95% Confidence Interval & p value

Nguyen Quang Vinh - Nguyen Thi Tu Van

Introduction

Statistics

- Descriptive:
 - Measures of central tendency
 - Measures of dispersion
- Inferential:
 - Estimation
 - Hypothesis testing

MEASURES OF CENTRAL TENDENCY

The Mean (arithmetic mean) Samplemean: $\bar{x} = \frac{\sum x}{n}$ Population mean: $\mu = \frac{\sum x}{N}$

- Uniqueness
- Simplicity
- Extreme value & The Mean (!)

The Median (Md)

- Uniqueness
- Simplicity
- Extreme value & The Median

The Midrange (Mr)
$$Mr = \frac{L+H}{2}$$

- Less popular than mean and median
- An easy to grasp
- Simplicity
- Extreme value & The Midrange (!)

Mode (Mo)

• Use for describing qualitativ e data

MEASURES OF DISPERSION (dispersion, variation, spread, scatter)

- 1. Range
- 2. Variance
- 3. Standard Deviation
- 4. Coefficient of Variance

MEASURES OF DISPERSION

3. Standard Deviation

Sample Standard Deviation, s:

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} = \sqrt{\frac{1}{n - 1} \left[\sum x^2 - \frac{(\sum x)^2}{n}\right]}$$

Population Standard Deviation, σ :

$$\sigma = \sqrt{\frac{\sum (x - \mu)}{N}}$$

4. Coefficient of Variation* :

$$C.V. = \frac{s}{\overline{x}}.100$$

* for data sets with extremevariation it is possible to obtain a C.V. > 100%

MEASURES OF DISPERSION (dispersion, variation, spread, scatter)

1. Range = H - L

2. Variance

Sample variance, s^2 :

$$s^{2} = \frac{\sum (x - \bar{x})^{2}}{n - 1} = \frac{1}{n - 1} \left[\sum x^{2} - \frac{(\sum x)^{2}}{n} \right]$$

Population variance, σ^2 :

$$\sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

SAMPLING DISTRIBUTION

The probability distribution of a sample statistic is its sampling distribution.

Standard error of the mean $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$

is called:

- the standard error of the mean (S.E.M.), or
- the standard error (S.E.), or
- the standard deviation of the probability distribution for the mean

Estimation

Estimator → Parameter

Parameters:

- Population mean
- Population proportion
- Population variance

- The difference between 2 means
- The difference between 2 proportions
- The ratio of 2 variances

Estimator → Parameter

 Each of these parameters: Point estimate
 Interval estimate

CONFIDENCE INTERVAL FOR A POPULATION MEAN

In general, an interval estimate is obtained by the formula

estimator ± (reliability coefficient) x (standard error)

In particular, when sampling is from a *normal* distribution with *known variance*, an interval estimate for μ may be expressed as:

$$\overline{x} \pm z_{\alpha/2} \sigma_{\overline{x}}$$

How to interpret the interval given by this expression

- In repeated sampling, from a normally distributed population, 100(1 α)% of all intervals of the form will in the long run include the population mean, μ
- The quantity 1 α, is called the *confidence coefficient*, &

The interval $\overline{x} \pm z_{\alpha/2} \sigma_{\overline{x}}$, is called the *confidence interval* for μ

The practical interpretation

We are 100(1 - α)% confident that the single computed interval

$$\overline{x} \pm z_{\alpha/2} \sigma_{\overline{x}}$$

contains the population mean, μ

• *E* = margin error = maximum error = practical / clinical acceptable error:

$$E = z_{\alpha/2} \sigma_{\overline{x}} = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

DISTRIBUTION OF THE SAMPLE MEAN \overline{X}

• Sampling is from a non-normally distributed population:

Central limit theorem:

Given a population of any non-normal functional form with a mean, μ , and finite variance, σ^2 ; the sampling distribution of \overline{X} , computed from samples of size n from this population, will be approximately normally distributed with mean, μ , and variance, σ^2/n , when the sample size is **large**.

How large does the sample have to be in order for the central limit theorem to apply?

- There is no one answer, since the size of the sample needed depends on the extent of non-normality present in the population.
- Rule of thumb: in most practical situations, a sample of size 30 is satisfactory.
- In general the approximation to normality of the sampling distribution becomes better and better as the sample size increases.

SAMPLING FROM NONNORMAL POPULATIONS

- \rightarrow Sampling from:
- Nonnormally distributed populations
- Populations whose functional forms are not known
- \rightarrow Taking large enough sample \rightarrow Central limit theorem

C.I. FOR THE DIFFERENCE BETWEEN 2 SAMPLE MEANS

When the **population variances are known**, the 100(1 - α)% confidence interval for μ_1 - μ_2 is given by

$$(\overline{x}_1 - \overline{x}_2) \pm z_{\alpha/2} \sigma_{\overline{x}_1 - \overline{x}_2}$$

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Sampling from nonnormal populations: taking large enough samples n_1 , $n_2 \rightarrow Central limit theorem$

C.I. FOR THE DIFFERENCE BETWEEN 2 SAMPLE MEANS

When the **population variances are unknown**, we distinguish between 2 situations:

(1) The population variances are equal

• If the assumption of equal population variances is justified, a *pooled estimate* of the common variance is given by the formula: $s_1^2 - (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2$

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

• The 100(1 - α)% C.I. for μ_1 - μ_2 is given by:

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

C.I. FOR THE DIFFERENCE BETWEEN 2 SAMPLE MEANS

(2) The population variances are not equal

• When one is reluctant to assume that the variances of 2 populations of interest are equal, the 100(1 - α)% C.I. for $\mu_1 - \mu_2$ is given by

$$(\overline{x}_{1} - \overline{x}_{2}) \pm t'_{\alpha/2} \sqrt{\frac{s_{1}^{2}}{n_{1}} + \frac{s_{2}^{2}}{n_{2}}} \qquad w_{1} = \frac{s_{1}^{2}}{n_{1}}$$
$$w_{2} = \frac{w_{1}t_{1} + w_{2}t_{2}}{w_{1} + w_{2}} \qquad t_{1} = t_{\alpha/2, n_{1}-1}$$
$$t_{2} = t_{\alpha/2, n_{2}-1}$$

 $t'_{\alpha/2}$ is called Cochran reliability factor

C.I. FOR A POPULATION PROPORTION

• The sample proportion \hat{p} is used as the point estimator of the population proportion p, then a C.I. is obtained by the general formula:

estimator ± (reliability coefficient) x (standard error)

• When np & n(1-p) are greater than 5, the sampling distribution of \hat{p} is # the normal distribution.

Therefore, the reliability coefficient is some value of z from the standard normal distribution.

• The standard error is $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$

Since *p* is unknown, we must used \hat{p} as an estimate. Thus σ is estimated by $\sigma_{\hat{p}} = \sqrt{\hat{p}(1-\hat{p})/n}$

C.I. FOR A POPULATION PROPORTION

- The 100 (1 α)% C.I. for *p* is given by $\hat{p} \pm z_{\alpha/2} \sigma_{\hat{p}}$ $\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1-\hat{p})/n}$
- Hence, the 95% C.I. for p is $\hat{p}\pm 1.96\sigma_{\hat{p}}$ $\hat{p}\pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$

C.I. FOR THE DIFFERENCE BETWEEN 2 POPULATION PROPORTIONS

$$(\hat{p}_1 - \hat{p}_2) \rightarrow (p_1 - p_2)$$

- When: $n_1 \& n_2$ are large & the population proportions, $p_1 - p_2$, are not too close to 0 or 1
- → the central limit theorem applies & normal distribution theory may be employed to obtain C.I.

$$S.E. = \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \frac{1}{n_2}$$
$$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \sqrt{\frac{1 + 100 (1 - \alpha)\%}{C.1. \text{ for } p_1 - p_2}}$$

Notes

- * As a rule σ^2 is unknown $\rightarrow \sigma^2$ has to be estimated
- * The most frequently used sources of estimates for σ^2 are the following:
 - 1. A pilot sample
 - 2. Previous or similar studies
 - σ ≈ R/4 (or R/6) (approximately normal distributed & some knowledge of the smallest and largest value of the variable in the population)
 - 4. s ≈ IQR/1.35

C.I. FOR THE VARIANCE OF A NORMALLY DISTRIBUTED POPULATION The $(100 - \alpha)$ % C.I. for $(n-1)s^2/\sigma^2$ $\chi_{\alpha/2}^{2} < (n-1)s^{2}/\sigma^{2} < \chi_{1-\alpha/2}^{2} < \\ \Leftrightarrow \frac{(n-1)s^{2}}{\chi_{1-\alpha/2}^{2}} < \sigma^{2} < \frac{(n-1)s^{2}}{\chi_{\alpha/2}^{2}} < \\ \end{cases}$ The $(100 - \alpha)$ % C.I. for σ^2 $\Leftrightarrow \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}$ The (100 - α)% C.I. for σ

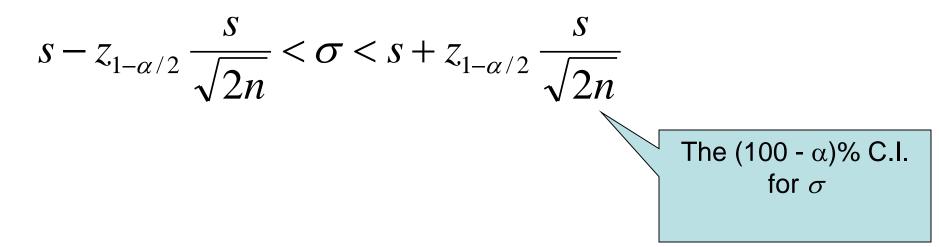
C.I. FOR THE VARIANCE OF A NORMALLY DISTRIBUTED POPULATION Drawbacks

Although this method of constructing C.Is. for σ^2 is widely used, it is not without drawbacks:

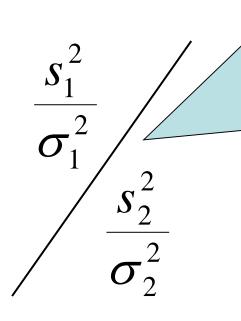
- The assumption of the *normality* of the population from which the sample is drawn is <u>crucial</u>.
- The estimator is not in the center of the C.I., because the χ^2 distribution, unlike the normal, is not symmetric.

C.I. FOR THE VARIANCE OF A NORMALLY DISTRIBUTED POPULATION

If the sample size is large :

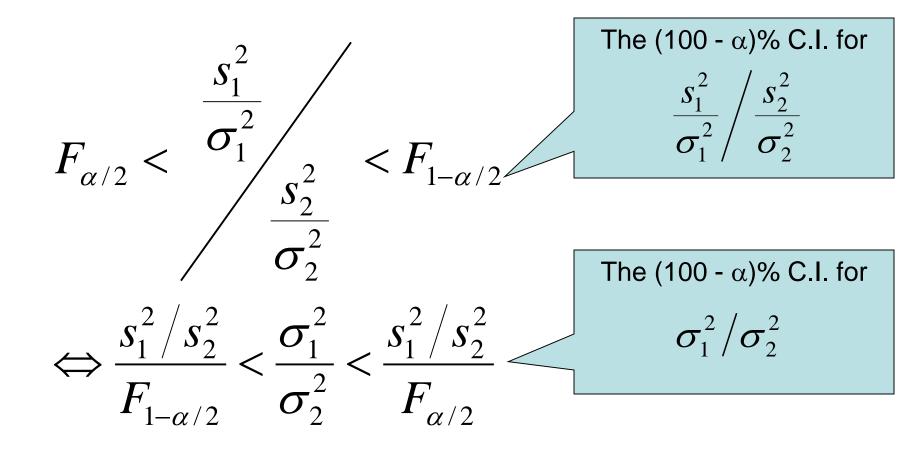


C.I. FOR THE RATIO OF THE VARIANCES OF 2 NORMALLY DISTRIBUTED POPULATIONS



follows the *F* distribution. With the assumptions: $s_1^2 \& s_2^2$ are computed from *independent* samples of size $n_1 \& n_2$, respectively, drawn from 2 *normally* distributed populations

C.I. FOR THE RATIO OF THE VARIANCES OF 2 NORMALLY DISTRIBUTED POPULATIONS



Note $F_{(1-\alpha),\nu_1,\nu_2} = \frac{1}{F_{\alpha,\nu_2,\nu_1}}$

Where:

 $v_1 = n_1 - 1$ (numerator degrees of freedom) $v_2 = n_2 - 1$ (denominator degrees of freedom)

Hypothesis Testing

Introduction

- Reaching a decision concerning a population by examining a sample from that population
- Two types of hypotheses:

(1) Research Hypotheses:

- The conjecture or supposition
- It may be the results of years of observation
- Leads directly to Statistical H.

(2) Statistical Hypotheses:

Hypotheses are stated in such a way that they may be evaluated by appropriate statistical techniques

Conditions under which type I & type II errors may be committed (the four possibilities)		Truth in the population	
		Association between predictor & outcome (H _o false)	No association between predictor & outcome (H _o true)
The results in the study sample \rightarrow Conclusion:	Reject H _o	Correct decision	Type I error
	Fail to reject H _o	Type II error	Correct decision

Note

- H_0 , $H_A \& \alpha$ must be defined <u>before</u> we observe any data In other words, do not let the data dictate our hypotheses
- The smaller α is, the large β is \rightarrow if we want β to be small, we choose a large value of α
- For most situations the range of acceptable α values is .01 to .1
- If there is no significant difference between the effects a type I error vs. a type II error, researchers often choose α = .05

The Five-Step Procedure for Hypothesis Testing

- Step 1: Set up H_o H_A
- Step 2: Define the test statistic, and its distribution
- Step 3: Define a rejection region: having determined a value for α
- Step 4: Calculate the value of the test statistic, and carry out the test → p value
 State our decision: to reject H_o or to fail to reject H_o
- Step 5: Give a conclusion free of statistical jargon.

- If H_0 is rejected, we conclude that H_A is true.
- If H₀ is not rejected, we conclude that H₀ may be true.

- We avoid using the word "accept" in the case that the H_o is not rejected, we should say that the H_o is "not rejected"

Summary

- (1) Hypothesis testing, in general, is not a procedure to proof a hypothesis It merely indicates whether the null hypothesis is supported or not supported by the available data
- (2) What we expect to be able to conclude as a result of the test usually should be placed in the H_A
- (3) The H_O is the hypothesis that is tested
- (4) The $H_0 \& H_A$ are complementary