Introduction to Confidence Intervals

1. General idea, the CI based on Z
2. The t-Student distribution
3. t-Student's CI for the population mean
4. Two (independent) samples: t-Student's CI for the mean difference

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
26.04.2023

CI when population standard deviation σ is known

- If either $X_1, \ldots, X_n$ are normal distributed, or n is so large that the central limit theorem starts to work, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

- This means that we can produce a 95% confidence interval as follows:

$$95\%\ \text{CI} = \left(\bar{X} - 1.96 \times \frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{\sigma}{\sqrt{n}}\right),$$

  where 1.96 is the two-sided **5% point** of the standard normal distribution.

- Small σ or large n gives more narrow intervals and means that $\overline{X}$ is more likely to be similar to the true mean.

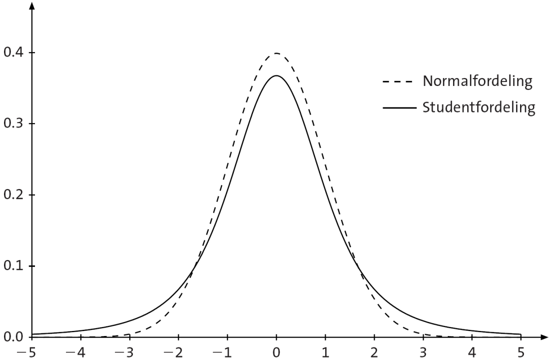
Unknown σ and the Student t-distribution

- The previous situation with known σ is rare in practice,
- We will use the previous strategy, but replace σ with the empirical standard deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

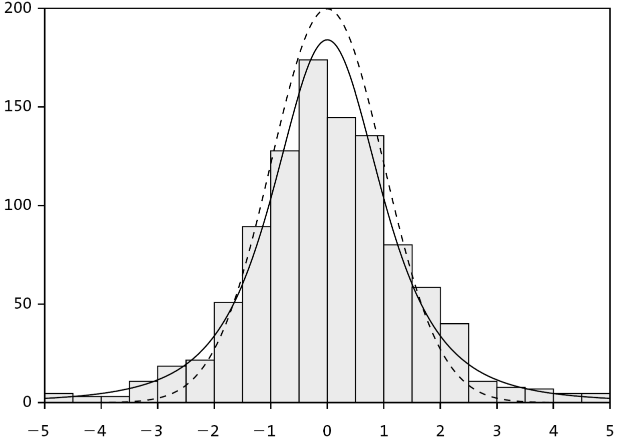
- The added uncertainty means that $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ is not normal distributed anymore, but distributed according to the so-called *Student t-distribution with n-1 degrees of freedom*, written

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1).$$



Figur 8.3 Standardnormalfordelingen er tegnet inn sammen med Studentfordelingen med 3 frihetsgrader. En ser at den siste fordelingen er mer spredt ut enn den første

- The higher degree of freedom the closer the stundent t is to the standard normal distribution N(0,1)



Figur 8.4 Det er gjort 1000 utvalg, hvert på fire tall, fra et sett med data over mannlig kroppshøyde. Størrelsen t er beregnet i hvert av utvalgene, og histogrammet viser fordelingen av t-verdiene. Som sammenlikning vises den standardiserte normalfordelingen (prikket kurve) og Studentfordelingen med 3 frihetsgrader (heltrukket kurve)

95% confidence interval when σ is unknown

- Can use previous strategy, but we cannot use the percentage point 1.96 anymore, since $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is not normally distributed
- We must use the percentage point for the Student t-distribution with n-1 degrees of freedom instead, t'
- We still need that either $X_1, \ldots, X_n$ are normally distributed, or n is so large that the central limit theorem ensures that $\bar{X}$ is normally distributed
- A 95%-confidence interval is then given by:

$$\mathrm{CI} = \left(\bar{X} - t' \times \frac{s}{\sqrt{n}}, \bar{X} + t' \times \frac{s}{\sqrt{n}}\right)$$

where s denotes the empirical standard deviation, and $\bar{X}$ denotes the sample mean.

Example 6.3 in Kirkwood & Sterne

The **numbers of hours of relief** obtained by six arthritic patients after receiving a new drug are recorded.

Patient no.	1	2	3	4	5	6
Number of hours	2.2	2.4	4.9	2.5	3.7	4.3

The sample mean is $\bar{X} = 3.3$ hours, the empirical standard deviation is $s = 1.13$ hours and the estimated standard error of the sample mean equals $s/\sqrt{n} = 0.46$ hours. The number of degrees of freedom is $(n - 1) = 5$

The 95% **confidence interval** (in hours) for the average number of hours of relief for arthritic patients in general is

$$(3.3 - 2.57 \times 0.46, 3.3 + 2.57 \times 0.46) = (2.1, 4.5),$$

where 2.57 is the two-sided 5% point of the t distribution with 5 degrees of freedom

We can simply use the normal distribution when n is large

- When n is large (150 or larger), then t' is almost the same as 1.96 for practical purposes,
- Reflects that the Student t-distribution is almost identical to the standard normal distribution for large degrees of freedom,
- The 95% **confidence interval** for the population mean is then given by

$$95\%\ \mathrm{CI} = \left(\bar{X} - 1.96 \times \frac{s}{\sqrt{n}}, \bar{X} + 1.96 \times \frac{s}{\sqrt{n}}\right).$$

Example 6.1 in Kirkwood & Sterne

We want to estimate the amount of insecticide that would be required to spray all the 10000 houses in a rural area as part of a malaria control programme. A random sample of 100 houses is chosen and the sprayable surface of each of these is measured. The **mean** sprayable surface area for these 100 houses is $\bar{X} = 24.2$ m^2, and the estimated **standard deviation** is $s = 5.9$ m^2. The estimated **standard error** of the sample mean is $s/\sqrt{n} = 5.9/\sqrt{100} = 0.6$ m^2.

The 95% **confidence interval** is:

$$(24.2 - 1.96 \times 0.6, 24.2 + 1.96 \times 0.6) = (23.0, 25.4)$$

The upper 95% **confidence limit** is used in budgeting for the amount of insecticide required per house. One litre of insecticide is sufficient to spray 50 m^2 and so the amount (in litres) budgeted for is:

$$10000 \times \frac{25.4}{50} = 5080$$

Small sample sizes

- The central limit theorem says that $\bar{X}$ is normally distributed, even if the individual observations are not
- For smaller samples, we need that the individual samples are normally distributed
- This can be easily checked with a normality plot in **R**
- When the distribution in the population is markedly *non-normal*, it may be desirable to
  - use a **transformation** on the scale on which the variable X is measured, or
  - calculate a **non-parametric** confidence interval, or
  - use **bootstrap** methods

  More on this in day 1 of week 2!

Confidence interval vs. reference range

- If the population distribution is approximately normal, the 95% **reference range** is given by

$$\text{95\% reference range} = (\mu - 1.96 \times \sigma, \mu + 1.96 \times \sigma),$$

  where μ is the population mean and σ is the population standard deviation
- There is a clear distinction between the CI and the reference range:
  - the **reference range** describes the variability between individual observations in the population
  - the **confidence interval** is a range of plausible values for the population mean, given the sample mean and its standard error

Since the sample size $n > 1$, *the confidence interval will always be narrower than the reference range*.

What if we have more than one Sample?

Two Independent Samples

2 groups: 1 measure for each individual, each which corresponds to a group (for example sick/healthy people)

Group 1		Group 2	
Ind.	Measure	Ind.	Measure
1	X_{11}	1	X_{12}
2	X_{21}	2	X_{22}
...		...	
		14	$X_{14\,2}$
15	$X_{15\,1}$		

The mean difference of two independent samples

- We want to compare the mean outcomes in two separate exposure (or treatment) groups: *group 0* and *group 1*
- In clinical trials, these correspond to the *treatment* and *control* groups,
- We will then build a **two-sample confidence interval**,
- Notation:
  - n_i is number of individuals in group i,
  - $X_{1,i}, \ldots, X_{n_i,i}$ observations in group n_i,
  - $\overline{X}_i$ average in group i,
  - μ_i mean in group i,
  - σ_i standard deviation in group i,
  - s_i empirical standard deviation in group i.

Assumptions and the Student t-distribution

- Independent individuals,
- Normal distributed averages, i.e. either
  - Large enough samples such that averages become normal distributed, or
  - Normal distributed observations,
- Equality of the two population standard deviations, σ_1 and σ_0

This means that

$$t = \frac{\bar{X}_1 - \bar{X}_0}{s\sqrt{(1/n_1 + 1/n_0)}} \tag{1}$$

is t-distributed with $n_1 + n_0 - 2$ degrees of freedom, where

$$s = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{(n_1 + n_0 - 2)}\right]} \tag{2}$$

Confidence interval for the mean difference $\mu_1 - \mu_0$

- The **confidence interval** gives a range of likely values for the difference in population means, $\mu_1 - \mu_0$, based on the difference in sample means, $\bar{X}_1 - \bar{X}_0$:

$$\mathrm{CI} = (\bar{X}_1 - \bar{X}_0) \pm t' \times s\sqrt{1/n_1 + 1/n_0},$$

where the common estimate, s, of the population **standard deviation** is given by (also at the previous slide):

$$s = \sqrt{\left[\frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{(n_1 + n_0 - 2)}\right]},$$

and t' is the appropriate **percentage point** of the t distribution with $(n_1 + n_0 - 2)$ degrees of freedom

Example: 7.2 in Kirkwood & Sterne

We consider the **birth weights** (in kg) of children born to 14 heavy smokers (**group 1**) and to 15 non-smokers (**group 0**), sampled from live births at a large teaching hospital

Heavy smokers (group 1)	**Non-smokers (group 0)**
3.18	3.99
2.74	3.89
2.90	3.60
3.27	3.73
3.65	3.31
3.42	3.70
3.23	4.08
2.86	3.61
3.60	3.83
3.65	3.41
3.69	4.13
3.53	3.36
2.38	3.54
2.34	3.51
	2.71
$\bar{X}_1 = 3.1743$	$\bar{X}_0 = 3.6267$
$s_1 = 0.4631$	$s_0 = 0.3584$
$n_1 = 14$	$n_0 = 15$

The **difference between the means** is given by

$$\bar{X}_1 - \bar{X}_0 = 3.1743 - 3.6267 = -0.4524,$$

and the **standard deviation** is given by

$$s = \sqrt{\frac{13 \times 0.4631^2 + 14 \times 0.3584^2}{14 + 15 - 2}} = 0.4121$$

with $(14 + 15 - 2) = 27$ degrees of freedom. The **standard error** of the difference is given by

$$\widehat{\text{s.e.}} = 0.4121 \times \sqrt{(1/14 + 1/15)} = 0.1531$$

The 95% **confidence interval** for the difference between the mean birth weight is given by

$$(-0.4524 - 2.05 \times 0.1531, -0.4524 + 2.05 \times 0.1531)$$
$$= (-0.77, -0.14),$$

where 2.05 is the 5% point of the t distribution with 27 degrees of freedom