CSDA 5320 Module 2 Statistical Inference Crash Course

N. Kasarov, Webster University

January 2025

1 List of Terms

The following is a list of terms which will be used in this text.

"Statistics": a mathematical discipline concerned with collecting and analyzing data
"Statistic": a function of the data
"Data": some empirical recording of facts
"Parameter": a variable in a statistical model related to the unknown quantity of interest
"Hypothesis": a statement which is yet to be evaluated for its truthfulness
"Probability": a philosophical and mathematical concept which was formalized in the 20th century by Russian mathematician Andrey Kolmogorov
"Inference": a conclusion based on facts, evidence, or reasoning
"Experiment": some repeatable and empirically measurable occurrence, the outcome of which is not known in advance
"Model": a formula or some set of mathematical concepts which explain the relationship between variables in a dataset
"Sample": data taken from a population of interest which is to be studied
"Population": a set of real or abstract objects which share some common trait(s) and are sampled for analysis

2 Introduction

When we talk about statistics, one common thing comes to mind: some mixture of percentages and charts which, unless they are related to something we are personally invested in, bores all to death. What we miss with this stereotype is that this is only one aspect of the discipline. If describing past data is the realm of descriptive statistics, making statements about unknown (either future or unobservable) events is the art of inferential statistics.

Inferential statistics all involve one thing: the inference. Unlike their descriptive counterparts, inferential statements are not as easy to make. If we were to record the grades of 100 students and 30 of them were graded "B", nobody would dispute statements like "30 percent of the students got B". If we were to change the verbiage and say "30 percent of all students are graded B", eyebrows would be raised. One was just a description of established facts, the other was a statement about a greater, unknown population.

To do data analysis (i.e., to analyze facts), one needs an appropriate mastery of both descriptive and inferential methods.

2.1 Statistics in Data Analysis

Statistics is the discipline concerned with the collection, description and reasoning about empirically measurable facts. Its role in data analysis is indispensable, providing standardized procedures on how to interpret different types of data. This standardization, in turn, allows for the automation of many procedures (through programming languages like Python or R), saving analysts both time and effort, and allowing them to focus on the interpretations themselves.

2.2 Analyzing the Unknown

The problem with reasoning about the unknown is obvious: How can one be sure about their conclusions if the concept is unfamiliar? Guessing is one way, but most would agree that it is not reliable. Asking an expert would be feasible, but there is no guarantee that one would be available. Writing down some known facts about it and using them to reach a conclusion seems reasonable. One problem, though: What happened yesterday is no promise for what will occur tomorrow. This is known as the problem of induction. In logic, if one knows something about a greater whole, they can infer things about parts of the whole:

"All Wesbter University students are excellent."
"John is a student at Webster University."
"Therefore, John is an excellent student."

This conclusion is known as deduction - it draws conclusions based on known premises. Deductive reasoning is how the Pythagorean theorem was proved and how one knows that $2 + 2 = 4$. It is a powerful reasoning method but rarely achievable when every premise is not known in advance. Induction is the reverse: drawing general conclusions from single observations. Obviously, things get more complicated in this case:

"I saw 100 dogs."
"All 100 were friendly."
"Hence, all dogs are friendly."

The problem with this conclusion is obvious: 100 dogs sharing that trait does not mean that all dogs are friendly.
The problem of induction was first analyzed by Scottish philosopher David Hume in his "A Treatise of Human Nature". The actual "problem" is that, according to Hume, there is no rational justification for inductive inferences. This is because while one may know all the facts that have occurred before, there is no way to guarantee that the next occurrence will be the same as those before. His thesis is that since people do not have access to the future, they have no way of verifying facts until they actually occur. Because they cannot be verified, there is no rational justification for one inductive method over another, since none can be proven right until their prediction happens or fails, he concludes.

2.3 Reasoning About Facts

The question then becomes "How does one reason about the data?" The answer depends on what the reasoning is aimed at.
If everything was known in advance, all questions would be answered with certainty. Reality, of course, does not allow people to know everything for certain. Hence, uncertainty is introduced every time one is presented with something unknown. Probability is the philosophical and mathematical concept which allows the quantification of this uncertainty. It is the proper use of this quantification that governs good reasoning about data. While it does not solve the problem of induction, it has proven itself to be a powerful tool for making data-based conclusions. Choosing the right approach is what makes a good data analyst stand out.

3 Probability

3.1 Basics

In mathematics, probability is a real number between 0 and 1 which shows how likely some event A is to occur:

$$\mathbf{P}(A) = x, 0 \leq x \leq 1$$

A probability of 1 indicates a guaranteed outcome while a probability of 0 suggests an impossible event. Anything in-between suggests uncertainty about the occurrence of the event. This uncertainty is the prime interest of both probability theory and statistics (and, naturally, data analysis).
While quantifying this lack of certainty is interesting, assigning some number to an event does not itself suggest what actions should be taken or what conclusions should be reached with that knowledge. One reasonable question is "If an event has probability 0.3, does that means it occurs 3 out of 10 times?" Another one could be "If the probability for this event is 0.6, should I believe it over its alternative?" While probability theory does define the mathematics behind

the concept, it does not offer an answer to either of these questions. Clearly, this is a problem as the practical application of such findings requires some interpretation of said "probability".

3.2 Probability Interpretations

How one interprets (or perhaps, refuses to interpret) probability has direct implications on what kind of data analytical procedure is appropriate to explain the facts at hand. More explicitly, an analyst cannot just perform an analysis and claim whatever they please about it. The whole point of an analysis is that it is reproducible and verifiable by others (hence, scientific and trustworthy). Being aware of the types of facts and knowledge available is key to choosing the appropriate type of procedure and inference which best suit the situation.

3.2.1 Frequency

If an event occurs m out of N times when an experiment is performed, its relative frequency is defined as m/N. If the same experiment gets repeated once again and the event occurs, the relative frequency is (m + 1)/(N + 1); if it does not occur, the updated relative frequency is now m/(N + 1). The more the experiment repeats, the more this frequency will be updated.
This interpretation of probability is sometimes called "Frequentist". Most of modern statistical methods are built around that interpretation. Frequentism enjoys popularity among the sciences due to its close ties to Karl Popper's 20th century influence on the philosophy of science (logical positivism, empiricism, and falsifiability, to name a few). It is also applied in most industrial, economic, governmental, and other applications due to its availability and huge selection of inferential procedures.

3.2.2 Belief

Dan has an old coin and he wants to check if it is fair. He flips it ten times just to be sure, 8 out of 10 flips are heads and he concludes the coin is biased. While Dan initially believed the probability of an outcome to be 0.5, some empirical data shook his perception. He then changed his probability opinion about the coin because it made sense to do so.
This is a subjective interpretation of probability which describes how previous knowledge can be updated in light of empirical evidence. It is known as Bayesianism and is a quickly developing methodology in many fields like robotics, AI, personalized medicine, and law because it allows for a rational update of prior beliefs (i.e., learning).

3.2.3 Plausibility

A common feature between Frequentist and Bayesian interpretations of probability is that they do, in fact, provide one. This is all weel and good, but some might ask: Is it really neceaasry? Both use the same probability calculus, how

different could they be? It is not uncommon for one to find desirable traits in both approaches, as the different interpretations lead to obviously different statements at the end of the analysis. Another common characteristic between them is their mutual usage of both probability theory and statistics, making use of common mathematical tools.
If one takes a pragmatic (and somewhat reductionist) approach and agrees that the probability of something is the output of a probability function, then it does not matter what the probability interpretation is. Building on this, if one then accepts that

$$\mathbf{P}(H_1) > \mathbf{P}(H_2)$$

suggests explanation (or hypothesis) H_1 explains facts better than explanation H_2 when performing an experiment, then one has statistical evidence favoring H_1 over H_2. This approach is concerned with direct assessments of facts (data) and their proposed explanatory mechanisms (i.e., mathematical models) and is known as Likelihoodism. While it does not provide an explanation of what those probabilities actually refer to (like beliefs, frequencies or something else), it provides a well-defined way to compare empirical facts objectively, focusing solely on data and model.

4 Inference

When something gets modeled, it is essentially reduced to a simpler concept which can be studied with the aid of mathematics. Inferential statistics does just that: it collects data and tests how well probability models explain the relationships between variables in a dataset. The keyword here is "test". Given some arbitrary column of written-down numbers, an analyst could choose a multitude of methods to "crunch the numbers" with, especially with the aid of modern software like Python or R. It would all be an exercise in futility, however, if the numerical results were not supported by some explanation (the whole point of an "analysis").

Whenever an inference is made, it is based on a hypothesis (i.e., an unconfirmed statement to be tested). In statistics, a hypothesis is some statement about the parameters (i.e., different variables which are not part of the recorded data) of a statistical model. How this hypothesis is evaluated is what separates the good analyses from the useless one.

4.1 Connecting Probability and Inference

In an experiment which involves randomness (i.e., the different outcomes cannot be predicted with certainty), one needs a list of all possible outcomes of said experiment (if not all are known, the exercise is pointless as the uncertainty is not quantifiable). Call it Ω:

$$\Omega := \{\omega_i : i \in B \subseteq R\}$$

This "list" is called a set in mathematics. This set is equipped with a probability function which assigns a probability number to each event of interest that this set (or "list") can generate.
Depending on how complex the experiment is, Ω can be quite arbitrary and calculating the probability of some abstract event may become impractical. For example, if a doctor was to try to account for every possible eventuality which might cause a patient a heart attack, the list could possibly go on forever. Easy to state mathematically, difficult to implement in practice.

If the doctor has a heart monitor, they could measure patient heart rates, record them (and maybe some other facts like demographics), and then study how patients with or without the condition behave on the monitor. Consequently, they could make a decision or conclusion based on the observed facts. This is what statisticians call a random variable (mathematicians just call it a measurable function):

$$f : \Omega \mapsto S \subseteq R$$

Anything which can be empirically measured and varies can be modeled as a random variable: the height of individuals in a given country, the grades a class of students receive on a test, the number of customer complaints a call center receives a month.

Because randomness is involved, these functions involve variables which account for that variability (and sometimes may be completely unrelated to the actual data). These variables are the model parameters: they affect the probability of a measurement and are thus of interest to hypothesize about. Call them Θ:

$$\Theta := \{\theta : \theta \in G \subseteq R\}$$

This set Θ is sometimes called the parameter space and Ω is called the sample space. A fundamental effort of statistical inference is using values from Ω (the data) to infer about Θ (the unknown) by using f (the model).

4.2 Frequentist Inference

Frequentist statistics interprets probability as the long-run frequency of an event, given an arbitrary number of repetitions of the experiment which produces the event ("experiment" here does not have to mean a strict lab experiment; it can be any occurrence which can produce the event of interest). The keywords here are "long-run" and "frequency". "Long-run" can be in the thousands, millions or hundreds of trillions. The goal of a Frequentist analysis is to evaluate how a statistical statement about the model (i.e., the hypothesis) will be correct during this long-run repetition of the experiment.

Doing so requires knowledge about how an estimate behaves under continuous repetition (i.e., how the mean age of persons will change when sampled from different neighborhoods, how the maximum salary changes from city to

city). This is known as the sampling distribution of a statistic: the set of all possible values it can take. By judging how these values behave in the long-run, Frequentist analysis can show how probable the observed measurement is, given the hypothesis. It conducts its statistical inference by connecting Ω with Θ through the sampling distribution (i.e., the long-run behavior) of f.

4.3 Bayesian Inference

Bayesian statistics interprets probability as the rational and subjective degree of belief of an individual/agent. As such, it aims to incorporate the individual prior knowledge about an event and refine those beliefs by analyzing data.

The agent starts with previous knowledge about the unknown object of interest. This knowledge is to be formalized in statistical terms for the analysis to work, more specifically in a hypothesized distribution. Statisticians call this a prior distribution. Once the prior distribution is selected and data collected, the probabilities of observing some hypothetical values of the parameters are calculated using the Bayes theorem (mathematical theorem in probability theory discovered by English monk and probabilist Thomas Bayes). Once the prior gets "updated" through the data, the result of the analysis is a probability distribution about the possible values of the parameter(s) (which is, frankly, what most expect to hear when asking about specific hypotheses).
Bayesian analysis can show how probable a certain value of the unknown object (parameter) of interest is. It draws its inference by connecting Ω and Θ through the prior distribution, model, and data.

4.4 Likelihoodist Inference

Whatever the objective or subjective of probability may be, the probability numbers remain outputs of a mathematical function. Once a model is chosen and data is collected, the goal then becomes to study how different versions of the model (i.e., different values θ for the parameter space Θ) explain the data generated by Ω. It almost seems like common sense to say that if one value of the parameter has a higher probability than another, given the same data:

$$\mathbf{P}(\theta_1|Data) > \mathbf{P}(\theta_2|Data)$$

, then θ_1 does a better job at explaining the empirical findings than θ_2. This is known as the Law of Likelihood. If this "law" is accepted, then one has a formal mathematical way to compare different parameter values of the same model, given the observed data. This comparison can be done by calculating the likelihood ration for any pair of $\theta_{i,j}$:

$$LR_{\theta_i/\theta_j} := \mathbf{P}(\theta_i|Data)/\mathbf{P}(\theta_j|Data)$$

If the LR = 1, both values are the same and equally plausible; if LR exceeds 1, the probability of the numerator exceeds the denominator and vice versa. The

bigger the ratio, the more statistical evidence one has for the numerator over the denominator.
Likelihoodist analysis draws its inference by evaluating the evidence for different parameter values $\theta_i \in \Theta$ against the observed data $x \in \Omega$. Unlike Frequentists and Bayesian analysis, it does not provide probability statements about the long-run behavior of an estimate or the parameter values. It is, however, the only form of statistical inference to date that offers a formalized concept of statistical evidence.

5 Inferring about Data

Data analysis, as shown, is a diverse field which borrows heavily from already established methods. What the analysis will be depends on the analyst, data, analytical question, and availability of analytical tools. The stages of a good analysis will (ideally) involve:

1. Understanding what is asked
2. Formalizing it to something measurable
3. Choosing an appropriate model
4. Collecting measurements
5. Running analyses
6. Making conclusions

For the purposes of the course:
1 is understanding the assignments
2 is figuring out which variable holds the measurement or how to calculate it
3 is choosing the correct analytical procedure
4 is loading the dataset
5 is programming the Python code needed to perform the procedure
6 is stating the correct conclusions, given the dataset, measurements, and procedure of choice