27 2025-03-31 | Week 11 | Lecture 26

27.1 Point estimation

Section 6.1 Suppose the fraction of people in Hawaii who like broccoli is p. This is an unknown parameter. We don't know what p is, but we would like to estimate it. In principle, we could ask everyone in the state, but that would be impractical. Instead, we take a random sample: we choose randomly n individuals and ask each of them whether they like broccoli or not. Let

$$\widehat{p} := \text{ fraction of our sample who like broccoli}$$

Suppose we sample $n = 100$ individuals and 20 of them like broccoli. Then

$$\widehat{p} = \frac{20}{100} = .2$$

Here, $\widehat{p}$ is an example of a *point estimate* of p.

Definition 83 (Point estimate and point estimator)**.** In statistics, we think of "data" as consisting of n IID random variables $X_1, \ldots, X_n$, called a *random sample*.

(i) A *point estimator* is any function $F(X_1, \ldots, X_n)$ of a sample.

(ii) A *point estimate* $\widehat{\theta}$ is any single number, computed from the data, which can be regarded sensible value of some unknown parameter θ.

In general, an *estimator* is a function of the sample, while an *estimate* is the realized numerical value of the estimator that is computed after the sample is actually observed. The standard convention is to refer to point estimates and point estimators by the same letter $\widehat{\theta}$.

For our broccoli example, the data consists of n random variables $X_1, \ldots, X_n$ were

$$X_i = \begin{cases} 1 & : \text{ i-th sampled person likes broccoli} \\ 0 & : \text{ they don't} \end{cases}$$

Our point estimate is $\widehat{p} = 0.2$ and the point estimator is the function $\widehat{p} = \frac{X_1 + \ldots + X_n}{n}$.

Taking a step back, we can think of our estimate as

$$\widehat{p} = p + \text{error of estimation}$$

It is perhaps clear that if we increase the sample size n, then the estimation error should decrease, so that

$$\widehat{p} \to p.$$

But we would like to precisely quantify the accuracy of the estimate. This brings us to the next subject.

27.2 Confidence intervals

Section 7.1

Definition 84 (confidence interval)**.** An interval $(\widehat{p} - \epsilon, \widehat{p} + \epsilon)$ is called *95% confidence interval* for the unknown parameter p if $\epsilon > 0$ is chosen large enough that

$$\mathbb{P}\left[|\widehat{p} - p| < \epsilon\right] \geq .95 \tag{29}$$

In words, the interval contains the true parameter with probability at least .95.

Can we come up with a 95% confidence interval for our broccoli poll? We already have $\widehat{p} = .2$. But we don't yet know how big ϵ needs to be. (Of course we want to pick the smallest ϵ possible, as that would give us a more informative confidence interval). To figure out how big ϵ must be, we'll have to use the central limit theorem again.

Formally, let

$$\widehat{p} = \frac{X_1 + \ldots + X_n}{n}$$

where

$$X_i = \begin{cases} 1 & : \text{ i-th sampled person likes broccoli} \\ 0 & : \text{ they don't} \end{cases}$$

Then

$$\widehat{p} \sim \text{Bin}(n, p)$$

We don't know what p is, but we can still say that by the central limit theorem,

$$\widehat{p} \approx p + \frac{\sigma}{\sqrt{n}} Z \tag{30}$$

where $Z \sim \mathcal{N}(0,1)$ and $\sigma^2 = \text{Var}(X_1) = p(1-p)$. We also don't know what σ^2 is, since it depends on the unknown parameter p. That's okay, we'll cope. It will be enough to observe that for any vale of $p \in [0,1]$, we have

$$\sigma^2 = p(1-p) \leq \frac{1}{4}$$

and therefore that

$$\sigma \leq \frac{1}{2}. \tag{31}$$

We will also use the following fact (and future homework problem):

$$\boxed{\mathbb{P}\left[|Z| < u\right] = 2\Phi(u) - 1} \tag{32}$$

By Eq. (30),

$$|\widehat{p} - p| = \frac{\sigma}{\sqrt{n}} |Z|$$

Therefore

$$\begin{aligned} \mathbb{P}\left[|\widehat{p} - p| < \epsilon\right] &\approx \mathbb{P}\left[\frac{\sigma}{\sqrt{n}} |Z| < \epsilon\right] \\ &= \mathbb{P}\left[|Z| < \frac{\epsilon\sqrt{n}}{\sigma}\right] \\ &\geq \mathbb{P}\left[|Z| < \frac{\epsilon\sqrt{n}}{1/2}\right] \qquad \text{(since } \sigma \leq 1/2\text{).} \\ &= \mathbb{P}\left[|Z| < 2\epsilon\sqrt{n}\right] \\ &= 2\Phi(2\epsilon\sqrt{n}) - 1 \end{aligned}$$

In conclusion we have:

$$\boxed{\mathbb{P}\left[|\widehat{p} - p| < \epsilon\right] \geq 2\Phi(2\epsilon\sqrt{n}) - 1} \tag{33}$$

Thus in order for the left hand side to be greater than .95, it will suffice for

$$2\Phi(2\epsilon\sqrt{n}) - 1 \geq 0.95$$

or equivalently

$$\Phi(2\epsilon\sqrt{n}) \geq 0.975.$$

Using a computer, we can determine that $\Phi(u) \geq 0.975$ whenever $u \geq 1.96$. Thus, we need

$$2\epsilon\sqrt{n} \geq 1.96.$$

For our problem, $n = 100$, so $\sqrt{n} = 10$, so we need ϵ to be at least as big as

$$\epsilon \geq 1.96/20 = 0.098.$$

Thus, our confidence interval is

$$(\widehat{p} - \epsilon, \widehat{p} + \epsilon) = (0.2 - 0.098, 0.2 + 0.098) = (.102, .298).$$

We conclude that with 95% confidence, the true proportion of people in Hawaii who like broccoli is between 10.2% and 29.8%. Our point estimate is 0.2, and our *margin of error* is 0.098.

If we had sampled more people, our margin of error would be smaller. How many people would we have to sample to reduce the margin of error to 0.05? What about to 0.01?