each individual test. Common adjustment methods include Bonferroni, Holm-Bonferroni, Šidák's correction, and many others. If you find yourself in this situation, I recommend that you talk to a statistician.

2.6 Carrying Out the Experiment/Study

At this point in time, you've reached the stage where you can either become trapped in an endless cycle of revisions or you can go ahead and carry out your first-order stochastic process to collect your data. Make sure that you follow your methods as closely as possible, and keep a record of any deviations from the method.

2.7 Analyzing Your Data

Once your data is collected, now you begin the process of analysis. There are two parts of this step: cleaning and analyzing.

2.7.1 Clean Your Data

Yes, even if you are the one who has done everything from the genesis of the SRQ through data collection, you still need to clean your data. Remember, that cleaning your data is primary way you build your understanding of the data (as well as providing an opportunity for you to identify/correct any mistakes).

In addition to cleaning data, this is also where you can start building data visualizations and looking at the values of descriptive/summary statistics. You are forming the data narrative which you will then use as the backdrop for the analysis portion.

2.7.2 Analyze

This is the portion where you'll actively try to answer your SRQ (i.e., Statistical Inference). This is where you'll now carry out your second-order stochastic process. You start this process by checking the assumptions for the method you selected. Then, if you are going to use Permutation or Bootstrapping, this is where you'll set everything up in the computer and then patiently wait for the simulation to run. Otherwise, you'll carry out the Parametric or Nonparametric Shortcut.

2.7.2.1 Checking Assumptions

You'll want to think through the assumptions associated with your selected method to get the sampling distribution. Make use of data visualizations to help you check these methods as the formal tests are often not robust, resistant, or smooth.

2.7.2.2 The Inference Step

Once you have carried out the simulation/shortcut, you'll need to make the inference step. That is to say, you'll now need to make the conclusion using the data and methods as your evidence and reasoning.

- ***p*-value Method:** You'll select the appropriate p-value for your hypotheses and compare this probability value to α_{UT}. If the p-value is less than (or equal to) α_{UT}, then we have observed an event that is "unusual" given the null hypothesis. We then take that data as being evidence *against* the null hypothesis. Our inference is that we "reject" the null hypothesis.
- **Confidence Interval Method:** You'll look at the appropriate confidence interval that reflects your alternative hypothesis. You'll then use the value you specified in the null hypothesis (e.g.,, θ_0, 0, 1, 2, μ_0) and look to see if that value is contained within the interval. If the value IS NOT in the interval we say that data are *not consistent* (i.e., against) with the null hypothesis. Our inference is that we "reject" the null hypothesis.
- **Critical Value Method (Archaic):** We can also find what is called the "Critical Value" for our estimator and then compare our observed value (J^*) to this critical value (J^{CV}). If $J^* \geq J^{CV}$, then we say that our data *do not support* the null hypothesis. (IMPORTANT NOTE: the inequality must match your alternative hypothesis.) Our inference is that we "reject" the null hypothesis.

If you find yourself with a p-value greater than α_{UT}, a confidence interval that contains θ_0, or $J^* < J^{CV}$, then your inference is "fail to reject" the null hypothesis. Some individuals prefer to say "accept" rather than "fail to reject" while others are against this idea. I have no strong feelings one way or the other as long as you realize that "accept" and "fail to reject" are meant to be interchangeable in this setting.

2.7.2.3 Effect Size

The last step of analysis is to move from the "Is there a statistically significant difference?" question to the "How much of a difference is there?" question. This is the transition from the Yes/No statistical significance question to the question of practical significance. Often times researchers aren't actually interested in statistical significance, rather they want to know about the pragmatic impacts of the research. We refer to this as **effect size**. Every hypothesis test that we cover will have at least one measure of effect size. However, getting to that measure depends upon your access to different software packages. Unfortunately, Minitab Express does not include Effect Size measures. (JMP, JMP SE, and JMP Pro can provide measures through a StatsTools add-in; R can get you effect sizes through various additional packages). Perhaps the effect size that you're most familiar with is correlation; Pearson's correlation is the measure of how strong the linear relationship between two attributes is. As a final step of any analysis where you've rejected the null hypothesis, you should calculate and record the values of the appropriate effect sizes.

2.8 Writing Up and Report Your Findings

Just as you should end your describing data phase of the Statistical Work Flow with a written data narrative, you'll need inference phase methods with another written narrative. This time you're writing up your inference. You do not need to go into the nitty-gritty details of what you've done, nor do you have explain what each value means. (You get to make some assumptions about the reader's level of knowledge.) You'll want to be sure to include the following:

- A mention of your level of significance (i.e., α_{UT}) or confidence (the complement of α_{UT})
- What method you used used (permutation, bootstrapping, Parametric/Nonparametric Shortcut–give the specific name for the shortcut used) and your chosen statistic/estimator
- That you checked the assumptions and what you found; data visualizations can be useful here if there are potential problems that you need to discuss
- The observed value of your chosen statistic/estimator
- Either the p-value or the confidence interval you've found
- State your inference
- If rejecting the null, give effect sizes
- Repeat with any Post Hoc analyses

With time you'll get faster and better at writing up your findings as well as weaving them into the larger data narrative.

3 Examples

In this section, I will go through several examples of the process laid out above. To save space, I won't necessarily layout all of the details.

3.1 Baseball Caps and Height

Suppose we're interested in whether people who wear baseball caps regularly are shorter than people who don't. This would give us the SRQ "Are people who regularly wear baseball caps shorter than people who do not regularly wear baseball caps?" This kinds of question is a two sample location problem as we're ultimately wanting to compare the location parameter for height between two populations (regular baseball cap wears and those who aren't). We will let's use the SAM as the basis for a difference statistic and use a parametric shortcut (Welch's t test). We'll update our choice of estimator to Welch's t statistic. Our hypotheses will be:
H_0 : regular baseball cap wearers are just as tall or taller than non-regular baseball cap wearers
H_1 : regular baseball cap wearers are shorter than non-regular baseball cap wearers

For any individual, we know that we need to log two attributes' values: the individuals height (inches) and whether or not they are a regular baseball cap wearer. Let's state that someone is a regular baseball cap wearer if they wear a baseball cap at least once a day, at least four times a week. We will go ahead a restrict our population to students at our university, modifying our question: "Are students at our university who regularly wear baseball caps..." Since we're only doing one hypothesis test, we'll control CER and we'll set $\alpha_{UT} = 0.05$.

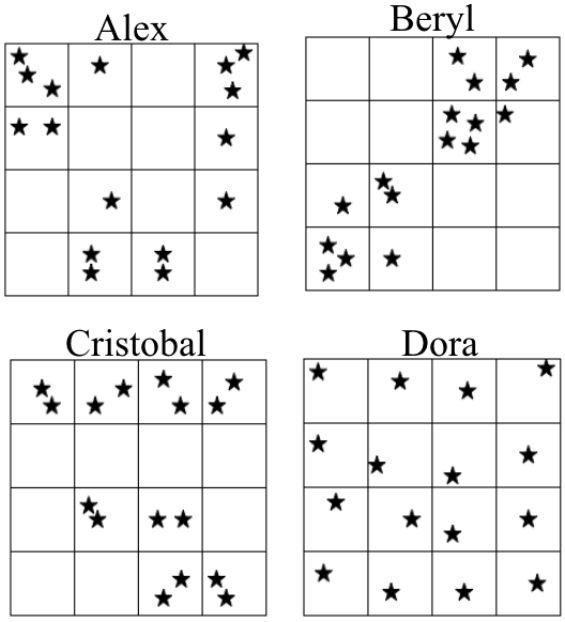
Our first-order stochastic process would then consist of getting a listing of students from the Registrar's Office to be our sampling frame. We would then construct

a lottery. We might want to ensure that we get representation for each year of school (first-year, sophomore, junior, senior, graduate) as well as gender (male, female, other). We'll revise our single lottery into 15 separate lotteries reflecting the unique pairings of each year and each gender. For each stratum, we will select an individual and see if they are willing to participate. If yes, record their height and how often they wear baseball caps.

For our second-order stochastic process, we'll make use the Welch's t as a parametric shortcut. After plugging the data into a software program, and calling the t-test option, we can calculate a p-value for our one-directional test. If we reject the null, we would then want to calculate the appropriate effect size measure. (I don't have data for this question so I can't go much further.)

3.2 Figuring Out Who Cheated

Investigating whether or not some cheated at a game or on exam is a common application of statistical inference. Here, we'll look at whether or not four students cheated at playing the Star Game. The game is played by drawing out a numbered piece of paper from a paper bag and drawing a star in the corresponding numbered box of a grid. The piece of paper is returned and the player repeats this process 15 more times. Here are the four game boards for the students:



Our question of "Who cheated?" does indicate that we will need to make more than one inference. Thus, we really have four SRQs, one for whether each student cheated. This gives us four sets of hypotheses:

1) Did Alex cheat?

H_0 : Alex did not cheat.

H_1 : Alex did cheat.

2) Did Beryl cheat?

H_0 : Beryl did not cheat.

H_1 : Beryl did cheat.

3) Did Cristobal cheat?

H_0 : Cristobal did not cheat.
H_1 : Cristobal did cheat.

4) Did Dora cheat?

H_0 : Dora did not cheat.
H_1 : Dora did cheat.

Since we know that we're going to need to do four hypothesis tests, we should control our Type I Errors with a stronger rate than CER; let's control EER and set our initial $\alpha_{UT} = 0.15$. We'll go ahead an use Bonferroni's correction to adjust our individual tests' thresholds, $\alpha^*_{UT} = \frac{0.15}{4} = 0.0375$.

While our data comes to us automatically, we'll still think through the generation of the data (also serves a refresher of the context): a person gets a bag with the number 1–16 on squares and a numbered game board. The person reaches into the bag and draws out a single square; the person than places a star in the corresponding numbered space on the game board. The person then returns the square to the bag. The bag is shaken. Repeat until the person has drawn a total of 16 times. Our questions about whether each student cheated falls into the problem type of Distribution/Density Problem. In essence, we can compare each student's game board to the long-run behavior of the game.

Individual	Value of Kolmogorov's D	P $\left[D^* > D_0 \middle\vert \text{assumptions}\right]$	Is p-value $\leq \alpha^*_{UT}$?
Alex	0.0696	1.0000	No
Beryl	0.1321	0.4838	No
Cristobal	0.2348	0.1024	No
Dora	0.3676	0.0033	Yes

To test these hypotheses, we will make use of a nonparametric shortcut for Distribution Problems using a statistic called Kolmogorov's D. This statistic measures the discrepancy (as a distance) between a particular named distribution and the actual accumulation of outcomes. In this situation, the long-run behavior of the game is a Poisson distribution with a unit rate of success equal to 1. Since we're using a shortcut our second-order process involves invoking a particular distribution's CDF to get a p-value.

We can now see that our data support the inferences that Alex, Beryl and Cristobal did NOT cheat while the data support the idea that Dora did cheat. (We don't worry about effect size in this setting.) I should quickly remark here that while this is evidence supporting the model that Dora cheated at the Star Game, we do not know if she *actually* cheated; she could legitimately get that game board which is a very rare occurrence (happens 0.33% of the time you play the game forever). Jumping straight to saying that she did cheat puts you on the fringes of a Type IV error.

3.3 Types of Oreos and the Amount of Stuf

Not counting different flavors of creme filling nor flavors of wafers, there are four major types of Oreos: Regular, Double Stuf, Mega Stuf, and Thins. Suppose that we want to know whether or not there actually is a difference between the amount

of creme filling (the stuf) based upon the type of Oreo. Since we're looking at four different types of Oreos, we are in a four sample location problem and will need to use the oneway layout. We need to track the mass of creme filling (in grams) and the kind of Oreo. The kind of oreo will be our factor. We will make use of the following baseline model where Y_{ij} represents the amount of cream filling for oreo i of type j: $Y_{ij} = \mu_{..} + \alpha_j + \epsilon_{ij}$. (We'll let 1 represent Double, 2 represent Mega, 3 represent Regular, and 4 represent Thins.) This will let us then use the following hypotheses:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 \text{ vs. } H_1 : \text{ Not } H_0$$

As we think about our first-order stochastic process, we will want to first generate a list of stores that sell all four types of Oreos. Our sampling frame will only consist of such stores; for my convenience, this will be further restricted to stores in Tempe, AZ. After selecting the store via a lottery, I will drive to said store and go to the appropriate aisle. For each type of Oreo, we will assign each package a unique ID and conduct a quick lottery to select the one package of each type to be used. (This is an example of tiered sampling.) I will then purchase the cookies and return home. We will measure each cookie's mass of creme filling (we're doing cluster sampling). As we record the type and mass of cream filling, we will also want to record which row the cookie came from and the measurement position. This will helps decide whether or not Oreo cream filling is actually independent; they aren't: https://youtu.be/jS_et4PP0P8?t=36.

Since we're in a Oneway Layout (ANOVA setting), we ideally would want to use the F statistic. We also will control the EER Type I Error Rate; let us set $\alpha_{UT} = 0.05$. However, we do have an issue with independence as well as normality. The parametric F test while robust to slight violations of the Normality Assumption, is not robust to violations of the Independence Assumption. However, bootstrapping is more robust. Thus, we will do the following for our second-order stochastic process:

1. Calculate the observed value of the F statistic; that is, $F^* = 1889.134$
2. We will bootstrap 1000 replicates of the F statistic using 416 as our random seed.
3. We will then look at the 95% Bias Corrected Confidence Interval; $(1347.49, 2594.73)$

I've added the results above. If you would like to follow along, you may do so by accessing the data here:

> JMP: https://www.dropbox.com/s/wharwik1xgv5mnk/Oreo.ANOVA.jmp?dl=1
>
> Minitab: https://www.dropbox.com/s/pkm0bhe8vd50z1w/Oreo.ANOVA.mpjx?dl=1

To make our decision, we would need look for the value of F under the null hypothesis; in case this would be the expected value of the F distribution with 3 and 123 degrees of freedom or 1.0165 (note: this is beyond this course to know how to calculate). That value is well outside of our confidence interval, thus our data are not consistent with the null hypothesis.

Since there is a statistically significant difference in the effect of Oreo type on the mass of creme filling, we should also look at effect sizes for the omnibus test (i.e. ANOVA) as well as for follow up Post Hoc tests. The alternative model that the type of Oreo contributes to the mass of creme filling accounts for 98% of the all of the variation in the mass of the stuff ($\eta^2 = 0.97876$, $\omega^2 = 0.97807$).

All of the types of Oreos are statistically different from one another with large effect sizes; Cohen's d varies from 8.758 when comparing Mega to Thin down to 1.348 for Regular to Thin. The smallest probability of superiority is 0.83 for a Regular Oreo beating a Thin Oreo for amount of stuf; the others are essentially 1. Thus, Mega stuf has the more stuf than Regular, Double, and Thin; Double stuf has more than Regular and Thin but less than Mega, Regular has more stuf than Thin, but less than Double and Mega, and Thin has less stuff than the other three. (Effect sizes *do* let us go beyond the standard yes/no for interpretations.)