An example of interval estimation

Problem 47 from Section 6.7 of Durrett's book:

> 47. Of the first 10,000 votes cast in an election, 5,180 were for candidate A. Find a 95% confidence interval for the fraction of votes that candidate A will receive.

Solution. We try to mimic the reasoning in the voltage measurement example:

1. We would like to estimate the fraction p of votes for candidate A among all the votes cast in the election. Whether the vote of a randomly picked person is for candidate A or not can be modelled with a Bernoulli random variable X with parameter p (why?).

2. The available statistical evidence is the votes of 10,000 individuals, that is, a random sample of size 10,000 from the entire population of those who cast votes. This can be modelled as a collection

$$X_1, X_2, \dots, X_{10000}$$

of 10,000 i.i.d. Bernoulli random variables each with parameter p.

3. A reasonable point estimator for p is the fraction

$$\widehat{p} = \frac{X_1 + X_2 + \cdots + X_{10000}}{10000}$$

of the sampled individuals who vote for candidate A. In our model, this is a random variable. Its observed value is $5180/10000 = 0.5180$

4. What can we say about the distribution of $\widehat{p}$?

   - Since $\widehat{p}$ is the average of a relatively large number of i.i.d. random variables, by the CLT, the distribution of $\widehat{p}$ is approximately normal.
   - The parameters of the approximating normal distribution are the mean and variance of $\widehat{p}$, that is,

$$\mu = \mathbb{E}[\widehat{p}] = \mathbb{E}\left[\frac{X_1 + X_2 + \cdots + X_{10000}}{10000}\right] = p \,,$$
$$\sigma^2 = \mathbb{V}\mathrm{ar}[\widehat{p}] = \mathbb{V}\mathrm{ar}\left[\frac{X_1 + X_2 + \cdots + X_{10000}}{10000}\right] = \frac{p(1-p)}{10000} \,.$$

   (Here, we have used the fact that the variance of a Bernoulli random variable with parameter p is $p(1-p)$.)

   - Alternatively, we can observe that

$$N = 10000\,\widehat{p} = X_1 + X_2 + \cdots + X_{10000}$$

   is a binomial random variable with parameters $n = 10000$ and (unknown) p.

5. We would like to use $\widehat{p}$ to construct an interval estimator for p with a confidence level pf 95%. The interval estimator will (often) be of the form $\widehat{p} \pm \varepsilon$. We have to choose ε such that $\mathbb{P}(p = \widehat{p} \pm \varepsilon) \approx 95\%$.

   Since $\widehat{p}$ is approximately distributed according to $\mathsf{N}\left(p, \frac{p(1-p)}{10000}\right)$, its normalized version

$$\frac{\widehat{p} - p}{\sqrt{p(1-p)/10000}} = \frac{\widehat{p} - p}{\sqrt{p(1-p)}/100}$$

   is approximately $\mathsf{N}(0,1)$. Hence, for every $a > 0$, we have

$$\mathbb{P}\left(-a < \frac{\widehat{p} - p}{\sqrt{p(1-p)}/100} < a\right) \approx \Phi(a) - \Phi(-a) = 2\Phi(a) - 1 \,,$$

where Φ is the cdf of the standard normal distribution. Setting $2\Phi(a) - 1 = 95\%$, we get $\Phi(a) = 0.975$, so we can use a computer (or the App for the normal distribution) to find $a \approx 1.95996 \approx 1.96$. Thus,

$$\mathbb{P}\left(-1.96 < \frac{\widehat{p} - p}{\sqrt{p(1-p)/100}} < 1.96\right) \approx 95\% \ .$$

Here, there is a slight complication: if we naively proceed as in the measurement example, we obtain

$$\mathbb{P}\left(p = \widehat{p} \pm 1.96 \times \sqrt{p(1-p)}/100\right) = 95\% \ ,$$

but the interval

$$\widehat{p} \pm 1.96 \times \sqrt{p(1-p)}/100 \qquad (\clubsuit)$$

depends on the parameter p that we are trying to estimate!

To resolve this issue, we can follow at least three different approaches:

Approach 1: (rough)
In $\sqrt{p(1-p)}/100$, quietly replace p with $\widehat{p}$ and pretend nobody notices. This way, we obtain the interval estimator

$$\widehat{p} \pm 1.96 \times \sqrt{\widehat{p}(1-\widehat{p})}/100$$

Since the sample size is large, $\widehat{p}$ is expected to be close to p, so the answer we obtain using this approach should not be too off the mark. In other words, the confidence level of the above estimator should be close to 95%.

Approach 2: (conservative but rigorous)
Note that $p(1-p)$ is maximized when $p = 1/2$ (why?). Hence, the interval ($\clubsuit$) is included in the interval

$$\widehat{p} \pm 1.96 \times \sqrt{1/2(1-1/2)}/100 = \widehat{p} \pm 0.0098 \ .$$

The confidence level of this interval estimator will be larger than 95%.

Approach 3: (solving the quadratic inequality)
The event $p = \widehat{p} \pm 1.96 \times \sqrt{p(1-p)}/100$ can equivalently be written as

$$p - \widehat{p} = \pm 1.96 \times \sqrt{p(1-p)}/100$$

or equivalently,

$$(p - \widehat{p})^2 < 1.96^2 \times \frac{p(1-p)}{10000} \ .$$

The latter can in turn be written in the standard form $Ap^2 + Bp + C < 0$ for some A, B, C. (Here, A is a constant, while B and C depend on $\widehat{p}$.) Solving this inequality for p, we obtain $P_1 < p < P_2$ where P_1 and P_2 are the roots of the quadratic polynomial $Ap^2 + Bp + C$. Thus, we obtain an interval estimator (P_1, P_2) with (almost accurate) confidence level 95%.

6. Plugging in the observed value 0.5180 for $\widehat{p}$ in the interval estimator, we obtain an interval estimate for p.

Approach 1:	$p = 0.5180 \pm 0.009794$	(with roughly 95% confidence)
Approach 2:	$p = 0.5180 \pm 0.0098$	(with at least 95% confidence)
Approach 3:	$p = 0.5179931 \pm 0.009791769$	(with 95% confidence)

(In the third approach, the quadratic polynomial has the roots 0.5082013 and 0.5277849. We have written the interval $(0.5082013, 0.5277849)$ in the form $0.5179931 \pm 0.009791769$ for the sake of comparison with the other two approaches.)

7. The interpretation of the 95% confidence level is as follows: If we take many many similar random samples of size $10{,}000$ from the voters, and each time use the same interval estimator to form a confidence interval, then for about 95% of such samples, the obtained interval will contain the true value of the fraction p.

$\square$