Applied Statistics for Data Scientists with R

Class 16: Sampling Distribution of Mean, One and Two-samples t-test

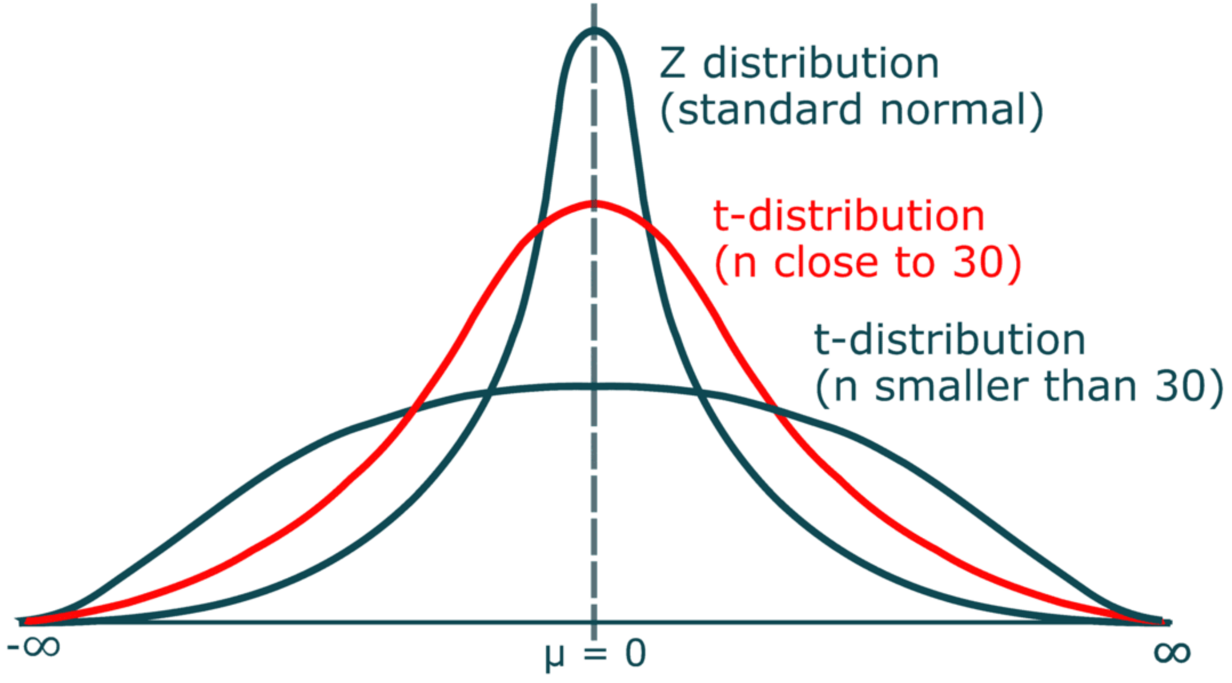
Central Limit Theorem

- Regardless of the population’s distribution, the sampling distribution of the sample mean tends to be normal as the sample size increases ($n \geq 30$ is often sufficient).

Student's t distribution

- Describes the distribution of mean of samples when the standard deviation is unknown.
- Used to compare means.
- Has degrees of freedom, $df = n - 1$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}}$$



Fun fact: William Sealy Gosset's pseudonym was Student

One sample *t*-test

- Used to:
  - Test the statistical difference between the mean value of a sample and a specific known or hypothesized population mean.
  - So, a mean value of a sample is compared to a fixed hypothesized value.

One sample t-test: Hypothesis

H_0: The sample mean is equal to X (may be any fixed value).
H_a: The sample mean is not equal to X (may be any fixed value).

Mathematically,

$$H_0 : \mu = \mu_0$$
$$H_a : \mu \neq \mu_0$$

One sample *t*-test: Assumptions

1. Sample data should be independent.
2. Sample data is randomly selected.
3. Data should come from a population distribution that follows normal distribution. (Using Shapiro-Wilk Test)

One sample t-test: Procedure

- Calculate mean $\bar{X}$
- Calculate sample standard deviation s
- Calculate test statistic

$$t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}} \longrightarrow SE = \frac{s}{\sqrt{n}}$$

- Calculate degrees of freedom, $df = n - 1$
- Calculate confidence interval

$$\bar{X} \pm t_{\alpha/2,df} \times SE$$

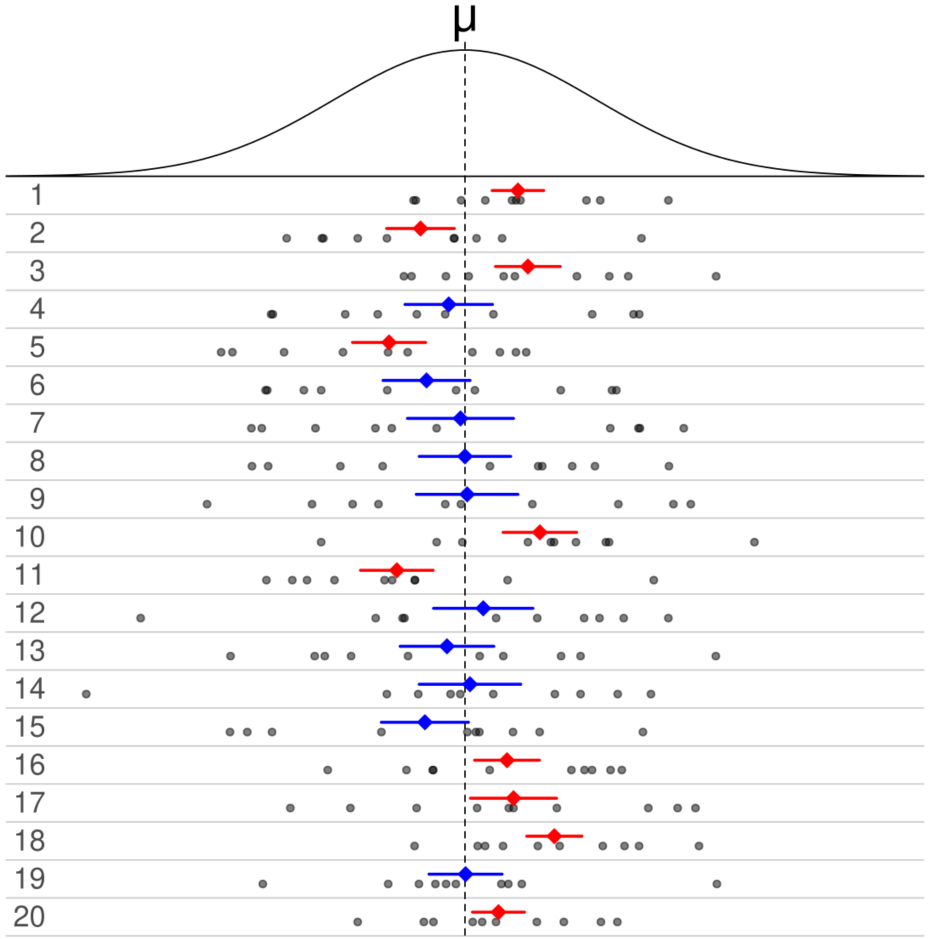
Critical value from the t-distribution for given confidence level and degrees of freedom

Confidence Interval

- Provides a range of plausible values for an unknown population parameter.
- General formula,

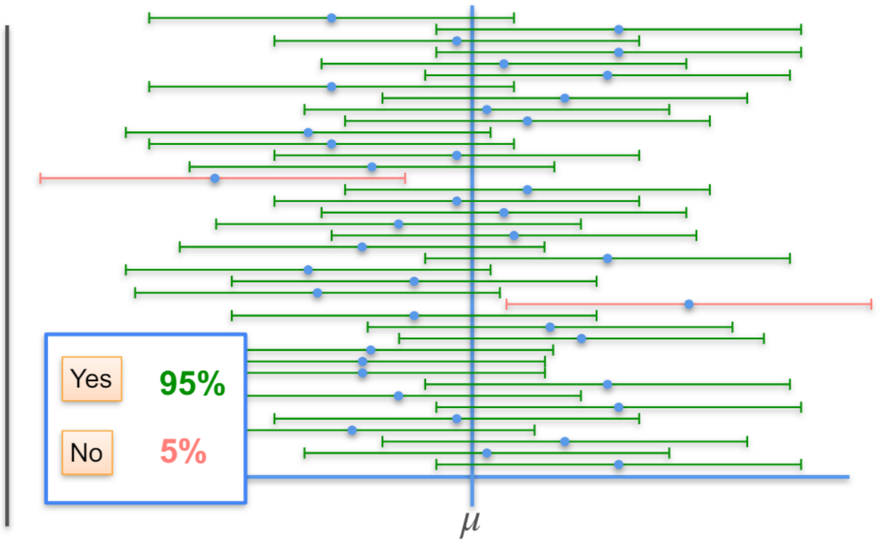
$$CI = Statistic \ \pm Mergin\ of\ Error$$
$$CI = Statistic \ \pm Critical\ value \times SE$$

$$\bar{X} \pm t_{\alpha/2,df} \times \frac{s}{\sqrt{n}} \qquad \hat{p} \pm Z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$



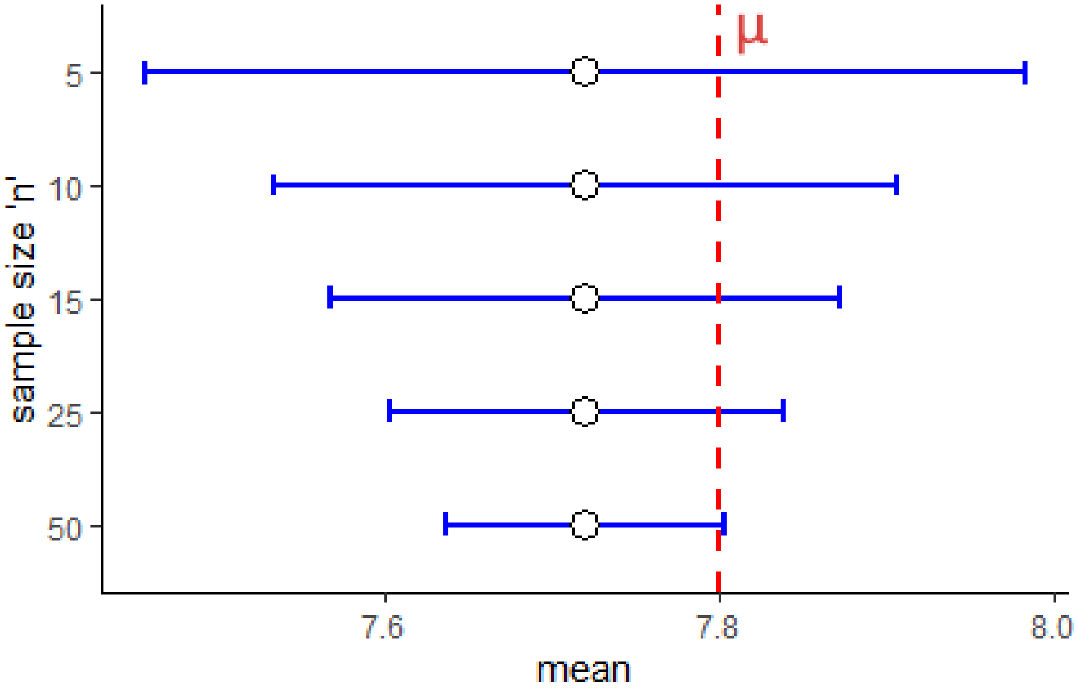
Confidence Interval (cont.)

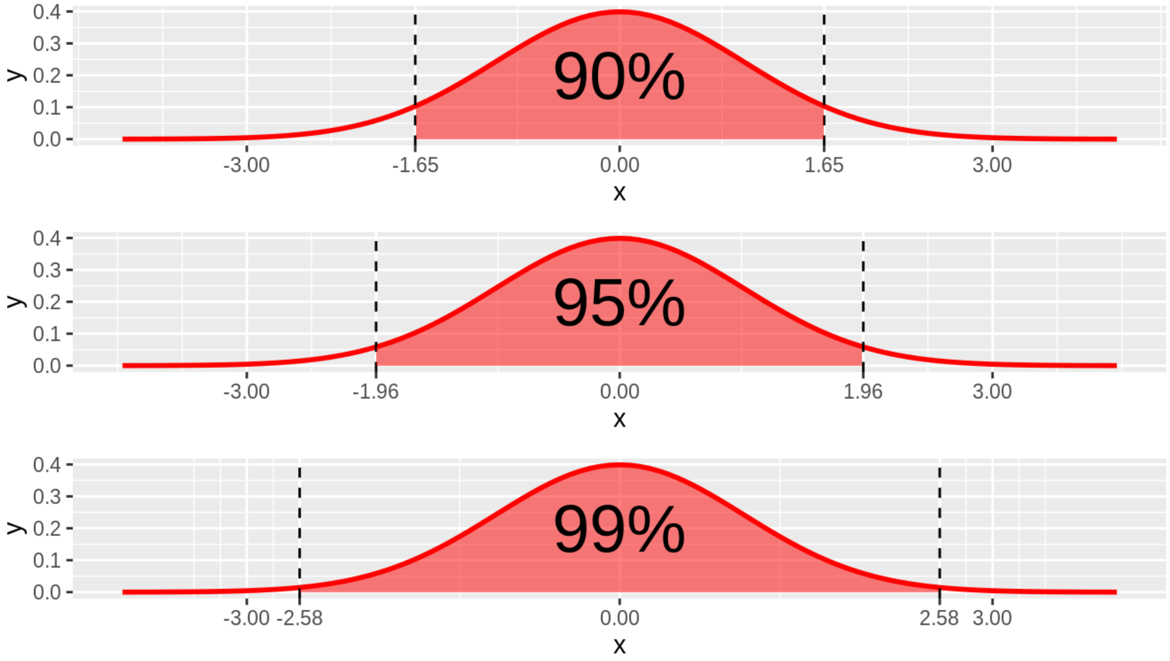
- A 95% confidence interval means that if we repeatedly take samples and compute CIs, 95% of them would contain the true parameter.
- A CI **does not** mean there is a 95% probability that the true mean is inside the interval.
- Increasing sample size decreases the width of the CI.
- Higher confidence levels (99%) make the interval wider.



Confidence Interval (cont.)

- Increasing sample size decreases the width of the CI.
- Higher confidence levels (99%) make the interval wider.





Independent Samples *t*-test

- Also known as two-sample t-test
- Used to:
  - Test the statistical difference between the means of two groups.
  - So, the dependent variable is a numeric variable, and the independent variable is categorical variable.
  - The dependent variable (values) depends on the independent variable (two groups).

Independent Samples t-test: Hypothesis

H_0: The two population means are equal
H_a: The two population means are not equal.

Mathematically,

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

H_0: The difference between the two population means is equal to 0
H_a: The difference between the two population means is not equal to 0.

Mathematically,

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_a: \mu_1 - \mu_2 \neq 0$$

Independent Samples *t*-test: Assumptions

1. Sample units are drawn randomly from the population.
2. The observations in the two samples are independent of each other. That is repeated measurements on the same individual is not taken.
3. The population from which the sample is drawn is assumed to be normally distributed.
4. The two samples come from population distributions that may differ in their mean value, but not in the standard deviation (homogeneity of variance).

Independent Samples *t*-test: Steps

- Collect data from two groups
- Calculate means → $\bar{X}_1, \bar{X}_2$
- Calculate variances of each group and then pooled variance → $s_p^2 = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$
- Calculate test statistic → $t = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ → $SE = \sqrt{s_p^2 \left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}$
- Calculate degrees of freedom → $df = n_1 + n_2 - 2$
- Calculate confidence interval → $(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2,df} \times SE$

Critical value from the t-distribution for given confidence level and degrees of freedom

Unequal variance: Welch's *t*-test

- Also known as Welch's unequal variances *t*-test.
- Uses ***Welch-Satterthwaite equation*** to compute modified degrees of freedom.

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

Test for Homogeneity of Variance

- Recommended: Levene's Test
- Other tests:
  - Brown-Forsythe Test (more robust than Levene's test)
  - Bartlett's Test (sensitive to normality assumption)

Hypothesis:

H_0: The variances of all groups are equal.
H_a: At least one group has a different variance.

Paired Samples t-test

- Used to compare the means of two related (paired) groups.
- Examples: Before & After measurements

Subject	Before	After	Difference (D = Before - After)
1	140	135	5
2	150	145	5
3	160	158	2
4	130	128	2
5	135	132	3
6	145	140	5
7	155	150	5
8	138	136	2
9	148	142	6
10	152	148	4
		Average	3.9

- Hypothesis: $H_0: \mu_D = 0$
  $H_a: \mu_D \neq 0$
- Test statistic, $t = \dfrac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$
- Degrees of freedom, $df = n - 1$
- Confidence interval, $\bar{D} \pm t_{\alpha/2,df} \times \dfrac{s_D}{\sqrt{n}}$