CSDA 5320 Module 6 Statistical Hypotheses Crash Course

N. Kasarov, Webster University

January 2025

1 Introduction

In science, theories are built on the consistent explanations of observable facts. A theory can be shown to be credible if it can produce falsifiable statements about the system it describes. Falsifiability, as defined by Austrian philosopher Karl Popper, is the ability of a theory/hypothesis to be logically contradicted by some empirical test. If it cannot be logically contradicted by an empirical test, there is no way to empirically evaluate its claims. If something cannot be empirically evaluated, it is of no interest to science.

1.1 Statistics and Science

Modern advancements in the sciences are in no small part due to the drive towards empiricism. Empiricism is a philosophical paradigm according to which conclusions must be based on sensory experiences (like sight or smell) and empirical evidence (results produced by experimentation). Statistics, especially its Frequentist paradigm, provides the methods to do just that: empirically evaluate (or test) theories and hypotheses.

1.2 Statistical Hypotheses

In everyday life, a hypothesis means an assumption which can be tested to see if it is correct. In statistics and data analysis, the notion is the same: a statistical hypothesis is some claim about the parameters used in a given model. A hypothesis can be simple or composite. Simple hypotheses make specific statements about the parameters: "The average individual height in the city is 190 centimeters". A composite hypothesis suggests more than one value for the parameter of interest: "The average headcount in the building at noon is more than 50". Choosing the right hypothesis to evaluate is a critical part of data analysis.

2 Statistical Hypothesis Testing

2.1 Background

Modern statistical methods owe much of their existence to the 20th century statisticians who first pioneered the wide usage of such procedures. The problem then was the same as it is now: there is no universal way to analyze facts and disagreements were rampant. Modern Frequentists statistics (the inferential method studied in this course) are the result of a combination between Fisher's significance testing and Neyman-Perason's hypothesis testing. The result is a mix of the two.

2.2 Modus Operandi

Per Popper, a hypothesis needs to be falsifiable to be of scientific relevance, i.e., it can be logically contradicted by an empirical test. Because logical contradiction boils down to a binary choice, the results of the test should suggest whether the hypothesis is usable or not (i.e., a binary decision).

2.3 Null Hypothesis Significance Testing (NHST)

Null Hypothesis Significance Testing (NHST) is a fundamental procedure in statistics used to assess whether there is enough evidence in a sample of data to infer that a particular hypothesis about a population is true or false. The core idea is to test whether the observed data are consistent with a specific assumption (the null hypothesis), and then decide whether to reject that assumption in favor of an alternative hypothesis.

Begin with a "null" hypothesis H_0: there is no effect, no difference, or no relationship between variables in the population. Example: "The mean weight of apples in a given orchard is 150 grams ($\mu = 150$)". This in turn defines the alternative hypothesis H_a which is the logical alternative to the null and suggests that there is an effect, a difference, or a relationship in the population. It is what the analysis aims to unearth evidence for. Example: "The mean weight of apples is not 150 grams ($\mu > 150$)".

Next, choose a significance level α. The alpha level is the probability threshold used to decide when to reject the null hypothesis. The most common choice is 0.05. Next, collect data and compute the relevant test statistic (e.g., a t-statistic or z-statistic). The statistic (which is some function of the data) will produce a certain value. Because different samples can produce different values of the same statistic, one needs to know the range (or "region") of values the statistic can produce (the sampling distribution). Some of these values will be appropriate if H_0 was assumed to be true, some will not. Call the values which support H_0 the "critical region".

Next, determine where the value of the statistic for the current sample is (in or outside of the critical region). To do so, determine the p-value, which is the probability of obtaining the observed (or a more extreme) statistic result, if the null hypothesis was assumed to be true. If this p-value is below α ($p < \alpha$), the null hypothesis H_0 is rejected; otherwise, it is not.

2.4 Falsifiability of NHST

When performing NHST, the analyst engages in a probabilistic version of Modus Tollens:

If A, then B
Not B, therefore not A

A is H_0, B is the critical region. If H_0 is true, the the test statistic will fall within the critical region. If the statistic does not fall in the critical region (not B), H_0 is rejected (not A).