Notes: Statistics

Aaron Trowbridge

Contents

0 Introduction

In the beginning there was a population, and over that population, there was a distribution. As mortals, and not gods, if we want to try to understand this distribution the best we can do is attempt to observe features of the distribution–hopefully at random–and use that data to build a blurry picture of the truth lying behind our perception.

Dramatization aside, this is the principle assumption and goal of statistics: infer knowledge about a hidden (parameterized) distribution by taking random samples and applying mathematical techniques to the data obtained. Usually we are either trying to estimate the *true* hidden parameters or test hypotheses we might have about those parameters.

In these notes I intend to condense the essential methods needed to piece together an actionable representation of the uncertain world around us. Statistical analysis of data is an essential tool in almost every field imaginable: from physics to psychology to finance, and everywhere in between.

Throughout these notes I will include custom visualizations and code snippets, for which I will be using the Julia programming language. Code for everything in this document, including the LaTeX source code, can be found on my github.

Part I

Data: Production, Visualization, and Description

1 Producing Data

There are many ways to get data, not all are equal, and, even more importantly, not all are equally as useful. *Bias* can seep into data in a number of ways depending on how the data is obtained.

The first step in getting data is to decide on the *population* we are interested in, which is typically very large (if it weren't we could just sample the whole population and there would be no need for statistics!).

The next step is to decide what kind of data we are interested in: namely either *numerical* (e.g. height or weight) or *categorical* (e.g. political party or favorite flavor of ice cream). Denoting our *random* variable of interest as X:

Definition 1 (Numerical Variable) *A feature obtained from an observation which has a random continuous numerical value according to the underlying distribution, e.g.* $X \sim N(\mu, \sigma)$*.*

Definition 2 (Categorical Variable) *A feature obtained from an observation which takes it's value from a set of possible categories* $\{x_1, x_2, \ldots, x_k\}$ *and* $P(X = x_i) = p_i$

The final step is to decide how we plan on getting our data; procedures for which can be broken up into two categories: *sampling* and *experimenting*

1.1 Sampling

Since populations are so large, and we can't just issue a census, we need a way to get a representative *sample* (say of size n) of the population, that we can analyze. There are various ways to sample and each comes with it's advantages and disadvantages. Here we list some of the "good" sampling methods:

Simple Random Sample (SRS)

This is the ideal sample and involves enumerating all of the possible members of the population and randomly selecting n numbers in the range of the population size.

Survey Sample

Here n individuals are chosen using a SRS (ideally) and asked to complete a survey, which can result in numerical or categorical data or both. This is a "good" sample if all who receive a survey complete it.

Stratified Random Sample

In this set up the population is segmented into some number of *strata* and then a SRS used to gather data from each.

Multistage Sample

Iterated grouping and random sampling with multiple layers.

Bias, where certain outcomes are systematically favored, can accrue when the sampling method is not random. Here I will list some of the types of bias that can arise:

Convenience Bias

Individuals are chosen based on ease of access.

Response Bias

Individuals are issued a survey and have an option not to complete the survey.

Undercoverage Bias

When some individuals in a population are not accessible due to the design of a survey.

1.2 Experimenting

2 Data Visualization

3 Descriptive Statistics

Definition 3 (Unbiased Estimator) *A statistic (e.g. $\bar{x}$) used to estimate a population parameter (e.g. μ) s.t. $E(\bar{x}) = \mu$*

Part II

Probability Theory

4 Probability

5 Random Variables

6 Expectation

7 Important Inequalities

8 Convergence Results

9 Distributions

9.1 Normal Distribution

9.2 t-Distribution

9.3 χ^2-Distribution

Part III

Statistical Inference

Thus far we have looked at how to extract data from a population, describe the data with summary statistics, and visualize the data with various types of plots. We will assume the obtained data had to come from a distribution that we don't have direct access to – we can only sample from it. Another assumption we will make is that the hidden distribution can be *parameterized*, that is there is a finite set of numbers that perfectly describe the distribution that we just don't know (a seemingly flimsy assumption but one that is necessary and fairly powerful). Now we will turn or attention to the principle task of statistics: inferring knowledge about the underlying distribution's parameters from a sample.

10 Sampling Distribution

Let's assume that the distribution we are after can be described by it's position, or mean μ, and it's standard deviation σ. Which means our random variable $X \sim \mathit{Distribution}(\mu, \sigma)$ will give us values of x_i for the i-th measurement. If we take a sample of size n and calculate the *sample mean*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

The central limit theorem tells us that as n gets large,

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n}), \tag{2}$$

which implies that

$$z = \frac{\bar{x} - \mu}{\sigma\sqrt{n}} \sim N(0, 1) \tag{3}$$

has a standard normal distribution which we can use to easily do calculations.

11 Parameter Estimation: Confidence Intervals

If our goal is to estimate the value of some parameter, the population mean μ for instance, we can take a sample and calculate the sample mean $\bar{x}$, and use the fact stemming from it's distribution (2), that $\mathbb{E}(\bar{x}) = \mu$, to construct a range around $\bar{x}$ s.t. we can be confidant that μ lies within that range. In general a confidence interval will look like

$$\text{estimate} \pm \text{margin of error} \tag{4}$$

One sample z-confidence interval

Given a sample $\{x_1, x_2, \ldots, x_n\}$ taken from a population with mean μ and known standard deviation σ, we can calculate a γ-level confidence interval around $\bar{x}$, given by

$$\boxed{\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}} \tag{5}$$

Here, z^* is the critical value defined implicitly by

$$\gamma = \text{area under standard normal curve between } -z^* \text{ and } z^* \tag{6}$$

For example, if $\gamma = .6 = 60\%$, then $z* = 1$, by the 60-95-99.7 rule.

One sample t-confidence interval

If we don't know σ then all is not lost, we simply use the sample standard deviation

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{7}$$

in place of σ and

$$t = \frac{\bar{x} - \mu}{s\sqrt{n}} \tag{8}$$

in place of z which has a t-distribution, with $\nu = n - 1$ degrees of freedom. The γ-level confidence interval in this case is then given by

$$\boxed{\bar{x} \pm t^* \frac{s}{\sqrt{n}}} \tag{9}$$

where t^* is defined analogously to z^*.

Two sample t-confidence interval

Given two samples of size n_1 and n_2, respectively drawn from two populations with means μ_1 and μ_2, we can find a γ-level confidence interval for $\mu_1 - \mu_2$, given by

$$\boxed{(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{10}$$

where t^* is found using the t-distribution with $\nu = \min\{n_1 - 1, n_2 - 1\}$

12 Hypothesis Testing

Part IV

Statistical Models and Methods

13 Regression

14 Classification