A simple random sample of size n is selected from a population...

(a) The ***expected value*** for the sample average is

$$EV_{avg} = \text{Population average},$$

i.e., (average of all possible sample averages) = population average.

(b) The ***standard error*** for the sample average is

$$SE_{avg} = \frac{SD_{pop}}{\sqrt{n}} \approx \frac{SD_{sample}}{\sqrt{n}}.$$

(c) If the sample size, n, is large enough, then the distribution of (all possible) sample averages has an approximately normal distribution, i.e.,

$$z = \frac{\text{(sample average)} - \text{(population average)}}{SE_{avg}}$$

follows the normal curve closely.

(d) So... if n is large enough,

$$P(\text{sample avg} - 2SE_{avg} < \text{pop. avg} < \text{sample avg} + 2SE_{avg}) \approx 95\%.$$

Example 1. A simple random sample of 1050 households in a certain city is surveyed.

Statistics: Sample average household income: \$3200; sample standard deviation \$2800.

Find a 95%-confidence interval for average monthly household income in the city.

(*) $SE = \frac{SD}{\sqrt{n}} = \frac{2800}{\sqrt{1050}} \approx 86.41.$

(*) 95%-confidence interval: $(\text{sample average} \pm 2SE) \approx (3200 \pm 173)$.

(*) Interpretation: There is an approximate 95% chance that the interval $(\$3027, \$3373)$ covers the average monthly household income in the city.

(*) Objection: The income data for the city doesn't follow the normal curve (how can we tell?), so we can't use it to figure probabilities!

(*) Justification: The sample averages ***do*** follow the normal curve (once the sample size is big enough). The confidence interval for the population average is constructed based on the distribution of sample averages.

Example 1. (cont.) There were 1950 children age 10 or younger in the sample households.

Statistics: The average amount of daily screen-time for these children was 3.3 hours, with a standard deviation of 2.5 hours.

Find a 95%-confidence interval for average daily amount of screen-time for children age 10 or under in the city.

(*) $SE = \dfrac{2.5}{\sqrt{1950}} \approx 0.057.$

(*) Confidence interval: (3.3 ± 0.114).

(*) Objection: The income data for the city doesn't follow the normal curve (how can we tell?), so we can't use it to figure probabilities!

(*) Justification: The sample averages ***do*** follow the normal curve (once the sample size is big enough). The confidence interval for the population average is constructed based on the distribution of sample averages.

Example 1. (cont.) There were 1950 children age 10 or younger in the sample households.

Statistics: The average amount of daily screen-time for these children was 3.3 hours, with a standard deviation of 2.5 hours.

Find a 95%-confidence interval for average daily amount of screen-time for children age 10 or under in the city.

(*) $SE = \dfrac{2.5}{\sqrt{1950}} \approx 0.057$.

(*) Confidence interval: (3.3 ± 0.114).

Something is wrong

(*) The sample of children is ***not*** a simple random sample. It is a ***cluster sample*** of children.

$\Rightarrow$ The method we use to construct a 95%-confidence intervals for population average based on a simple random sample is incorrect for cluster samples.

Example 1. (cont.) The average daily screen time for heads-of-household in the sample households is 4.8 with $SD = 3.8$.

(*) $SE = 3.8/\sqrt{1050} \approx 0.117...$

Is 4.8 ± 0.234 hours a 95%-confidence interval for the average daily amount of screen time for heads-of-household in the city?

Yes — A simple random sample of households is also a simple random sample of heads-of-household.

The Gauss model for measurement error.

When repeated, independent measurements are made of the same quantity, the observed values may be composed of two or of three components:

$$\text{Observed value} = \begin{cases} \text{true value} \quad + \quad \text{chance error} \\ \textbf{\textit{or...}} \\ (\text{true value} + \text{bias}) \quad + \quad \text{chance error} \end{cases}$$

(1) The *true value* is constant. Sometimes known, sometimes not.

(2) The chance errors in different measurements are (assumed to be) *independent* of each other. I.e., the chance errors are like random draws with replacement from a box of tickets: the ***error box***.

- The *error box* has average 0.
- The SD of the error box is generally unknown, but is estimated by the SD of the measurements.
- It is often assumed that the error box follows the normal curve.

(3) The bias is a nonzero constant.

We can use this model to estimate the value of a quantity being measured (assuming no bias), based on the following principle:

The average of a sequence of measurements almost certainly yields a more accurate estimate for the value being measured than any single measurement.

(*) n measurements are made, v_j is the j^{th} observed value, ε_j is the chance error in the j^{th} measurement and $\mathcal{V}$ is the true value of the quantity being measured. So, assuming no bias:

$$v_j = \mathcal{V} + \varepsilon_j \qquad \text{for } 1 \le j \le n.$$

(*) The average of the n measurements:

$$\overline{v} = \frac{v_1 + v_2 + \cdots + v_n}{n} = \frac{(\mathcal{V} + \varepsilon_1) + (\mathcal{V} + \varepsilon_2) + \cdots + (\mathcal{V} + \varepsilon_n)}{n} = \mathcal{V} + \overline{\varepsilon}$$

(*) $\overline{\varepsilon}$ is the (unknown) average of the n errors, and

$$\Rightarrow \quad \mathcal{V} = \overline{v} \pm |\overline{\varepsilon}|.$$

Next: estimate $|\overline{\varepsilon}|$

(*) The standard error for the average of the errors is

$$SE_\varepsilon = \frac{SD(\text{error box})}{\sqrt{n}} \approx \frac{SD(\text{observed values})}{\sqrt{n}}.$$

(*) The average of the error box is 0 so if n is large enough,

$$P(|\overline{\varepsilon}| < 2SE_\varepsilon) = P(|\overline{\varepsilon} - 0| < 2SE_\varepsilon) \approx 95\%.$$

I.e., there is a 95% chance that $-2SE_\varepsilon < \overline{\varepsilon} < 2SE_\varepsilon$.

Conclusion: If there is no bias, then $(\overline{v} \pm 2SE_\varepsilon)$ is a 95%-confidence interval for $\mathcal{V}$, the true value of the quantity being measured.

(*) This is useful (and accurate) because SE_ε is generally *small*.

Example. The concentration of a saline solution is measured repeatedly, $n = 200$ independent measurements are made. We want to estimate the concentration of salt in this solution.

Statistics: Average concentration: $\bar{c} = 9.2$ gram/liter; Standard deviation: $SD = 0.27$ gram/liter.

(*) Standard error: $SE = \dfrac{0.27}{\sqrt{200}} \approx 0.019.$

Confidence interval: (9.2 ± 0.038).

Interpretation: There is a 95% chance that the interval $(9.162, 9.238)$ covers the true concentration of salt in the solution (grams/liter).

Observation: If bias is present, then the interval

$$\text{average(measurements)} \pm 2SE\text{(measurements)}$$

is a 95% confidence interval for (true value)+(bias).

(*) If the true value is *known*, then we can use these ideas to test for bias (in the measuring procedure).

Example. A new scale is being evaluated at the CDFA, Division of Measurement Standards. A one kilogram checkweight (true weight = 1.00031 kg) is weighed $n = 50$ times.

Statistics: The average of the measurements is 1.008 kg and the standard deviation of the measurements is 0.023 kg.

Using these numbers we construct the following interval:

$$\text{Avg} \pm 2SE = 1.008 \pm 2 \cdot \frac{0.023}{\sqrt{50}} = 1.008 \pm 0.0065.$$

What does this tell us?

Analysis: If there is no bias — if the scales are well-calibrated — then the interval $(1.0015, 1.0145)$ is a 95% confidence interval for the true value of the checkweight.

⇒ This interval *does not cover* the *known* true value, 1.00031. *Why?*

(*) One possibility is that chance error causes this, but this is not very likely. Chance error will cause the confidence interval to miss its target in only 5% of the procedures.

(*) The other possibility is that there is bias — the scales are not well-calibrated.

(*) If the probability that the 'miss' was caused by chance error is *low enough*, then we conclude that the 'miss' was caused by bias.

This type of analysis is streamlined a bit in a procedure known as a *test of significance.*