STATISTICAL INFERENCE

ETHAN LEVIEN

Contents

1. Estimators

The basic question of statistical inference can be framed as follow: We have a statistical model for a variable Y, e.g.

$$Y \sim \text{Normal}(\mu, \sigma), \tag{1}$$

but we don't know the parameters (in this case μ and σ). Why do care about the parameters? We need knowledge of these in order to make predictions.

Now imagine we have some samples of Y (either from simulations or data), $Y_1, Y_2, \dots, Y_N$. **What are our best estimates of these parameters, and how accurate are they?** We've already tackled the first part of this problem for a number of distributions. The solution relies on two key observations:

(1) Both μ and σ can be represented as means over the distribution of Y. For example $\mu = \mathbb{E}[Y]$.
(2) If we have enough samples the sample average should be close to the actual average. That is, $1/N \sum_{i=1}^N Y_i \approx \mathbb{E}[Y]$. The central limit theorem tells us how accurate this estimate is.

To make this procedure more precise and generalizable, let's introduce some definition and notation. We will let $\hat{\theta}$ denote an **estimator** of a parameter θ from a sample if $\hat{\theta}$ is some function of our sample which is meant to approximate θ. For example

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^N Y_i \tag{2}$$

is an estimator of μ in the Normal model. **Remember:** the estimator is a property of the data. That is, $\hat{\mu}$ **depends on the specific data we collect** or simulation we run. However, in classical statistics, it is meant to approximate something, μ which is a property of our statistical model. μ **does not depend on the data**.

Since $\hat{\mu}$ depends on the data, different replications of our sample will generate different values of $\hat{\mu}$. We can therefore think of $\hat{\mu}$ as a random variable. We call the distribution of $\hat{\mu}$ over many replications of our data the **sample distribution**. This is different than the distribution of Y, rather it is the distribution of $Y_1, Y_2, \dots, Y_N$. For example, Y follows the Normal distribution given above, the sample distribution of $\hat{\mu}$ is

$$\hat{\mu} \sim \text{Normal}(\mu, \sigma/\sqrt{N}) \tag{3}$$

1.1. Standard errors.

A natural way to quantify the uncertainty in our estimate is the standard deviation of the $\hat{\theta}$ under the sample distribution. We call the resulting quantity the **standard error**, which is our **estimate** of the standard deviation of the sample distribution For the Normal model,

$$\text{se}(\hat{\mu}) = \frac{\hat{\sigma}}{\sqrt{N}}. \tag{4}$$

Date: April 2022.

This tells us how much our estimate will vary between different experiments (or surveys/simulations). The 95% **confidence interval**, or $95\% CI$ is the interval

$$[\hat{\theta} - 1.96\text{se}(\hat{\theta}), \hat{\theta} + 1.96\text{se}(\hat{\theta})] \tag{5}$$

In classical statistics, the interpretation of CI is subtle: it is not saying there is 95% chance or parameter will be in this interval. To understand why, note that the parameter has a 95% chance to be in the interval

$$[\theta - 1.96\text{std}(\theta), \theta + 1.96\text{std}(\theta)] \tag{6}$$

The standard errors and confidence intervals can help us decide how much data we need to collect.

Example 1. *Using standard errors to design an experiment*

There is an alternative way to think about the CI: as a measure of our belief in the parameter value. This interpretation is more natural for me, and we will discuss how to formalize it in the context of Bayesian statistics.

Exercise 1: *Standard errors for binomial model*

1.2. Bias and consistency.

There must be some properties we would like the estimator to have. At a minimum, it should be in some way informed by the data. We express this with the assumptions that: The more data we have (e.g. the larger N) the closer we expect $\hat{\theta}$ to be to the true value. What do we mean by "closer" when we are talking about random things. This turns out to be technical, but for our purposes we will say $\hat{\theta}$ is **consistent** if

$$\mathbb{E}[\hat{\mu}] \to \mu \ \text{ and } \text{se}(\hat{\theta}) \to 0 \quad \text{as} \quad N \to \infty. \tag{7}$$

To better understand the notation of consistence, let's consider two rather silly ways to estimate q in a Bernoulli distribution. Let $\hat{q}_1$ and $\hat{q}_2$ be two other estimators of q defined by

$$\hat{q}_1 = \frac{Y}{N} + \frac{1}{N} \tag{8}$$

$$\hat{q}_2 = \frac{y_1 + y_2}{2} \tag{9}$$

Example 2. *Understanding consistency*

This example demonstrates that consistency is not the only property we look for in an estimator, since $\hat{q}_1$ seems inferior to $\hat{q} = Y/N$. To this end, we say that an estimator $\hat{\theta}$ is **unbiased** for some N (not just very large N), the average over the sample distribution is equal to the actual value under the model distribution; that is,

$$\mathbb{E}[\hat{\theta}] = \theta. \tag{10}$$

Exercise 2: *Understanding bias*

2. Maximum Likelihood

Sometimes it is quite clear what the estimator for a parameter should be, for example, this is the case for q in the Bernoulli distribution. However, we will find this is not always the case, so it is useful to have a **more systematic way of finding estimators.**

Recall that the probability distribution for the binomial distribution is

$$p(Y) = \binom{n}{Y} q^Y (1-q)^{n-Y} \tag{11}$$

In statistics, we sometimes call this the **likelihood**. More generally, the likelihood is defined as the probability we say a data set as a function of the parameters.

Equation (11) tells us how likely it is to observe k YES among n people surveyed. Then, it seems reasonable that this number should not be very small, since that would mean our survey results are an anomaly. More generally, the larger $\mathbb{P}(Y|q)$ is the more likelihood our results are. This suggests one a way to estimate determine q: We can take as our estimate $\hat{q}$ the value which makes $\mathbb{P}(Y|q)$ largest. In other words, we are finding the value of q which makes the data the most likely, and we will call this the **maximum likelihood estimate**.

You can do this using calculus (if you know how, I suggest you give it a try) to determine that the value of q which makes (11) largest is

$$\hat{q}_{\text{MLE}} = \frac{Y}{n} \tag{12}$$

For a Normal distribution with mean and variance μ and σ, the MLE estimators are the usual sample mean and standard deviation which we have already been exposed to.

3. Estimators for parameters in a regression model

Consider the example of a clinical trial conducted as follows. Suppose N people participate in the trial, and are randomly assigned to the the control group (C) and treatment group (T) with probability $1/2$. People in T are given a drug whose effects is measure by a percent. We can model the distribution of blood pressure before and after treatment as

$$Y_C \sim \text{Normal}(\mu_C, \sigma) \tag{13}$$

and

$$Y_T \sim \text{Normal}(\mu_T, \sigma). \tag{14}$$

Our model for an individuals response can be framed as a regression model

$$Y = (\mu_T - \mu_C)X + \mu_C + \epsilon \tag{15}$$

where $X = 1$ if someone is in the treatment group and

$$\epsilon \sim \text{Normal}(0, \sigma_\epsilon). \tag{16}$$

How would we estimate μ_T, μ_C and σ from a sample? Assuming we know the group each patient has been assigned to, we observe that

$$\mathbb{E}[Y|X=1] = \mu_T, \quad \sqrt{\text{var}(Y|X=1)} = \sigma_\epsilon. \tag{17}$$

This means we can estimate these from sample averages of Y and X (and similar for $X = 0$). You might be tempted to say that the variance of Y is also σ_ϵ – but this is false! Why? (think back to the previous note when we looked at the marginal distribution of Y)

This means the sample distribution of the mean of the treatment group is

$$\hat{\mu}_T \sim \text{Normal}(\mu_T, \sigma/\sqrt{N}) \tag{18}$$

If $\Delta\mu = \mu_T - \mu_C$, then as estimator $\Delta\mu$ is

$$\Delta\hat{\mu} = \hat{\mu}_T - \hat{\mu}_C \tag{19}$$

which is really the slope of the regression line. The sample distribution of $\Delta\hat{\mu}$ is also Normal:

$$\Delta\hat{\mu} \sim \text{Normal}(\Delta\mu, \sigma_\epsilon/\sqrt{N}) \tag{20}$$

3.1. Covariance and correlations coefficient.

Now let's consider the general regression model

$$Y = aX + b + \epsilon \tag{21}$$

where X may be a continuous variable. **How would we estimate a?**

To understand how this can be done, we start by defining the covariance:

$$\text{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{22}$$

What is $\text{cov}(X, X)$?

$$\text{cov}(X, X) = \mathbb{E}[XX] - \mathbb{E}[X]\mathbb{E}[X] = \mathbb{E}\left[X^2 - \mathbb{E}[X]^2\right] = \text{var}(X). \tag{23}$$

If X and Y are independent, the covariance will be zero, it is possible for the covariance to be zero without the variables being independent.

Just like standard deviation and mean, estimates of covariance are obtained by replacing $\mathbb{E}$ with the sample average: That is, if we have samples $(x_1, y_1), (x_2, y_2), \ldots,$

$$\mathbb{E}[XY] \approx \frac{1}{N}\sum_i x_i y_i \tag{24}$$

In python, you can compute the covariance

```
> np.cov(x,y)[0,1]
```

The reason for the $[0, 1]$ is that the covariance function in numpy actually computes a 2D array (a Matrix), where the off diagonal entries are the covariance.

Example 3. *Covariance vs. independence*

3.2. **Linear regression and least squares.** For the linear regression model, we can show that

$$\mathrm{cov}(X,Y) = a\sigma_x^2 \tag{25}$$

Since $\mathrm{cov}(X,Y)$ and σ_x are both things that can be computed from a sample, this suggests an estimator for the slope variable

$$\hat{a} = \frac{\frac{1}{N-2}\sum_i (y_i - \hat{\mu}_y)(x_i - \hat{\mu}_x)}{\hat{\sigma}_x^2} \tag{26}$$

The $N-2$ comes from the fact that we need at least two points to fit our regression model. We can rewrite this as

$$\hat{a} = \frac{\sum_i (y_i - \hat{\mu}_y)(x_i - \hat{\mu}_x)}{\sum_i (x_i - \hat{\mu}_x)^2} = \sum_i \left[\frac{(y_i - \hat{\mu}_y)}{(x_i - \hat{\mu}_x)}\right]^2 \frac{(x_i - \hat{\mu}_x)}{\sum_i (x_i - \hat{\mu}_x)^2} \tag{27}$$

This is weighted average of the rise/run between different points and the mean. Notice that if there are only two values of x, it reduces to the formula in the previous section. We can also show that

$$\hat{b} = \hat{\mu}_y - \hat{a}\hat{\mu}_x \tag{28}$$

$\hat{a}$ and $\hat{b}$ are also called the **least squares estimator**, because it happens to be the values of a and b which minimize the squared distance to the estimated regression line.

Since the samples are Normally distributed around the line with variance σ, we can estimate σ as the sample variance of the difference between our data and the line $\hat{b} + x\hat{a}$. However, since we are replacing the actual values of a and b with the estimations, we need to account for this in our estimation of the uncertainty and divide by $N-2$ instead of N.

You never need to work with these formulas directly, as there is a Python package which does the computations for us. The code for fitting a linear regression model is

```
> X = sm.add_constant(x)
> model = sm.OLS(y,X)
> results = model.fit()
> print(results.summary())
```

The following example illustrates the use of this

Example 4. *Working with stats models*

Exercise 3: *More working with stats models*