Stats M254 - Lecture 1

Scribe: Chris Hartl

2015-04-01

Review of Statistics and Statistical Inference

"Inference" here is a term of art indicating that we want to infer the value, or possible values, of a particular model parameter. This includes both point estimates of parameters, as well as uncertainty or confidence regarding such values. That is to say

$$\text{Inference} = \text{Point Estimate} + \text{Uncertainty Estimate}$$

There are two complementary approaches, the frequentist and the Bayesian approach, and we will review both.

Example:

Let $X_1, \ldots, X_n$ be i.i.d. ("independent identically distributed": drawn from the same population, and unordered/exchangeable) observations with $X_i \sim N(\theta, \sigma^2)$. How do we estimate the (unknown) parameter θ?

By plugging the data $X_1, \ldots, X_n$ into the joint density function we will be able to perform this kind of inference. Each X_i follows

$$f(x; \theta, \sigma) = (2\pi)^{-1/2} \exp\left\{ -\frac{(x-\theta)^2}{2\sigma^2} \right\}$$

Because the X_i are independent, the joint distribution will be the product of each density. That is:

$$\begin{aligned} f(x_1, \ldots, x_n; \theta, \sigma) &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left\{ -\frac{(x_i-\theta)^2}{2\sigma^2} \right\} \\ &= (2\pi)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2 \right\} \end{aligned}$$

This function $f(x_1, \ldots, x_n; \theta, \sigma)$ is the joint density function of the data points. However, one can also consider $f(\cdot)$ as a function of θ and σ, e.g. $L(\theta, \sigma; X_1, \ldots, X_n) = f(x_1, \ldots, x_n; \theta, \sigma)$. This view of f (or L) is often termed the "likelihood" of the parameters. By *maximizing* L, i.e. (letting $\sigma = 1$)

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, L(\theta, 1; X_1, \ldots, X_n)$$

produces the *maximum likelihood estimate* (MLE) $\hat{\theta}$ of θ. This is the standard *frequentist* approach.

Homework: Show that the MLE of $N(\theta, 1)$ as above is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Under the MLE framework, it can be shown that, if $X_i \sim N(\theta, 1)$, then

$$\bar{X} \sim N(\theta, \frac{1}{n})$$

This provides the ability to form a (95%) confidence interval about this mean:

$$CI = (\bar{X} - 1.96/\sqrt{n}, \bar{X} + 1.96/\sqrt{n})$$

Note that the interpretation of this object is ***not***:

(WRONG) θ has a 95% chance to fall within this CI

The problem with this reasoning is that, from the frequentist perspective, θ is fixed. θ itself has no distribution, you don't get to "pick theta" multiple times and find that 95% of the time it was in the interval. In other words, this interpretation assigns randomness to the *parameter* as opposed to the *sample*!

The correct interpretation is to interperet the *data* as random. This gives the following (correct) interpretation ("principle of repeated sampling"): If we were to repeat this experiment 100 times, and sample n data points from $N(\theta, 1)$, then out of these 100 experiments, we would expect 95 of the resulting confidence intervals to contain the true value of θ. That is, we would expect 95% of *our experiments* to produce a CI which contains the true value of θ.

By contrast, the following is the Baysian approach. The Bayesian approach imagines the parameter θ to be *random*, and specifies the distribution of θ via a prior $p(\theta)$. Then the data density, given a specific value of θ is written as $f(x_1, \ldots, x_n|\theta)$.

However, we want the *joint* density $f(x_1, \ldots, x_n; \theta)$. Using Bayes' theorem:

$$f(x_1, \ldots, x_n, \theta) = \frac{f(x_1, \ldots, x_n|\theta)p(\theta)}{p(x_1, \ldots, x_n)} \propto f(x_1, \ldots, x_n|\theta)p(\theta)$$

,

where $p(x_1, \ldots, x_n)$ is the marginal density of $X_1, \ldots, X_n$. It is also, in this case, a normalization constant, just a think that ensures that the *total* probability is 1.

From the Bayesian approach, we want to understand the conditional distribution of θ *given* the data: $p(\theta|X_1, \ldots, X_n)$. Bayes' rule again provides the means to do this, writing:

$$p(\theta|x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n; \theta)}{p(x_1, \ldots, x_n)} \propto f(x_1, \ldots, x_n|\theta)p(\theta) = \text{likelihood} \times \text{prior}$$

It is the frequentist approach to maximize the *likelihood*. The Bayesian approach maximizes the *posterior*.

In this example, we have our $X_1, \ldots, X_n$ i.i.d with $X_i \sim N(\theta, 1)$. Let $\theta \sim N(\alpha, 1)$. For instance, imagine θ were the mean expression for some gene of interest. If we know, that mean gene expression across all genes is roughly α, we might use that as prior knowledge about θ. Plugging into our equation above, and not worrying about the multiplicative constants:

$$\begin{aligned}\text{posterior} \propto \text{prior} \times \text{likelihood} &\propto \exp\left\{-\frac{(\theta-\alpha)^2}{2}\right\} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2\right\} \\ &= \exp\left\{-\frac{(n+1)\theta^2 - 2(\alpha+\sum_{i=1}^{n}X_i)\theta + \alpha^2 + \sum_{i=1}^{n}X_i^2}{2}\right\} \\ &\propto \exp\left\{-\frac{n+1}{2}\left(\theta - \frac{\alpha+\sum_{i=1}^{n}X_i}{n+1}\right)^2\right\}\end{aligned}$$

This means that the posterior distribution is normal, that is

$$\theta|X_1, \ldots, X_n \sim N\left(\frac{\alpha+\sum X_i}{n+1}, \frac{1}{n+1}\right) = N\left(\frac{1}{n+1}\alpha + \frac{n}{n+1}\bar{X}, \frac{1}{n+1}\right)$$

This means that the prior information is *weighted* into the posterior information. It is, in this case, exactly a weighted average between the prior mean α and the data mean $\bar{X}$. Very clearly, *because* the mean of this posterior distribution is $\frac{1}{n+1}\alpha + \frac{n}{n+1}\bar{X}$ it must be that

$$\underset{\theta}{\text{argmax}}\, \text{posterior}(\theta) = \frac{\alpha + n\bar{X}}{n+1}$$

Why would one want to take the Bayesian approach?

i. Small sample size (say $n = 3$)

ii. Strong expectation about the parameter θ

iii. *Whenever* you would want to incorporate subjective opinion or prior knowledge

iv. Convenience (low sample size, or ease of computation due to a *conjugate prior*)

object	Frequentist	Bayesian
parameter (θ)	fixed	random
data (X)	random	random

Repeated Sampling

In the previous example, we had a clear model, the normal distribution. However, in most cases we do not know the underlying distribution of the X, and the data may not "look" like they follow any of the standard distributions. How do we perform statistical inference without knowing the underlying data generating distribution?

There are many approaches, of which I will discuss two: jackknife and bootstrap.

Jackknife (also known as Leave-One-Out or LOO):

Given the data points $X_1, \ldots, X_n$. Let $S_i = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n\}$ be the dataset with the ith data point left out. Let $\hat{\theta}_i$ be the estimator generated from S_i. There being n datasets, there are therefore n estimates $\hat{\theta}_1, \ldots, \hat{\theta}_n$, which can provide a measure of uncertainty about the statistic $\hat{\theta}$.

Bootstrap:

Here, we regard $X_1, \ldots, X_n$ as a *population*, and draw samples from it, with replacement. Let $S_i = \{X_1^{(i)}, \ldots, X_n^{(i)}\}$ be a sample of n data points draw, with replacement, from the original set of data points, and the estimate $\hat{\theta}_i$ be the estimator generated from the sample S_i. The *bootstrap sample* $S_1, \ldots, S_b$ produces b estimates $\hat{\theta}_1, \ldots, \hat{\theta}_b$.

This collection of estimates can then be used to produce *bootstrap estimates* of θ and its uncertainty. Specifically, the *bootstrap estimate* of θ is $\langle\theta\rangle_b = \frac{1}{b}\sum_{i=1}^{b}\hat{\theta}_i$. In addition, the *bootstrap estimate of the variance of* $\hat{\theta}$ is given by

$$\hat{\text{var}}(\hat{\theta}) = \frac{1}{b-1}\sum_{j=1}^{b}\left(\hat{\theta}_j - \langle\theta\rangle_b\right)$$

The bootstrap here is an attempt to replicate the central limit theorem. The CLT states that, under certain conditions about independence and existence of certain moments, any distribution satisfying those conditions with mean μ follows:

$$\sqrt{n}(\bar{X} - \mu) \to_d N(0, 1)$$

With the boostrap, we have $\sqrt{n}(\langle\theta\rangle_b - \hat{\theta})$ which mimics $\sqrt{n}(\hat{\theta} - \theta)$. Note, especially, the dependence on n on the bootstrap estimate, and not b!

In addition, the collection $\hat{\theta}_i, \ldots, \hat{\theta}_b$ can be used to provide nonparametric confidence intervals. By sorting these from smallest to largest, the 95% confidence interval will correspond to the 2.5% and 97.5% percentiles of this sorted set. See Freedman (1981) and Freedman & Bickel (1983) for discussion about necessary conditions for the effectiveness of a bootstrap estimate.