STATISTICAL INFERENCE

ETHAN LEVIEN

Contents

1. Statistical inference and maximum likelihood

Let's imagine we do in fact conduct a survey of $n = 20$ students and find $k = 17$ students respond YES to the question "do you identify as a male?" Can we make this a little more precise? Remember, any time we draw a conclusion we need a model. What is a model for the number of students who respond yes to the survey. Assuming that the samples are independent, we can treat y_i as a Bernoulli distribution. Then, the a number of people, k, who responded saying they are republicans follows a Binomial distribution,

$$Y \sim \text{Binomial}(n, q) \tag{1}$$

where n is the number of students in our survey and q is a parameter.

Recall that the probability distribution for the binomial distribution is

$$p(Y|q) = \binom{n}{Y} q^Y (1-q)^{n-Y} \tag{2}$$

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.

1.1. MLE, bias and consistency.

Equation (**??**) 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 (**??**) largest is

$$\hat{q}_{\text{MLE}} = \frac{Y}{n} \tag{3}$$

MLEs are very useful, but as we learn later on, they are only one of many ways to estimate a parameter in our model. Any number $\hat{q}$ which we use to approximate the parameter q is an **estimator**.

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 wouldn't want to set $\hat{q} = 1/2$ based solely on our intuition). The more data we have (e.g. the larger n) the closer we expect $\hat{q}$ to be to the true value. To make this precise, we define an estimator $\hat{q}$ to be **consistent** if $\hat{q}$ converges to q as n grows. But what does converges mean when we are dealing with random variables? We can understand this through simulations:

Example 1. *Plot $\hat{q}_{\text{MLE}}$ as a function of n for $n = 10, 20, 50, 100, 1000, 5000, 10000$.*

Solution:

```
> k_range = [5,10,20,50,100,1000,5000,10000]
> plt.plot(k_range,[np.random.binomial(k,0.3)/k + 1/k for k in k_range],"+")
> plt.plot(k_range,[np.random.binomial(k,0.3)/k + 1/k for k in k_range],"+")
> plt.plot(k_range,[np.random.binomial(k,0.3)/k + 1/k for k in k_range],"+")
```

Date: April 2022.

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

$$\hat{q}_1 = \frac{k}{n} + \frac{1}{n} \tag{4}$$

$$\hat{q}_2 = y_i \tag{5}$$

Exercise 1. *Are these consistent or not? Generate simulations to support your result.*

Solution:

```
> k_range = [5,10,20,50,100,1000,5000,10000]
> plt.plot(k_range,[np.random.choice([0,1],p = [1-0.3,0.3]) for k in k_range],"+")
> plt.plot(k_range,[np.random.choice([0,1],p = [1-0.3,0.3]) for k in k_range],"+")
> plt.plot(k_range,[np.random.choice([0,1],p = [1-0.3,0.3]) for k in k_range],"+")
> plt.plot(k_range,[np.random.binomial(k,0.3)/k + 1/k for k in k_range],"o")
> plt.plot(k_range,[np.random.binomial(k,0.3)/k + 1/k for k in k_range],"o")
> plt.plot(k_range,[np.random.binomial(k,0.3)/k + 1/k for k in k_range],"o")
```

This exercise demonstrates that consistency is not the only property we look for in an estimator, since $\hat{q}_1$ seems inferior to $\hat{q}_{\rm MLE}$. To this end, we say that an estimator is biased if an estimator is, on average, equal to the value of q used to generate the data. In other words, if we run many simulations, or take many different samples from a population and compute the estimator, then we should get the true value of q.

Exercise 2. *Determine whether $\hat{q}_1$ and $\hat{q}_2$ are biased using simulations.*

1.2. **Standard errors.** At this point, you understand that $\hat{q}_{\rm MLE}$, like all estimators, depends on the data we collect. If we had collected different data, e.g. surveyed a different class, we would get a different $\hat{q}_{\rm MLE}$. How much will $\hat{q}_{\rm MLE}$ vary between samples? In classical statistics, we measure accuracy using the standard error, denoted ${\rm se}(\hat{q})$. Roughly speaking, if we performed many experiments and measured $\hat{q}$, the measurements will typically differ by ${\rm se}(\hat{q})$.

$${\rm se}(\hat{q}) = \sqrt{\frac{\hat{q}(1-\hat{q})}{n}} \tag{6}$$

Example 2. *Run simulations to determine test this formula.*

2. Inference for a Normal distribution

Suppose have $y_1,\dots,y_n$ from a variable which follows a Normal distribution, that is

$$y_i \sim {\rm Normal}(\mu,\sigma) \tag{7}$$

What is our best estimate of μ and σ?

from a Normal distribution with mean and variance μ and σ, the MLE estimators are

$$\hat{\mu}_{\rm MLE} = \frac{1}{n}\sum y_i \tag{8}$$

and

$$\hat{\sigma}_{\rm MLE} = \sqrt{\frac{1}{n-1}\sum(y_i - \hat{\mu})^2} \tag{9}$$

Exercise 3. *Show with simulations that these are consistent and unbaised.*

3. Hypothesis testing

In statistics, we often infer parameters, such as q, not because we are interested in specific values, but rather because we would like to use them to make a decision. For example, whether a candidate drug is worth moving to the next step in clinical trials. Or perhaps whether there is gender bias in a given class or field. This problem is often framed in terms of **hypothesis testing**, in which we assign a probability to a particular hypothesis or its converse. In the context of a Bernoulli random variable, we might want to decide whether we can rule out $q = q_0 < 1/2$ – that is, the samples being fair – given our data. One way to access this is with a p-**value**. There many ways to define a p-value, but lets focus on the case where we are interested in understanding whether a drug has an effect. We have a control group whose response is y_c who is not treated and a treatment group y_t. We then look at the difference $Y = y_c - y_t$. Now uppose that the drug has no effect, then the mean of this should be zero, so testing o see if a drug had an effect essentially amounts to testing if the mean of Y is not zero – this is our null hypothesis.

For this problem, p-value is the chance that the actual value

Exercise 4. *Estimate the p-value using Monte Carlo simulations.*

We can understand the p-value using the Normal approximation to

Exercise 5. *Plot the p-value as a function n.*