STATISTICAL INFERENCE

ETHAN LEVIEN

Contents

1. Estimators

Previously, we say that if we survey N people in a population for which each individual has a chance q of answering YES (for example, because they are an unbiased sample from a larger population for which q is determined).

Question: If we find that Y people respond YES, what is our best estimate of the true value of q?

This should be Y/N, since for an individual response $x_i = 0, 1$, $\mathbb{E}[x_i] = q$ and

$$\bar{x}_i = Y/N \approx \mathbb{E}[x_i] = q. \tag{1}$$

At the same time, we expect that for any particular sample of the population

$$\frac{Y}{N} \neq q \tag{2}$$

since there is always a chance that we happen to sample more or less people who answer YES. In this case we call Y/N an **estimator** of q, and write $\hat{q} = Y/N$.

Can we quantify how much this two quantities will typically differ? This is related to the idea of uncertainty quantification, a central topic in statistics. The idea is that we want to quantify how confident we are in something we've inferred.

1.1. Standard errors.

In classical statistics, we measure accuracy using the standard error, denoted $\mathrm{se}(\hat{q})$. The standard error is the standard deviation of $\hat{q}$ taken over different replications of our experiment. If we are, say, flipping a coin and tallying the result to obtain Y, it is clear what this means. If Y is the vote share from a one-off election, then is becomes a bit puzzling to think about replicating the experiment. Fortunately, this is a philosophical problem, not a mathematical one. Mathematically, we can always define $\mathrm{se}(\hat{q})$ *within the context of our model* as

$$\mathrm{se}(\hat{\theta}) = \sqrt{\mathrm{var}(\hat{\theta}(Y))} \tag{3}$$

where the variance is taken over the distribution of Y. That is, we use the probability distribution $P(Y)$ to compute this variance. Roughly speaking, if we performed many experiments and measured $\hat{q}$, the measurements will typically differ by $\mathrm{se}(\hat{q})$.

Exercise 1: *Standard errors for binomial model*

Exercise 2: *Experimental design*

Date: April 2022.

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 wouldn't want to set $\hat{q} = 1/2$ based solely on our intuition). We express this with the assumptions that: 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 convergence mean when we are dealing with random variables? This turns out to be technical, as there are different things this can mean. For our purposes, we can understand converges as

$$\mathrm{se}(\hat{\theta}) \to 0 \quad \text{as} \quad N \to \infty. \tag{4}$$

Example 1. *Understanding consistency*

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

$$\hat{q}_1 = \frac{k}{n} + \frac{1}{n} \tag{5}$$

$$\hat{q}_2 = y_i \tag{6}$$

Exercise 3: *Understanding consistency*

This exercise demonstrates that consistency is not the only property we look for in an estimator, since $\hat{q}_1$ seems inferior to $\hat{q}_{\mathrm{MLE}}$. 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 4: *Understanding bias*

1.3. **MLE.** Recall that the probability distribution for the binomial distribution is

$$p(Y) = \binom{n}{Y} q^Y (1-q)^{n-Y} \tag{7}$$

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.

1.4. **Maximum Likelihood.** Equation (7) 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 (7) largest is

$$\hat{q}_{\mathrm{MLE}} = \frac{Y}{n} \tag{8}$$

MLEs are very useful, but as we learn later on, they are only one type of estimator.

2. Inference for a Normal distribution

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

$$y_i \sim \mathrm{Normal}(\mu, \sigma) \tag{9}$$

What is our best estimate of μ and σ?

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

$$\hat{\mu}_{\mathrm{MLE}} = \frac{1}{n} \sum y_i \tag{10}$$

and

$$\hat{\sigma}_{\mathrm{MLE}} = \sqrt{\frac{1}{n-1} \sum (y_i - \hat{\mu})^2} \tag{11}$$

Exercise 5: *Consistency of MLE for Normal distribution*

3. Hypothesis testing

In statistics, we might 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. This problem is often framed in terms of **hypothesis testing**, in which we assign a probability to a particular hypothesis or its converse.

Suppose for example, that we conduct a clinical trial as follows. Two groups of N people, the control group (C) and treatment group (T), are randomly selected from the population. People in T are given a drug which reduces their blood pressure and those control group are given a placebo. We can model the distribution of blood pressure before and after treatment as

$$Y_C \sim \text{Normal}(\mu_C, \sigma) \tag{12}$$

and

$$Y_T \sim \text{Normal}(\mu_T, \sigma) \tag{13}$$

This means or model for the sample distributions of the means are

$$\hat{\mu}_T \sim \text{Normal}(\mu_T, \sigma/\sqrt{N}) \tag{14}$$

If $\Delta\mu = \mu_T - \mu_C$ the sample distribution of $\Delta\mu$ is also Normal. We are ultimately interested in moving the drug to the next phase of a clinical trial, which means determining if $\Delta\mu = 0$.

The approach to this problem of frequentist statistics is to ask: How likely is it that we would observe an effect at last as large as we did if the null hypothesis was true. The p-value in this context is the chance we

$$p_v = P(\Delta\hat{\mu}/\text{std}(\Delta\hat{\mu}) > \Delta\hat{\mu}|\mu_C = \mu_T) \tag{15}$$

If the p-value is very small, then it is highly unlikely we would have observed what we did when the null hypothesis was true. In this case, we can REJECT the null hypothesis as false. Usually some threshold is set for this, and if the p_v is below that threshdold we say our result in statistically significant.

Exercise 6: *p-values*

We can understand statistical significance in terms of statdard errors as well.

3.1. Problems with p-values, hypothesis testing and statistical significance.

Despite the widespread use of p-values, classical hypothesis testing and statistical significance, these concepts have some problems. This does not mean they are not useful, rather it is important to understand how they might be applied in appropriately in practice.

First, typically the null hypothesis is never true, that is it is never the case that two subpopulations are exactly equal – that is, that there is no effect. If we have enough data, we can alsmost always rule out the null hypothesis.

Exercise 7: *Behavior of p-values in N and effect size.*

A major issue in who statistical significance is used in practice, is that is can create a selection bais in the published liturature, where effects sizes are almost always over estimates.

Exercise 8: *Bias in the liturature*

Finally, a philosophical problem with statistical signifiance is that te difference between statistically signficant.

Exercise 9: *Problems with statistical signficance*