35 Applied Inferential Statistics

David Housman

## Level of Significance

The *level of significance ()* is the probability at which we say that an improbable event does not occur. More concretely, is the acceptable probability for erroneously accepting (a) that a new drug is effective when it is actually ineffective, (b) that a person is guilty when they are actually innocent, or (c) that a company is lying about the quality of its product when they are actually telling the truth. In scientific literature, often used levels of significance are 0.05, 0.01, and 0.001.

## Hypothesis Test

A *hypothesis test* consists of null and alternative hypotheses. A *test statistic* that measures how far the data are from the null hypothesis is chosen. The *p-value* is the probability of obtaining a sample from the null hypothesized population that has a test statistic at least as supportive of the alternative hypothesis as obtained from the data. The p-value could be approximated with a simulation or calculated with a theoretically derived formula. If the p-value is smaller than the level of significance , then we reject the null hypothesis and accept the alternative hypothesis. If the p-value is at least as large as the level of significance, then we fail to reject the null hypothesis and fail to accept the alternative hypothesis.

The hypothesis test process is similar to jurisprudence in most western democracies. The accused is assumed to be innocent (the null hypothesis) and is determined to be guilty (the alternative hypothesis) if and only if the evidence (summarized by the test statistic and p-value) proves guilt beyond a reasonable doubt (the p-value is less than the level of significance). Although assumed to be innocent, the accused is either proclaimed guilty or not guilty (which could be considered different from proclaiming innocence).

A hypothesis test process is also similar to a mathematician’s proof by contradiction. To prove that is not rational (the alternative hypothesis), the mathematician supposes is rational (the null hypothesis), finds a contradiction (evidence that shows the p-value is zero), and concludes that is not rational. Evidence derived from a sample almost never yields a p-value of zero, and so we are willing to say that will a small enough p-value there is sufficient evidence to accept the alternative hypothesis.

We can compare reality and our decision in the following table (with rows corresponding to reality and columns corresponding to our decision).

|  | reject | don’t reject |
| --- | --- | --- |
| is true | type I error () | correct decision |
| is false | correct decision | type II error () |

We can see that the significance level is also the probability of making a type I error: rejecting a true null hypothesis. The probability of a type II error may be a function of the possible alternative hypotheses.

The hypothesis testing process provides a quantitative way to describe the intuition we can obtain by looking at plots and more intuitive statistics. Since multiple tests have been defined, users must make choices about which test to use in any particular circumstance. This is one of the many aspects of statistics that is more art than science, and there are regular controversies and different schools of thought in the professional statistics community. Thus, it is important to state what test was used and the actual p-value obtained in any inferential statistical analysis.

## Confidence Interval

A *confidence interval* is an estimate of a population parameter that conveys the precision of the estimate. A level of confidence equal to one minus the level of significance () is typically reported as a percentage confidence. The classical (parametric) approach is to include all null hypothesized values for the parameter that would not be rejected with a hypothesis test. A resampling (nonparametric) approach is to use the actual sample data as a best estimate for the population and determine a coverage interval for the estimator. Since people use a variety of different procedures and confidence levels, these whould be reported along with the confidence interval.

## Validity

Hypothesis tests and confidence intervals are only considered valid if the significance level and procedure is chosen **before** the data are collected and analyzed. This is how the United States Food and Drug Administration deals with drug testing for safety and effectiveness. In the examples that follow, we will often consider different procedures and signifance levels as an aid to understanding the concepts. This is often what researchers do in what is often called *exploratory data analysis*. We should be careful, honest, and transparent in reporting any results obtained through exploration.

## Fair Coin Claim

An actual penny was flipped 100 times resulting in 37 heads. Is the penny a fair coin?

## Normal(50, 3) Concrete Block Strength Claim

Acme Concrete claims that the compressive strength of its concrete blocks is normally distributed with a mean of 50 hundred pounds per square inch and a standard deviation of 3 hundred pounds per square inch.

To test the claim, ten concrete blocks are randomly chosen and their compressive strengths (in hundred pounds per square inch) are measured.

acme = tibble(x = c(49.2,53.9,50.0,44.5,42.2,42.3,32.3,31.3,60.9,47.5))

We compare this data to the claimed distribution (the null hypothesis) in a variety of ways.

## Normal Concrete Block Strength Claim

Acme Concrete claims that the compressive strength of its concrete blocks is normally distributed, but does not specify the two parameters.

## Gamma Concrete Block Strength Claim

Acme Concrete claims that the compressive strength of its concrete blocks is distributed as a gamma random variable.

## No Answer Copying Claim

. Two students were given different 20 question multiple choice (4 per question) exams. They claim to have randomly chosen their answers. But their answers were the same on 15 of the questions. The instructor suspects the students of copying off each other. Who is right? (Based on an actual academic integrity case at a liberal arts college.)

## Steel Bar Improvement Amounts

A steel manufacturer has been producing bars with a yield strength that is normally distributed with a mean of 900 MPa and a standard deviation of 50 MPa. The manufacturer hopes that a modified process will increase the mean yield strength of bars to 950 MPa and decrease the standard deviation. The new process is used to produce 15 bars whose yield strengths in MPa are given below.

steel = tibble(x = c(961, 962, 942, 934, 939, 920, 980, 972, 905, 985, 897, 970, 948, 888, 930))

What is the mean yield strength of bars made by the new process? What is the standard deviation yield strength made with the new process? Has the manufacturer’s hopes been fulfilled? (Based on a text book example but is typical of real-world questions.)

## Penny Side Flips

What is the probability of heads when a penny is balanced on its edge and then the surface is jostled? When this was done 100 times, 64 heads appeared.

## Teacher Differences

. Zumbrun surveyed 141 secondary mathematics teachers. She divided the teachers into two groups based on whether the teacher had ever taught a stand-alone high school statistics course. Each teacher was also asked several questions that had quantitative answers (sometimes via a 1-7 Likert-type scale). A summary of some of the results are given in the table below. For which questions were there statistically significant differences between the two groups? (Christina Zumbrun, Attitudes and Beliefs Towards Statistics: What are They and Why Do They Matter to All Teachers? Goshen College Science Speakers Presentation, April 4, 2018)

## Cricket Chirp Rates and Temperature Relationship

Consider the following data of the air temperature in degrees Fahrenheit (t) and cricket chirp rate in chirps per minute (c). What, if any, linear relationship exists between these two variables? How precise are predictions made by the model?

crickets = tibble(  
 t = c(46, 51, 54, 57, 59, 61, 63, 66, 68, 72),  
 c = c(40, 55, 72, 77, 90, 96, 99, 113, 127, 132))

## GapMinder

In February 2006, a Swedish physician and data advocate named Hans Rosling gave a TED talk titled [“The best stats you’ve ever seen”](https://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen) where he presented global economic, health, and development data from the website [gapminder.org](http://www.gapminder.org/tools/#_locale_id=en;&chart-type=bubbles). Describe lifeExp (life expectancy in years) as a function of continent (note that “Americas” includes countries in both North and South America and that Antarctica is excluded), pop (number of people living in the country), and/or gdpPercap (gross domestic product in US dollars).

## World Swim Records

Consider the data in swim.csv: world record times in the 100 meter freestyle long course swimming race. The variable year is the year in which the record was set, time is the record time in seconds, and sex is whether this is the women’s or the men’s record. Only the last instance of a record breaking time is recorded in each year. The information was gathered from https://en.wikipedia.org/wiki/World\_record\_progression\_100\_metres\_freestyle. Find a reasonable model for this data.

## Coal Differences

Random samples of coal were taken from two mines and the heat-producing capacity (in millions of calories per ton) for each specimen was obtained. (This data is from Richard A. Johnson, *Miller & Freund’s Probability and Statistics for Engineers*.)

mine1 = c(8260, 8130, 8350, 8070, 8340)  
mine2 = c(7950, 7890, 7900, 8140, 7920, 7840)

What is the difference between the mean heat-producing capacities of the two mines?

## Facility Usage Differences

A student who used the college recreational facilities was interested in whether there is a difference between the facilities used by men and women. A randomly selected sample of users is summarized in the table. Note that I seem to have forgotten the labels for the table (men vs. women, and facility A vs. facility B), but this actually does not matter for a test of independence.

## MIT Suicides

The number of days between MIT student suicides from 10/8/64 to 6/15/91: 38, 336, 151, 444, 868, 269, 978, 27, 30, 9, 428, 366, 138, 52, 622, 169, 1772, 295, 204, 696, 139, 16, 347, 1, 19, 169, 7, 61, 0, 852, 243, 0. (data obtained from Elaine Chew and Philip Greenspun, Is Suicide at MIT a Poisson Process?, document found at http://philip.greenspun.com/research/suicide-at-mit.pdf.) Is the population exponentially distributed?