Statistics Unit 5

Topic: Statistical Testing - review of the course

- Date: @March 21, 2024
- Lane: *Statistics*

What it covers

- What is statistics? Inference? Prediction?
- How do the subject covered in this class relate to statistics?
- What is the purpose of a statistical test? How is it performed?

Requires

Probability, Optimization, linear Algebra,

Required by

Next classes in the program

What is the purpose of statistical inference & prediction?

Statistics provides techniques for *scientific enquiry* — that means how one goes about proving if something is the case from data. One starts with a premise, called a *hypothesis* — in short an explanation of how one believes their data came about - this takes the form of a *statistical model* — and one wants to test if their actual data fits the model, and hence is consistent with the hypothesis. Since the data is uncertain, (or else statistical tests would not be needed) the results of the test typically are *not* conclusive, but only indicate a degree of support of the evidence (the data) for or against the hypothesis.

In machine learning a “hypothesis” may be just a test of an incremental change in a software application, as in “A/B testing” where typically one modifies, say a website with the hope to improve a measurable quantity, say a customer response rate. The point of the statistical test is actually quite modest; to estimate the chance that a change has occurred. A definitive proof in the face of uncertainty is a formidable undertaking, and often unnecessary for acting on the data.

The reasoning behind a statistical test.

The actual basic statistical hypothesis test has a modest purpose: to indicate the degree to which a measured quantity is not just a spurious result, but has changed from what is considered typical. To estimate the measured quantity of interest, one takes a *sample* of a given size *n* and applies a function to the *n* measurements to reduce them to a *statistic.* This is an application of the techniques in probability modeling and interval estimation learned in this course.

How does one actually go about this? Perhaps one has a presumption that the change in one measurable quantity causes the change in another. “If one marked-up the price of gasoline would people drive less?” Putting aside whether it can be proven that one causes the other; given a model and data to learn it, imagine the simple case of regression, where one wants to know if the change in the independent variable in the model — price — has an effect on the outcome— driving. This becomes a question about the the beta coefficient for that variable. The way this question is posed as a statistical test is to hypothesize the converse: “*Is this beta coefficient actually zero, and if so is the variation seen in it attributable to just noise?”* If not (e.g. this hypothesis is *rejected)* then it supports our initial presumption. This “safe” line of reasoning appears convoluted because it makes minimal assumptions about the reasoning underlying the test.

This is called a “*statistical hypothesis test.”* Specifically, since the test’s premise is that there is no effect, it is called a test of the “*null hypothesis”.* The point of the test is to “reject” the null hypothesis, indicating the test is “significant” and there is evidence for the desired effect. If not, either the experiment lacked enough data, or this line of investigation is going down the wrong track.

Statistical testing is built on statistical estimation, specifically on interval estimation. This illustrates the distinction between *prediction* and *inference*. Testing is tool for inference. Consider the example of regression. The regression optimizes the accuracy of it’s prediction. Given some level of predictive accuracy, *Inference* quantifies the accuracy of the model coefficients, to explain the model performance, and hence how it applies to the “real world”.

Creating a “confidence interval” from data.

Given the data experiment and the associated statistical model, a statistical test reduces to exactly the question of creating an estimation interval from the data. For a statistical test this interval is called the *confidence interval.* There’s an equivalence between statistical tests and interval estimation; For every statistical test there’s a interval estimation task, and vice versa.

In the simplest example, assume there’s a phenomenon one is interested in that typically has a statistic whose value is zero, and based on a sample of *n* measurements one wants to test if it’s deviated One creates a confidence interval as the estimation interval around the observed mean $\overline{x}$ using the standard deviation (”standard error”) of the sample mean:

$$[\overline{x} - 2\,\overline{\sigma}/\sqrt{(n)}\ ,\ \overline{x} + 2\,\overline{\sigma}/\sqrt{(n)}]$$

The $1/\sqrt{(n)}$ term is because one is looking for the deviation of the sample mean, not an individual datum. And the factor of 2 creates an (approximately - the true number is about 1.96 for a normal distribution) *95%* confidence interval. The 95% *critical value* is a standard convention, but one can use different probability intervals, depending on how severe a test is desired. The closer the probability is to *1,* the more challenging the test is. 1 minus this probability is known as the *p-value* ascribed to the test, so a smaller p-value implies a test less likely to mistake a random occurrence for an actual effect.

Alternately one can create a confidence interval by using *bootstrap* resampling, discarding the assumption that the statistic would be normally distributed. By looking at the empirical distribution of the resampled statistic one can specify a confidence interval. This is a good idea when working with complex statistical or other models.

“Failure to reject the null hypothesis” is one kind of failure of several possible ways an experiment may fail. e.g. the model or data may be inappropriate. The same can be said for the opposite, “failure not to reject the null hypothesis.”

Advanced tests

The test just presented is known as a “z-test”, and is one of many variations of how critical regions are created. For small samples, there are adjustments when the sample variance is uncertain. In some circumstances it makes sense to test with an asymmetric region. If there’s an identified alternative, then one can introduce a comparative “alternate hypothesis” to test against, as in A/B testing. Further complications arise when there are many alternate hypotheses to consider. Or one may consider sequential testing that allow early stopping before the full dataset is available. And each test has its Bayesian counterpart. Welcome to the world of statistics!

Readings

Statistical inference is a broad field with many complexities. Ironically the availability of large datasets and vast computational resources change the nature of conventional statistical methods, and tend to simplify the process.

Evans Chapter 5

ISLP_python, Ch.13.1

References

Several recent books in machine learning review statistical methods from a modern viewpoint.

- Gareth James , Daniela Witten , Trevor Hastie, Robert Tibshirani Introduction to Statistical Learning (2023)https://www.statlearning.com/resources-python (”ISLP” 2023)
- Bradley Efron, Trevor Hastie “Computer Age Statistical Inference” (”CASI” 2021) An advanced, readable text addressed to those already conversant in the advances in statistics brought on by computer science and machine learning.