Confidence Intervals

Reading

- Sections 4.1, 4.2

Practice Problems

4.6.2 (Page 206) 4.7, 4.8, 4.9, 4.10, 4.11, 4.12, 4.13, 4.14

Notes

Confidence Intervals

When we take a sample from a population, there two kinds of questions we aim at answering. The first of these is the idea of confidence intervals.

> **Confidence Interval Question**
>
> We have taken a sample from a population, and have computed its sample mean $\bar{x}$ and other information. What can we say about the population mean μ?
>
> Given the randomness of the sample, we should expect a certain degree of uncertainty in our answer. Confidence intervals make this more precise.

The idea of confidence intervals goes as follows:

- Once we have verified some conditions, we can say that $\bar{x}$ follows a normal distribution:

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- We decide on a percentage C that we will call the confidence level C (e.g. 90%).
- We locate the range in the normal distribution that covers C of the cases. We do this by computing a number m such that C of the cases are between $\mu - m$ and $\mu + m$. Note that μ is actually an unknown, but we will be able to know m.
- This means that C of the possible samples out there produce a $\bar{x}$ in the range between $\mu - m$ and $\mu + m$.
- So now make the assumption that the sample we took is in fact one of those C of them. We have a C chance of being correct in that assumption.
- With that assumption in place, the $\bar{x}$ from the sample, which we have computed, is within m from μ.

- Turning that around, we can say that μ must be within m from $\bar{x}$.

Let us summarize all this:

Confidence Interval for population mean

- Decide on a **confidence level C** (often 95%).
- Compute z^*.
  - This is the value in the standard normal distribution $N(0,1)$ such that C of the distribution is between $-z^*$ and z^*.
  - You can find the power endpoint as the z with $p = \frac{1-C}{2}$.
  - Or the upper endpoint as the z with $p = \frac{1+C}{2}$.
- Compute the **margin of error**:

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

- Then we say that we have for the population mean μ the **confidence interval** $(\bar{x} - m, \bar{x} + m)$, with confidence level C.
- What this means is this: We are claiming that the population mean μ is somewhere in the range $(\bar{x} - m, \bar{x} + m)$, and we are C confident of this fact, because we followed a process that has a C chance of being correct.

This last point is important: There are two sources of indeterminacy, if you like:

- We are claiming μ is somewhere in a range, rather than an exact value. The width of that range is controlled by m, and is a measure of the **accuracy** of our prediction.
- There is a chance we will be wrong. This is measured by C (or rather $1 - C$, and is a measure of our **confidence** in our prediction.

We will see in a moment that there is a tradeoff involved: We can increase our accuracy if we are willing to reduce our confidence, and vice versa.

In any case, there is always a $1 - C$ chance of being wrong. This is especially important if we want to compute multiple confidence intervals, for more than one variable. Suppose we take samples and compute confidence intervals for k different variables, independent of each other. If each confidence interval is taken at the C level, then the chances that all intervals are correct are C^k. The assumption of independence is rarely correct however, and other more complicated techniques have to be used in that case. For now, it is something to keep in mind when considering multiple confidence intervals.

in any case, a 95% confidence level for example means that in about one out of 20 times when we do this we'll be wrong.

Controlling the Margin of Error

A key quantity in a confidence interval is the margin of error m. We want it to be as small as possible, but it comes at a cost.

To reduce the margin of error

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

we can:

- Make z^* smaller. This means making C smaller.

  Cost: less confidence in our answer.
- Make σ smaller. We unfortunately do not have much control over σ, it is the population standard deviation. We can however try some techniques like *stratified sampling* that might result in a smaller σ at parts of the computation.

  Cost: Much more complicated sampling process and analysis phase.
- Make n larger. Because of the presence of the square root, you often need a disproportionate increase in the sample size (e.g. 100-fold increase in n for a 10-fold decrease in m).

  Cost: Considerable resource cost increase. Calling 4000 people rather than 40 is a lot more expensive, and might even be impossible depending on the population size.

As an example, suppose that $\sigma = 1$ and $n = 50$. Let us try to achieve a confidence level of $C = 90\%$. To find z^*, we will look up $p = \frac{1+0.9}{2} = 0.95$ in the table, and find $z^* = 1.645$. Therefore we find a margin of error:

$$m = 1.645 \times \frac{1}{\sqrt{40}} = 0.26$$

In other words, we will be predicting that the population mean μ is within 0.26 of whatever value our sample mean $\bar{x}$ has.

Suppose we want to achieve a margin of 0.1. Let us look at our options. Suppose first that we want to keep our confidence level fixed, and therefore z^* fixed. Then the only other thing we really have control over is n.

The sample size needed to achieve a given margin of error is:

$$n = \left(\frac{z^* \sigma}{m}\right)^2$$

In our case this means $n = \left(\frac{1.645\times 1}{0.1}\right)^2 = 270.6$, so 271 samples, almost 7 times more than what we started with.

The other alternative would be to reduce our confidence level. Let us compute what z^* should be:

$$0.1 = z^* \frac{1}{\sqrt{40}} = 0.158 \times z^*$$

Therefore $z^* = 0.633$. The corresponding p value is 0.737. Turning that into a C value would require $p = \frac{1+C}{2}$. We would get $C = 2p - 1 = 0.474$. So to achieve that margin of error we would need to drop down to a 47.5% confidence level. That is very low.