VI.23 The Accuracy of Averages (Confidence Interval for a Population Mean) (also: C.I. for Difference of Two Population Means for Different Populations)

- A. How can we describe the accuracy of \( \bar{X} \) for estimation of \( \mu \)?

A. In the same fashion that we describe the accuracy of \( \hat{p} \) for estimation of \( p \): a confidence interval.

Recall: the interval

\[
\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

has approximate confidence level \( 1 - \alpha \) (or \( 100(1-\alpha)\% \)).

The above interval is of the form

\[
\hat{p} \pm z_{1-\frac{\alpha}{2}} \text{SE}(\hat{p})
\]

To get a confidence interval for \( \mu \) based on \( \bar{X} \), we will follow this same approach, with some necessary modifications.

Actually, this is an estimate of \( SE(\hat{p}) \).

The actual \( SE(\hat{p}) \) is

\[
\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

which is unknown if \( p \) is unknown.
The Setting

- a population
- a chance variable $X$: a numerical measurement on a randomly selected item from the population (for example, height of persons, or diameter of trees)
- a "parent" probability distribution for the values of $X$ (for example, Poisson, or exponential, or Normal)
- a mean $\mu$ and a variance $\sigma^2$ for the parent probability distribution
- a random sample of values $X_1, X_2, \ldots, X_n$
- sample size: $n$
- sample mean: $\bar{X} = \frac{1}{n} (X_1 + X_2 + \ldots + X_n)$
- sample SD: $\sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$
- target parameter to be estimated: $\mu$
- additional parameter needing to be estimated: $\sigma$
- **Goal**: a $100(1-\alpha)%$ confidence interval for $\mu$ based on $\bar{X}$

**NOTE.** The population might be a box

```
[1 2 -3 100.63 3 20.7]
```
or a population of individuals $\{x, x, \ldots, x\}$, for example.
The Solution: a 100(1-\(\alpha\))% C.I. for \(\mu\) based on \(\bar{X}\)

Case 1. \(\sigma^2\) is known (typically, it is not known)

We use

\[
\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}
\]

because the SE of \(\bar{X}\) is \(\frac{\sigma}{\sqrt{n}}\). Based on the CLT, this interval has approximate confidence level 1-\(\alpha\).

Case 2. \(\sigma^2\) is unknown (this is the typical situation)

We modify the above in 2 ways:

(i) Replace \(\sigma\) by the sample SD

\[
\hat{\sigma} = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{X})^2}
\]

(ii) Widen the interval by replacing \(Z_{1-\frac{\alpha}{2}}\) by a slightly larger number.

Why? This is the price due to adding the uncertainty by replacing \(\sigma\) by a random estimator.

What larger number?

The analogue of \(Z_{1-\frac{\alpha}{2}}\) taken from the \(t_{n-1}\) distribution instead of from the \(N(0,1)\) distribution.

If \(n\) is moderate in size, say \(n \geq 30\), we can skip this step.
The "t" distributions

[see text, pp. 488-495; see table, p. A-105]

- For each positive integer \( m = 1, 2, \ldots \), we define
  the \( t\)-distribution with \( m \) degrees of freedom.

- The \( t_m \) distribution is similar to the \( N(0, 1) \) distribution except that the density curve is lower in the middle and higher in the tails:

\[ \text{N}(0, 1) \text{ curve} \quad \text{---} \quad t\text{-distribution curve} \]

\[ \text{Both are symmetric about 0} \]

- As the "degrees of freedom" \( m \) increases, the \( t_m \) curve gets closer and closer to the \( N(0, 1) \) curve.

- What chance variable has a \( t_m \) distribution?
  
  A. Take \( N(\mu, \sigma^2) \) as the parent distribution.

  Take a sample \( X_1, X_2, \ldots, X_n \) from this distribution.

  Then \( X \) has the distribution \( N(\mu, \sigma^2) \) exactly.

  Then

  \[ \frac{X - \mu}{\sigma / \sqrt{n}} \]

  has the distribution \( N(0, 1) \).

  And

  \[ \frac{X - \mu}{s / \sqrt{n}} \]

  has the distribution \( t_{n-1} \).

  \( \text{Replacing} \ \sigma \ \text{by} \ s \)

- The fact was first worked out by W. S. Gosset in 1908.
  He published it under the pseudonym "Student". So we also call these the "Student t" distributions 😊
We now complete the discussion of "Case 2: \( \sigma^2 \) unknown."

- Our 100(1 - \( \alpha \))% Confidence Interval for \( \mu \) is:

\[
\bar{X} \pm Z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \quad \text{if } n \geq 30
\]

or

\[
\bar{X} \pm t_{n-1, 1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \quad \text{if } n < 30
\]

- As \( n \) increases, the constant \( t_{n-1, 1-\frac{\alpha}{2}} \) approximates \( Z_{1-\frac{\alpha}{2}} \) more and more closely, because the \( t_{n-1} \) curve approximates the \( \mathcal{N}(0, 1) \) curve more and more closely.

**Example.** Take \( \alpha = 0.05 \). Then \( 1 - \frac{\alpha}{2} = 0.975 \).

<table>
<thead>
<tr>
<th>( n )</th>
<th>( t_{n-1, 0.975} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>2.78</td>
</tr>
<tr>
<td>10</td>
<td>2.26</td>
</tr>
<tr>
<td>30</td>
<td>2.05</td>
</tr>
<tr>
<td>60</td>
<td>2.00</td>
</tr>
<tr>
<td>( \infty )</td>
<td>1.96 ( \approx 2.00 )</td>
</tr>
</tbody>
</table>

**NOTE.** For \( n \geq 30 \), the value of \( t_{n-1, 0.975} \) is very close to 2.00.
Setting

- A parent distribution with mean $\mu$ and variance $\sigma^2$, with both parameters unknown.
- Target parameter: $\mu$.
- A random sample $X_1, \ldots, X_n$ from the parent distribution.

The Point Estimator

The point estimator is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$ 

Key property. By the Central Limit Theorem, the distribution of $\bar{X}$ is approximately $N\left(\mu, \frac{\sigma^2}{n}\right)$. This holds exactly if the parent distribution is Normal.

Approximate $100(1 - \alpha)\%$ Confidence Interval for $\mu$

(a) Default choice:

$$\bar{X} \pm z_{1-\alpha} \frac{s}{\sqrt{n}},$$

with $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$, the sample standard deviation.

(b) Modification for small $n$. If $n \leq 30$, then we should allow for more uncertainty due to the variability in $s$ as estimator of $\sigma$ and due to the lessened accuracy of the CLT. For this purpose, we may replace $z_{1-\alpha}$ by $t_{n-1,1-\alpha}$ based on the $t$ distribution with $n - 1$ degrees of freedom.

In the case that the parent distribution is Normal, this modification gives an exact rather than approximate $100(1 - \alpha)\%$ confidence interval and is therefore the preferred approach.

— RJS, 1/18/2011
Example (see text, pp. 417 - 419)

- population: people in a town
- chance variable \( X \): total years of schooling
- target parameter: the mean years of schooling for all persons age 25 and over: \( \mu \)

- sample of size \( n = 400 \) taken: \( X_1, \ldots, X_{400} \)
- data: the years of schooling for the sample individuals

\[
\bar{X} = \frac{\sum_{i=1}^{400} X_i}{400} = 11.6 \text{ years}
\]

\[
\text{SD for sample} = \sqrt{\frac{1}{400} \sum_{i=1}^{400} (X_i - 11.6)^2} = 4.1 \text{ years}
\]

\[
\text{SE}(\bar{X}) = \frac{1}{\sqrt{400}} \times 4.1 \text{ years} = 0.205 \text{ years}
\]

- approximate 95\% C.I. for \( \mu \):

\[
11.6 \pm 1.96 \times 0.205
\]

i.e., \( 11.6 \pm 0.4 \)

i.e., \( [11.2, 12.0] \) (years)

Interpretation: This random interval has chance 0.95 of falling onto the value \( \mu \).

- see text for further discussion of this example

THIS SPACE IS FOR YOU :) GET A 99% C.I. FOR \( \mu \).
Practical Guide for Estimation of Difference of Two Population Means Using $\bar{X}_1 - \bar{X}_2$

Setting

- Two parent distributions with respective means $\mu_1$, $\mu_2$ and variances $\sigma_1^2$, $\sigma_2^2$ with all four parameters unknown.
- Target parameter: $\mu_1 - \mu_2$.
- Two random samples:
  - $X_{11}, \ldots, X_{1n_1}$ from the 1st parent distribution,
  - $X_{21}, \ldots, X_{2n_2}$ from the 2nd parent distribution.

The Point Estimator

The point estimator is

$$\bar{X}_1 - \bar{X}_2 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2} \sum_{i=1}^{n_2} X_{2i}.$$

**Key property.** By the Central Limit Theorem, the distribution of $\bar{X}_1 - \bar{X}_2$ is approximately $N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$. This holds exactly if the parent distribution is Normal.

Approximate $100(1 - \alpha)\%$ Confidence Interval for $\mu_1 - \mu_2$

(a) Default choice:

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{1-\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

with $s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$, the respective sample variances, for $i = 1, 2$.

**NOTE.** If it is known that $\sigma_1^2 = \sigma_2^2$, then instead use

$$\left(\bar{X}_1 - \bar{X}_2\right) \pm z_{1-\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

with $s_p^2 = \frac{1}{n_1+n_2-2}[(n_1-1)s_1^2 + (n_2-1)s_2^2]$, the "pooled sample variance".

(b) Modification for small $n_1$ or $n_2$. If either $n_1 \leq 30$ or $n_2 \leq 30$, then we should allow for more uncertainty due to the variability in sample standard deviations and the lessened accuracy of the CLT. For this purpose, we may replace $z_{1-\alpha/2}$ by $t_{\nu,1-\alpha/2}$ based on the $t$ distribution with $\nu$ degrees of freedom, where $\nu$ is a sample-based value given by "Satterthwaite's approximation":

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1} \left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1} \left(\frac{s_2^2}{n_2}\right)^2}.$$

**NOTE.** If it is known that $\sigma_1^2 = \sigma_2^2$, then instead use $t_{\nu,1-\alpha/2}$ with $\nu = n_1 + n_2 - 2$.

-- RJS, 3/31/2011
Example (Extension of previous example, p. 7)

- 2 populations: Town A and Town B

- Chance variable: $X$: total years of schooling

- Target parameter: $\mu_A - \mu_B$,

where $\mu_A =$ mean years of schooling for all persons age 25 and older in Town A

$\mu_B =$ similar for Town B

- Two random samples, of sizes $n_A = 400$ and $n_B = 350$

  for Towns A and B, respectively

- Data: the years of schooling for the sampled individuals

  $\rightarrow$ reduced to: $\bar{X}_A$, $A_A = [11.6\ yrs, 4.1\ yrs]$ 

  $\bar{X}_B$, $A_B = [10.7\ yrs, 3.8\ yrs]$ 

- NOT ASSUMED THAT $\sigma_A^2 = \sigma_B^2$

- Approximate 95% C.I. for $\mu_A - \mu_B$:

$$\left( \bar{X}_A - \bar{X}_B \right) \pm 2.975 \sqrt{\frac{(4.1)^2}{400} + \frac{(3.8)^2}{350}}$$

1.e., $0.90 \pm 1.96 \sqrt{0.042 + 0.041}$

1.e., $0.90 \pm 0.56$

1.e., $[0.34, 1.46]$ (years)

Interpretation: This random interval has chance 0.95 of falling onto the value $\mu_A - \mu_B$.

This space is for you 😊. Get a 99% C.I. for $\mu_A - \mu_B$. 