UNIVERSITY OF LONDON

Goldsmiths
UNIVERSITY OF LONDON

Interval Estimation

Math Notes | Larry Cui

March 28, 2022

1 Variation of an Estimator

We can get the estimate for parameters based on a survey of n results, either by the method of maximum likelihood or moments. For example, we have $\hat{\lambda} = \overline{X}$ as the estimator for a Poisson parameter, λ. Now we need to consider another question: how close is this estimated λ_e to the real λ? The usual way to quantity the amount of uncertainty in an estimator is to construct a *confidence interval*. In principle, it's range of numbers that have a high probability of "containing" the unknown real parameter as an interior point.

Example 1: We have a random sample of size 4 (6.5, 9.2, 9.9, 12.4) from a normal pdf population:

$$f_Y(y;\mu) = \frac{1}{\sqrt{2\pi}(0.8)} e^{-\frac{1}{2}\left(\frac{y-\mu}{0.8}\right)^2}$$

To get the μ, first of all we know from method of maximum likelihood that the estimate is $\mu_e = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = (1/4)(38) = 9.5$.

Furthermore, we also know that according to the Central Limit Theorem (a very strong tendency), $\overline{Y}$ have a normal distribution[1] .

The variance of the population has been provided: 0.8^2, then

$$\text{Var}\left(\overline{Y}\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}\left(Y_i\right) = \frac{0.8^2}{n}$$

$$\sigma_{\overline{Y}} = \frac{0.8}{\sqrt{4}}$$

If we want a confidence of 95% for example, we do the Z-transformation and have the following interval:

$$P(-1.96 \leqslant Z \leqslant 1.96) = 0.95 = P\left(-1.96 \leqslant \frac{\overline{Y} - \mu}{0.8/\sqrt{4}} \leqslant 1.96\right)$$

[1]The population from which the sample is taken happens to be a normal distribution also is nothing but an coincidence. We will always have a normal $\overline{Y}$ irrespective of the underlying population.

One interpretation of the above equation is, 9.5, the sample mean $\overline{Y}$, in 95% probability, is within the interval around the real μ:

$$\mu - 1.96 \cdot 0.4 \leqslant \overline{Y} \leqslant \mu + 1.96 \cdot 0.4$$

On the other hand, we can also have an "inverting" probability statement about μ:

$$\overline{Y} - 1.96 \cdot 0.4 \leqslant \mu \leqslant \overline{Y} + 1.96 \cdot 0.4$$

We call the above range a 95% "confidence interval for" μ. In the long run, 95% of the time the range will contain the unknown real μ. From the construction of the range, we can see the range is not fixed, due to different $\overline{Y}$ from samples, it actually oscillates around the real μ. Maybe an illustration below will better convey the idea:

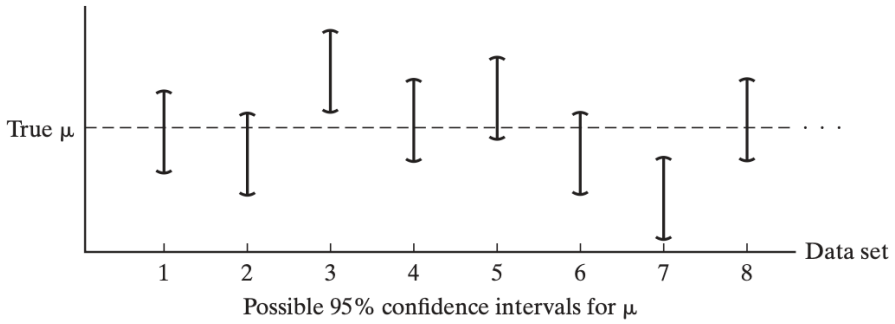

Figure 1: Inverting Probability

But no matter how the $\overline{Y}$ oscillates, the confidence interval constructed as above around any $\overline{Y}$ would be in 95% chance contain the true μ.

2 Confidence Interval for the Binomial Parameter

Sometimes we just don't know the variance of a population, so it becomes impossible to construct a confidence interval as showed above. Fortunately, many of populations are coming as binomial distribution, which we can deduce the $p = X/n$ and variance $\sigma^2 = np(1-p)$. The sample mean, $\overline{X}$, has a normal distribution pdf and can be Z-transformed:

$$\frac{X/n - p}{\sqrt{\frac{(X/n)(1-X/n)}{n}}}$$

Comment: We want use this equation to get a confidence interval for true p. But must bear in mind two issues:
(1) we have an estimate of p already, which is $p_e = X/n = \overline{X}$.
(2) by convention, we will use p_e to calculate σ. We cannot wait for the true p, which may never be known. When the sample size n is big enough, however, the error can be neglected as

minor.

Confidence Interval Let k be the number of successes in n independent trials, where n is large and $p = P(success)$ is unknown. An approximate $100(1-a)\%$ confidence interval for p is:

$$\left(\frac{k}{n} - z_{a/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}, \quad \frac{k}{n} + z_{a/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}\right)$$

When $a = 0.05$, $z_{a/2} = 1.96$.

There are many software package dedicated to generate random numbers according to certain pdf. However can we be sure the numbers generated follow the pdf? Well, we use confidence interval to evaluate the result.

A common approach is to see if the numbers fall "equally" to both sides of a pdf median (*median test*). The underlying pdf can be any kinds, the sample result, however, either in the upper part or in the lower part, is following a binomial distribution. This is why our equation for confidence interval can kick in.

Example 2: Suppose $y_i, y_2, \ldots, y_n$ denote measurements presumed to have come from a exponential pdf, $f_Y(y) = e^{-y}, y \geqslant 0$. Check if this sample pass the median test.

We need to find the median of the pdf first:

$$\int_0^m e^{-y}\,dy = -e^{-y}\Big|_0^m = 1 - e^{-m} = 0.5$$

$$m = 0.69315$$

We know from the sample test that there are 26 numbers out of the total 60 below the median. So we have $k/n = 26/60$, and $\sigma/\sqrt{n} = \sqrt{\frac{(k/n)(1-k/n)}{n}} = \sqrt{\frac{(26/60)(1-26/60)}{60}}$, the 95% confidence interval of p is:

$$\left(\frac{26}{60} - 1.96\sqrt{\frac{(26/60)(1-26/60)}{60}}, \quad \frac{26}{60} + 1.96\sqrt{\frac{(26/60)(1-26/60)}{60}}\right) = (0.308, 0.558)$$

The above interval contains 0.5, which means the dataset passes the median test. The software generates numbers of the pdf right.

3 Margin of Error

Let w denote the width of a 95% confidence interval for the true p. we have

$$w = \frac{k}{n} + 1.96\sqrt{\frac{(k/n)(1-k/n)}{n}} - \frac{k}{n} + 1.96\sqrt{\frac{(k/n)(1-k/n)}{n}}$$
$$= 3.92\sqrt{\frac{(k/n)(1-k/n)}{n}}$$

k may vary from sample to sample, but on a conservative stand point, we want w to be large enough to wrap up whatever p the total population would take. Apparently, the maximum number that $(k/n)(1-k/n)$ can have is $1/4$. Not a precise one, but a safe bet for w, would be

$$\max w = 3.92\sqrt{\frac{1}{4n}}$$

Margin of Error (denoted by d) is one-half of max w.

For a 95% confidence interval, it's $1.96/2\sqrt{n}$. Usually, d is represented in percentage, so $d = 0.05$ is a margin error of 5%.

The true p of the population would be within the scope of $(k/n - d, k/n + d)$ in 95% cases.

Example 3: A final poll of $n = 1000$ showed that 480 of the respondents were supporting candidate Max. If Max needs 50% to win the election, is the poll suggesting a possible winning?

We know d is the maximum distance from the center, i.e., $k/n = 0.48$. If we stick to 95% confidence interval, then

$$d = 1.96/(2\sqrt{n}) = 1.96/(2\sqrt{1000}) = 0.031$$

This tells us the 95% interval is $(0.449, 0.511)$:

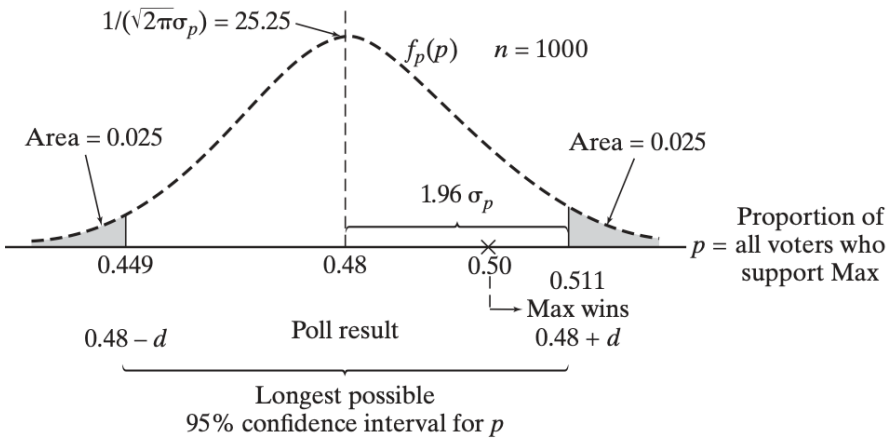

Figure 2: Sample Size of 1,000

Comment: As mentioned above, $p_e = k/n$ follows normal distribution because of CLT, the pdf of p_e:

$$f_p(p_e) = \frac{1}{\sqrt{2\pi}(\sigma)} e^{-\frac{1}{2}\left(\frac{p_e - p}{\sigma}\right)^2}$$

We may never know the true p and σ of the population. But we play safe by letting $p = 1/2$ in cdf to get the largest value for margin error d. Then, we get σ from d by re-writing the cdf for right part of the 95% interval:

$$\frac{d}{\sigma} = \frac{0.311}{\sigma} = 1.96$$

So we have a presumed $\sigma = 0.0158$, slightly larger than the true one, but in any cases that n is large, the difference is so small that can be ignored.

Another observation is, when we draw the pdf of $f_p(p_e)$, we actually are centered at $p_e = k/n = 0.48$. Again, had multiple samples been collected, the p_e might distribute around true p other than 0.48. But with only one sample of a large size n in hand, 0.48 is our best bet for the population.

Now we can flip around the question: if the true p is 0.48, what's Max's chance to win the election (get at least 50% ballets) ?

This question equals to finding the tail area to the right of 0.50 of the cdf graph in Figure 2.

$$\begin{aligned} P(\text{Max wins}) &= P[(p - 0.48)/0.0158 > (0.50 - 0.48)/0.0158] \\ &= P(Z > 1.27) \\ &= 0.1020 \end{aligned}$$

Comment: $d = 1.96/2\sqrt{n}$ tells us that margin error will become smaller if size n grows larger. In another word, the estimate p_e would be closer to the true one. We visit Max's case again, but this time we collect a poll of size 5000. Margin error is therefore $d = 1.96/2\sqrt{5000} = 0.014$, the pdf of p_e:

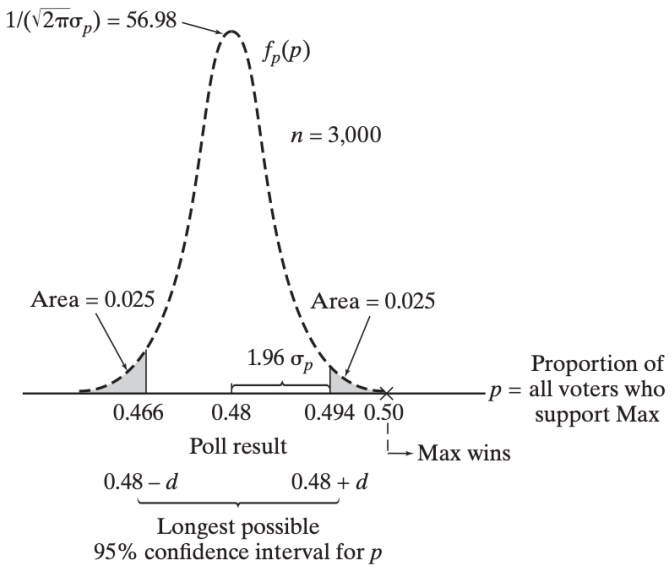

Figure 3: Sample Size of 5,000

σ of this new p_e is even smaller: $\sigma = 0.014/1.96 = 0.0071$. Only miracle happens can Max win the election now:

$$\begin{aligned} P(\text{Max wins}) &= P[(p - 0.48)/0.0071 > (0.50 - 0.48)/0.0071] \\ &= P(Z > 2.82) \\ &= 0.0024 \end{aligned}$$

4 Choosing Sample Sizes

Put it in simple words, the margin error d is the number, if divided by σ of the normal distributed p_e, will get a single side of the confidence interval. If we are looking for a 95% interval, the single side is 1.96.

We get p_e from k/n, and use p_e to approximate true p. It's no surprise to us that we will continue using p_e to calculate σ. Taking into account of the sample size n, we have:

$$\sigma = \sqrt{\frac{(k/n)(1 - k/n)}{n}}$$

For a 95% confidence interval, if we hold d fixed, we can get the size n for the exact interval:

$$\begin{aligned} \frac{d}{\sqrt{\frac{(k/n)(1 - k/n)}{n}}} &= 1.96 \\ \frac{d^2 n}{(k/n)(1 - k/n)} &= 1.96^2 \\ n &= \frac{1.96^2 (k/n)(1 - k/n)}{d^2} \end{aligned}$$

Comment: Note that we are using k/n to approximate the true p, so it might be over-

estimated or under-estimated. However, no matter what number the true p takes, the product of $(k/n)(1 - k/n)$ will not be greater than $1/4$. As a result, we can substitute $(k/n)(1 - k/n)$ with $1/4$, to get a larger, safer number of n:

$$n = \frac{1.96^2}{4d^2}$$

Comment: Most of the time, the size n derived from the above equation is good enough for practical purposes. Sometimes, however, when the survey is running on a tight budget, we want the sample size to be as small as possible without losing the accuracy. If we know the rough estimate of the true p from the population, we can furthermore decrease n. For example, if we know p is around 0.20, then:

$$n = \frac{1.96^2(0.20)(1 - 0.20)}{d^2} = \frac{1.96^2}{d^2} \cdot \frac{4}{25}$$

Comment: We claim the population is a binomial distribution and jump to the conclusion that the variance of it is:

$$\text{binomial:} \quad \sigma^2 = \frac{p(1-p)}{n}$$

But, the sample of any survey is carried out *without replacement*. This is in fact a hyper-geometric distribution. Binomial and hyper-geometric both have the same $p = k/n$, but their variance are different.

$$\text{hyper-geometric:} \quad \sigma^2 = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right)$$

The second part $\dfrac{N-n}{N-1}$ is called *finite correction factor*. In most cases, the population size N is much greater than sample size n that $\dfrac{N-n}{N-1} \to 1$, we do not need to include the finite correction factor. But when N is small compared to a sizeable n, it's worth correction for the σ^2.