Inference for the Mean of a Population

Inference when σ is known

- Tests and confidence intervals for the mean μ of a normal population are based on the sample mean

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

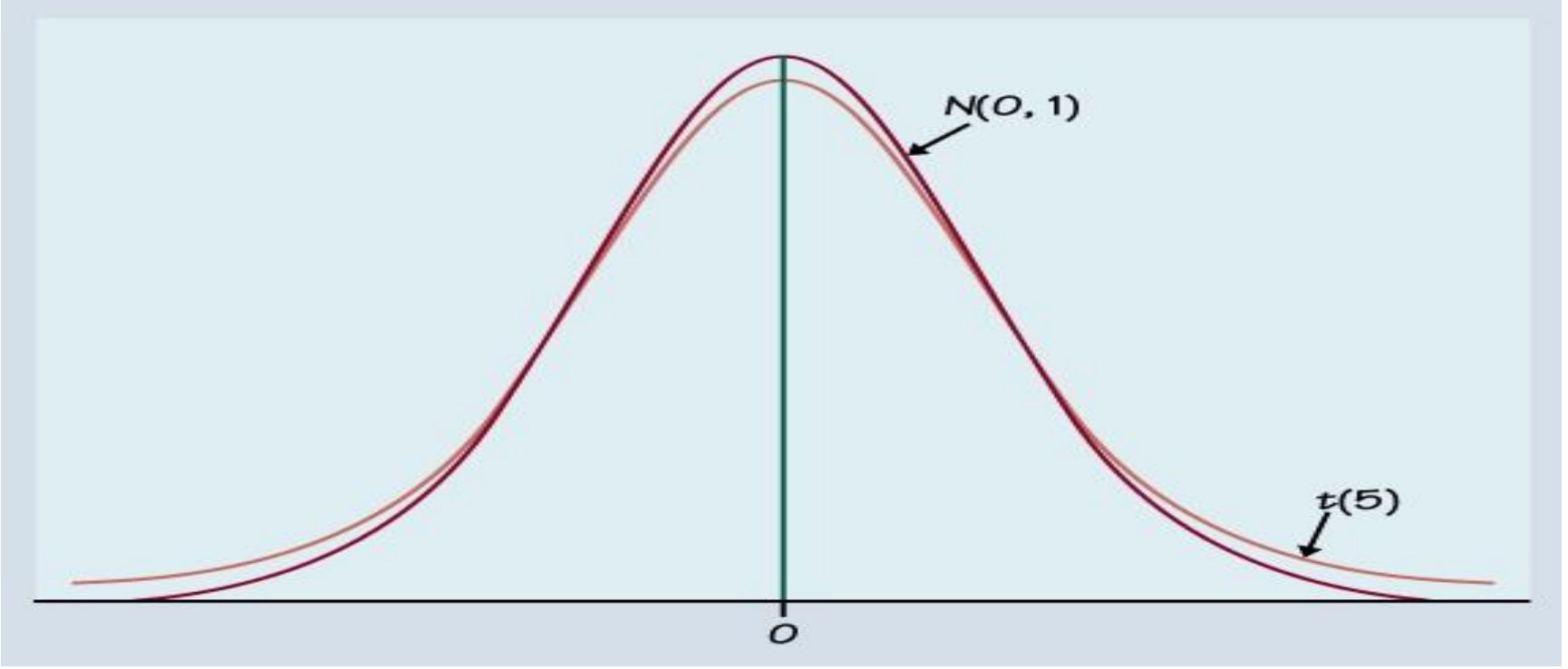
- Central limit theorem: the procedures are approximately correct for non-normal distributions when the sample is large.

When σ is unknown

- Use the one-sample *t* statistic with *n* –1 degrees of freedom

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

t vs z graph



One sample *t* procedures

A level C confidence interval for μ

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

where t^* is the upper (1-c)/2 critical value for the t(n-1) distribution

The interval has the form

$$\text{estimate} + t^* SE_{\text{estimate}}$$

Hypothesis test

Hypothesis testing of H_o:$\mu = \mu_o$

- State the hypothesis
- Calculate the test statistic
- Compute the *P*-value
- Compare the *P*-value to α and conclude.

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

or

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Robustness

- *Sample size less than 15*. Use *t* procedures only if the data are close to normal.

- *Sample size at least 15*. The *t* procedures can be used except in the presence of outliers or strong skewness.

- *Large samples (*n > 40). The *t* procedures can be used even for clearly skewed distributions.

Matched pairs

Use these one-sample procedures to analyze matched pair data by first taking the difference within each matched pair to produce a single sample.

Example 1

- A 95% confidence interval for the mean of a population is computed from a random sample and found to be 9 ± 3. We may conclude

  - A) that there is a 95% probability that the population mean is between 6 and 12.
  - B) That if we took many more random samples and computed a confidence interval from each, approximately 95% of these intervals would contain μ.
  - C) All of the above.

Example 2

- Many food manufacturers fortify their food products by adding vitamins. The following data are the amounts of vitamin C measured in mg/100g of corn soy blend for a random sample of size 8.

26 31 23 22 11 22 14 31

- Sample mean = 22.5 ; s = 7.19

- Find a 95% C.I. For the mean vitamin C content of the corn soy blend.
- The US Agency for International Development specifies a mean vitamin C content of 30 mg. Test the hypothesis that mean vitamin C content does not conform to these specifications at significance level 5%.

- 95% C.I. For the mean vitamin C content of the corn soy blend

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$
$$= 22.5 \pm t_{0.975,7} * 7.19/\sqrt{8}$$
$$= (16.5, 28.5)$$

- Hypothesis test:
  - H_0: Vc=30 Vs H_a: Vc≠ 30

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{22.5 - 30}{7.19/\sqrt{8}} = -2.95$$

p-value is 0.011, therefore reject null hypothesis at 5% significance level.

Two Sample Design

Comparative Studies

- To see if there is a difference between two groups, we can measure the same variable in both groups and compare the results

- The groups may receive different treatments in an experiment or we may be comparing two groups (such as males and females) in a survey

Two-sample problems

Goal: to compare the responses to two treatments or the characteristics of two populations

Ex 1: to study the effect of calcium on blood pressure, conduct a randomized comparative experiment. One group receives calcium and a control group gets a placebo

Ex 2: a psychologist develops a test to measure sensory integration. She compares the sensory integration scores of male and female preschoolers

Assumptions

- We have **two random samples**, from two distinct populations. The samples are **independent**. We measure the **same variable** for both samples.

- Both populations are **normally distributed**. The means and standard deviations of the populations are unknown.

Population	Variable	Mean	Standard deviation
1	x_1	μ_1	σ_1
2	x_2	μ_2	σ_2

Inference

- Compare the two population means Either
  - Using a confidence interval for their difference $\mu_1 - \mu_2$
  - or testing the hypothesis of no difference, H_0: $\mu_1 = \mu_2$
- Sample data:

Population	Sample size	Sample Mean	Sample standard deviation
1	n_1	$\bar{x}_1$	s_1
2	n_2	$\bar{x}_2$	s_2

To do inference on $\mu_1 - \mu_2$ we use $\bar{x}_1 - \bar{x}_2$.

Example

21 healthy men took park in a double blind blood pressure experiment.

10 received calcium supplement

11 received a placebo

Response variable is the change in systolic pressure after 12 weeks. An increase appears as a negative response

Group 1 (calcium):

7 -4 18 17 -3 -5 1 10 11 -2

Group 2 (placebo):

-1 12 -1 -3 3 -5 5 2 -11 -1 -3

Summary

Group	Treatment	n	$\bar{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	-0.273	5.901

Is this good evidence that calcium decreases blood pressure more than a placebo does?

The hypotheses are:

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

We can also write these hypotheses as:

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

To do inference on $\mu_1 - \mu_2$, we use

Samplingdistribution of $\bar{x}_1 - \bar{x}_2$

- Mean = $\mu_1 - \mu_2$
- Standard deviation = $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
- Distribution is normal

z-statistic:

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

t-statistic

σ_1 and σ_2 are unknown

Substitute s_1 and s_2.

t-statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Inference procedures: CI

Confidence interval for $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where t^* is the upper (1-C)/2 critical value.

C.I. has the form

estimate $\pm$ t^* S.E.(estimate)

Inference procedures: Testing

To test the hypothesis H_0: $\mu_1 - \mu_2 = 0$ compute the two-sample t statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Degrees of Freedom

Use the two-sample *t* statistic with the *t*(*k*) distribution.

Conservative degrees of freedom:

$$k = \min(n_1 - 1, n_2 - 1)$$

More accurate degrees of freedom:

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

Use conservative degrees of freedom when doing calculations by hand. Software (such as Excel) uses the more accurate formula.

Blood pressure example

Group	Treatment	n	$\bar{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	-0.273	5.901

a) Compute a 95% confidence interval for $\mu_1 - \mu_2$. Use the conservative df calculation.

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$= (5.00 - (-0.273)) \pm \sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}$$

$$= (-2.16, 12.71)$$

Hypothesis Test

Group	Treatment	n	$\bar{x}$	s
1	Calcium	10	5.000	8.743
2	Placebo	11	-0.273	5.901

Determine whether there is good evidence that calcium decreases blood pressure more than a placebo. Test

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

$$H_a: \mu_1 - \mu_2 > 0$$

Test statistic:

$$t = \frac{(5.00 - (-0.273)) - 0}{\sqrt{\frac{8.743^2}{10} + \frac{5.901^2}{11}}} = 1.6$$

Computer Output

Two Sample T-Test and Confidence Interval

Two sample T for Calcium vs Placebo

	N	Mean	StDev	SE Mean
Calcium	10	5.00	8.74	2.8
Placebo	11	-0.27	5.90	1.8

95% CI for mu Calcium - mu Placebo:

(-1.7, 12.3) ***Note: the software uses the exact calculation for the df (slide 24)***

T-Test mu Calcium = mu Placebo (vs >):

T = 1.60 P = 0.065 DF = 15

Robustness

Two sample procedures are appropriate if

- The data are random samples from the populations of interest
- $n_1 + n_2 < 15$. Use *t* procedures if the data are close to normal. Do not use *t* if the data are clearly non-normal or outliers are present
- $15 \leq n_1 + n_2 < 40$. The *t* procedures can be used except in the presence of outliers or strong skewness.
- $n_1 + n_2 \geq 40$. The *t* procedures can be used even for clearly skewed distributions.