Lecture Series on Biostatistics

Inferential Statistics-Estimation

By

Dr. Bijaya Bhusan Nanda,
M. Sc (Gold Medalist) Ph. D. (Stat.)
Topper Orissa Statistics & Economics Services, 1988
bijayabnanda@yahoo.com

CONTENTS

- Introduction
- Confidential interval for a population mean
- The t distribution
- Confidence interval for the difference between two population mean
- Confidence interval for a population proportion
- Confidence interval for the difference between two population proportion

Introduction

Statistical Inference

- It is the procedure by which we reach a conclusion about a population on the basis of information contained in the sample drawn from that population.
- Two broad areas of statistical inference
  - Estimation
  - Hypothesis testing.
- **Estimation**- It is the process of calculating some statistics based upon the sample data drawn from a certain population. The statistics is used as an approximation to the population parameter.

Types of Estimate

For each of parameters, we can have

- **Point Estimate**
- **Interval estimate**

- A point estimate is a single numerical value used to estimate the corresponding population parameter.
- An interval estimate consists of two numerical values defining a range of values that, with a specified degree of confidence, includes the parameter being estimated.

Choosing an Appropriate Estimation

- A single computed value is an estimate.
- An estimator usually presented as a formula. For example

$$\bar{x} = \frac{\sum x_i}{n}$$ is the estimator of population mean

- The single numerical value that results from evaluating this formula is called an estimate of the parameter μ.
- E(T) is obtained by taking the average value of T computed from all possible sample i.e. $E(T) = \mu_T$

Criteria of a good estimator:

- There can be more than one estimator for the parameter. For example population mean can be estimated by the sample median or sample mean.
- But a good estimator should fulfill certain criteria. One such criterion of a good estimator is unbiasedness.
- An estimator T of the parameter θ is said to be an unbiased estimator of θ if $E(T)= \theta$
- Sample mean is an unbiased estimator of population mean.
- Sample proportion is an unbiased estimate of population proportion.
- The difference between two sample means is an unbiased estimate of difference between the population mean.
- The difference between two sample proportions is an unbiased estimate of difference between the population proportion.

Sampled Population and Target Population

- Sampled population is the population from which one actually draws a sample.
- The target population is the population about which one wishes to make inference.
- Statistical inference procedure allows one to make inference about sampled population (provided proper sampling methods have been employed)
- Only when sampled population and the target population are the same, it is possible to reach statistical inference about the target population.

Random and Non random Samples

- The strict validity of the statistical procedures discussed depends on the assumptions of random sample.

- In real world applications it is impossible or impractical to use truly random samples.
- If the researchers had to depend on randomly selected materials, very little research of this type would be conducted.
- Therefore, non statistical considerations must play a past in the generalizations process.
- Researchers may contend that samples actually used are equivalent to simple random samples, since there is no reasons to believe that material actually used is not representative of the population about which inferences are desired.
- In many health research projects, samples of convenience, rather than random samples are employed.

- Generalization must be made on non statistical consideration.
- Consequences of such generalization, however may range from misleading to disastrous.
- In some situation it is possible to introduce randomization into experiment even though available subjects are not randomly selected from well defined population.
- Example: Allocating subjects to treatments in a random manner.

Confidential interval for a population mean

- Say $\bar{x}$ is the sample mean from a random sample of size n drawn from a normally distributed population.
- $\bar{x}$ is an unbiased estimator of the population mean
- $$E(\bar{x}) = \mu$$
- But because of sampling fluctuation $\bar{x}$ can't be expected to be equal to μ.
- Therefore, we should have an interval estimate μ.
- For the interval estimate we must have knowledge of sampling distribution of $\bar{x}$.
- Sampling distribution of $\bar{x}$ for a normal population is as follows

$$\bar{x} \approx N(\mu_{\bar{x}}, \sigma_{\bar{x}}) i.e. N(\mu, \sigma/\sqrt{n})$$

- Following the normal probability distribution the interval estimate for μ is as follows

95% confidence interval for $\mu = \mu \pm 2\sigma_{\bar{x}}$

since we don't know μ with the help of sample mean $\bar{x}$ we have the 95% confidence interval for μ as

$$\bar{x} \pm 2\sigma_{\bar{x}} = \bar{x} \pm 2\sigma/\sqrt{n}$$

If we don't know σ the 95% confidence interval for $\mu = \bar{x} \pm 2s/\sqrt{n}$ where s= the standard deviation based on the sample.

Interval estimate Component

- In general an interval estimate may be expressed as follows

  estimator ±(reliability coefficient)× (standard error)

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

$$\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$$

where z(1-α/2) is the value of z to the left of which lies 1- α/2 and to the right of which lies α/2 of the area under its curve.

Interpreting Confidence Interval:

- **Probabilistic Interpretation:** In repeated sampling, from a normally distributed population with a known standard deviation, 100(1- α) percent of all intervals of the form $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$ will in the long run include the population mean μ.
- **Practical Interpretation:** When sampling is from a normally distributed population with known standard deviation, we are 100(1- α) % confident that the single computed interval, $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$, contains the population mean μ.
- **Precision:** The quantity obtained by multiplying the reliability factor by the standard error of the mean is called the precision of the estimate. This quantity is also called the margin of error.

Example

A physical therapist wished to estimate, with 99 percent confidence , the mean maximum strength of a particular muscle in a certain group of individuals. He is willing to assume that strength scores are approximately normally distributed with a variance of 144. A sample of 15 subjects who participated in the experiment yielded a mean of 84.3.

Solution:

The z value corresponding to a confidence coefficient of 0.99 is found in Table D to be 2.58. This is our reliability coefficient. The standard error is $\sigma_{\bar{x}} = 12/\sqrt{15} = 3.0984$ .Our 99%

Confidence interval for μ, then

$$84.3 \pm 2.58(3.0984)$$

$$84.3 \pm 8.0$$

$$76.3, 92.3$$

We say we are 99% confidence that the population mean is between 76.3 and 92.3 since, in repeated sampling, 99% of all intervals that could be constructed in the manner just described would include the population mean.

The t distribution

- Estimation of confidence interval using standard normal z distribution applies when population variance is known. Usually, the knowledge of population variance is not known.
- This condition presents a problem to construct confidence intervals.

Although the statistic $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ is normally distributed

or approximately normally distributed when n is large, regardless of the functional form of the population, we can't make use of this fact because σ is not known.

- In this case we may use of $S = \sqrt{\sum (x_i - x)^2 / (n-1)}$

  as an estimator of σ. When sample size is large, say, greater than 30, s is a good approximate of σ is substantial and we may use normal distribution theory to construct confidence interval.
- When the sample size is small, ≤30, we make use of Student's t distribution.

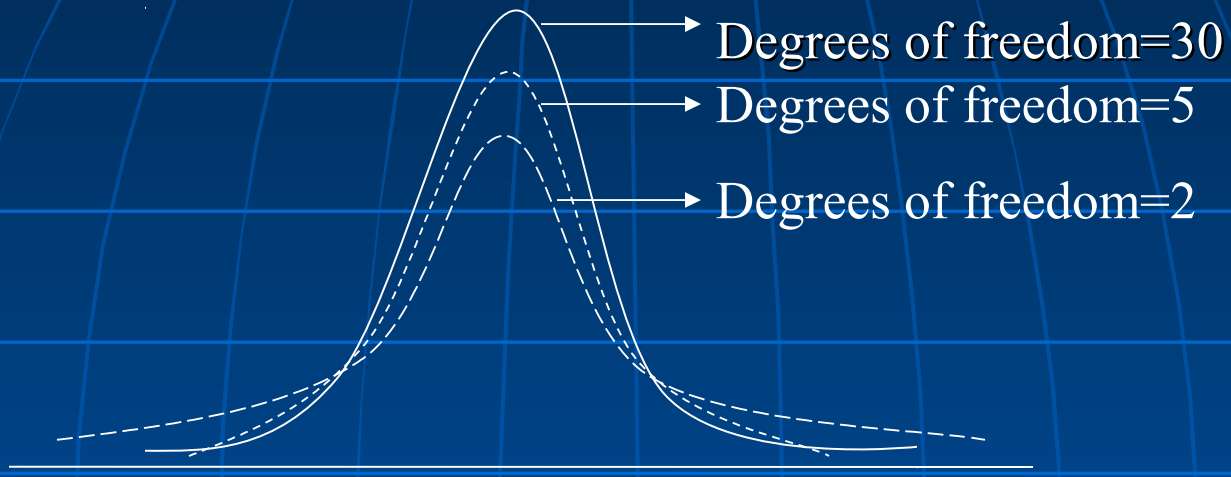
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

The quantity $\frac{s}{\sqrt{n}}$ follows this distribution.

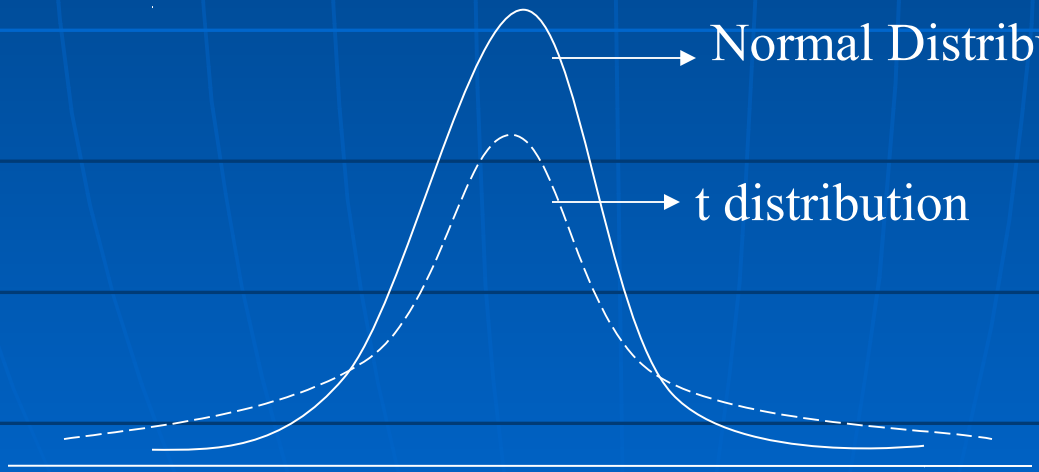
Properties of the t distribution:

- It has a mean of 0.
- Symmetrical about the mean.
- In general, it has a variance greater than 1, but the variance approaches 1 as the sample size becomes large. For df >2, the variance for t distribution is df /(df-2).
- The variable t ranges from $-\infty$ to $+\infty$.
- The t distribution is really a family of distributions, since there is a different distribution for each sample value of n-1, the divisor used in computing s^2. We recall that n-1 is referred to as degrees of freedom.
- Compared to normal distribution is less peaked at center and has higher tails t distribution approaches the normal distribution an n-1 approaches infinity.
- The t distribution approaches the normal distribution as n-1 approaches infinity.

The t distribution for different degrees of freedom



Comparison of normal distribution and t distribution



Confidence Intervals Using t:

estimator ± (reliability coefficient) (standard error)

The source of reliability co-efficient is 't' distribution rather than standard normal z distribution.

- When sampling is from a normal distribution whose standard deviation , σ, is unknown the $100(1-\alpha)\%$ confidence interval for the population mean μ, is given by

$$\bar{x} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

- A moderate departures of the sampled population from the normally can be tolerated in using the 't' distribution. An assumption of mound shaped population distribution be tenable.

Example:

Maureen McCauley conducted a study to evaluate the effect of on the job body mechanics instruction on the work performance of newly employed young workers (A-1). She used two randomly selected groups of subjects, an experimental group and a control group. The experimental group received one hour of back school training provided on occupational therapist. The control group did not receive this training. A criterion referenced Body Mechanics Evaluation Checklist was used to evaluate each workers lifting, lowering, pulling, and transferring of objects in the work environment. A correctly performed task received a score of 1. The 15 control subjects made a mean score of 11.53 on the evaluation with a standard deviation of 3.681. We assume that these 15 controls behave as a random sample from a population of similar subjects. We wish to use these sample data to estimate the mean score fro the population.

Solution:

We may assume sample mean, 11.53, as a point estimate of population mean but since the population standard deviation is unknown, we must assume the population of values to be at least approximately normally distributed before constructing a confidence interval for μ.

Let us assume an assumption is reasonable and that a 95% confidence interval is desired. Our estimator $\bar{x}$ is and our standard error is

$$s/\sqrt{n} = 3.681/\sqrt{15} = 0.9564$$

We need to find reliability coefficient, the value of t associated with a confidence coefficient of 0.95 and n-1=14 degrees of freedom.

Confidence interval is equally divided in to two tails i.e.0.025 .In Table E the tabulated value of t is 2.1448, which is our reliability coefficient.

Then our 95% confidence interval is as follows

$$11.53 \pm 2.1448(0.9504)$$

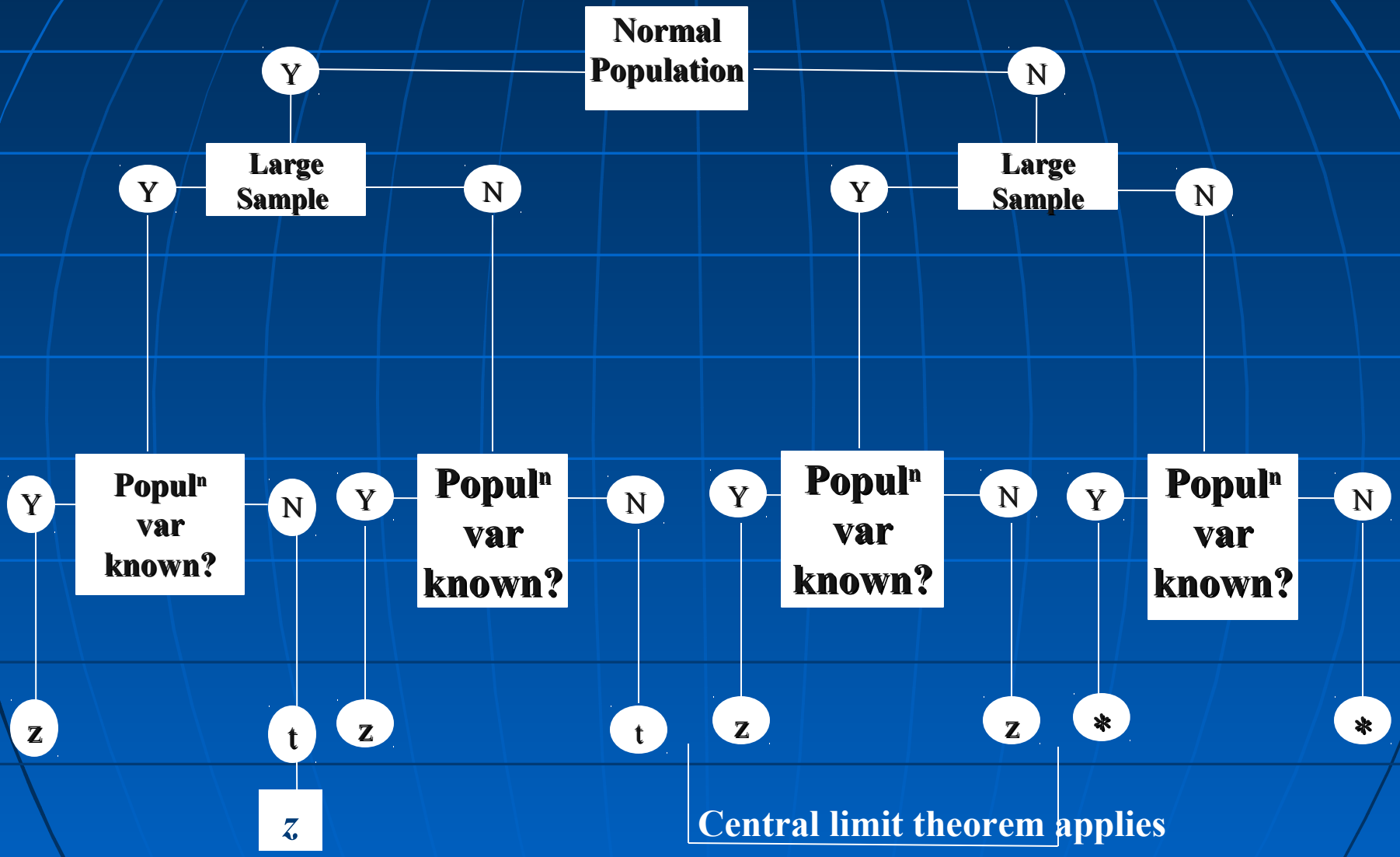
$$11.53 \pm 2.04$$

$$9.49, 13.57$$

Deciding Between z and t:

To make an appropriate choice between z and we must consider whether the sampled population is normally distributed, and whether the population variance is known.

Flowchart for use in deciding between z and t when making inferences about population means



* = use nonparametric Procedure

Confidence interval for the difference between two population means

Normal Population

- Say μ_1-μ_2 = the difference between two normal population means.
- $\bar{x}_1 - \bar{x}_2$ = the difference between two independent sample means drawn from the two populations respectively.

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

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

Where σ_1^2 & σ_2^2 are the variance of the population 1&2 and n_1, n_2 are the size of the SRS drawn from population 1&2 respectively.

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

$$(\bar{x}_1 - \bar{x}_2) \pm z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- When the confidence interval includes zero, we say that the population means may be equal when it doesn't includes zero we say that the interval provides evidence that the two population means are not equal.

Example:

A research team is interested in the difference between serum uric acid levels in patients with and without Down's syndrome. In a large hospital for the treatment of the mentally retarded, a sample of 12 individuals with Down's syndrome yielded a mean of 4.5mg/100ml. In a general hospital a sample of 15 normal individuals of the same age and sex were found to have a mean value of 3.4. If it is reasonable to assume that the two populations of values are normally distributed with variances equal to 1 and 1.5, find the 95% confidence interval for $\mu_1 - \mu_2$.

Solution:

For a point estimate of $\mu_1 - \mu_2$ we use

$$\bar{x}_1 - \bar{x}_2 = 4.5 - 3.4 = 1.1$$

The reliability coefficient corresponding to 0.95 is found in Table D to be 1.96. The standard error

$$\sigma(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1}{12} + \frac{1.5}{15}} = 0.4282$$

The 95 percent confidence interval, then, is

$$1.1 \pm 1.96(0.4282)$$

$$1.1 \pm 0.84$$

$$0.26, 1.94$$

The difference $\mu_1 - \mu_2$ is some where between 0.26 and 1.94, because, in repeated sampling, 95% of the intervals constructed in this manner would include the difference between the true means. Since the interval does not include zero, we conclude that the two population means are not equal.

Sampling from Nonnormal Population:

If sample sizes n_1 and n_2 are large, we may construct the confidence interval the same way as we do incase of normal population in accordance with the Central Limit Theorem.

Example:

Motivated by an awareness of the existence of a body of controversial literature suggesting that stress, anxiety, and depression are harmful to the immune system, Gorman et al. (A-5) conducted a study in which the subjects were homosexual men and some of whom were HIV positive and some of whom were HIV negative. Data were collected on a wide variety of medical, immunological, psychiatric, and neurological measures, one of which was the number of CD4+ cells in the blood. The mean number of CD4+ cells for the 112 men with HIV infection was 401.8 with a standard deviation of 226.4. For the 75 men without HIV infection the mean and standard deviation were 828.2 and 274.9, respectively. We wish to construct a 99% confidence interval for the difference between population means.

Solution:

Here we are using z statistic as the reliability factor in the construction of our confidence interval. Since the population standard deviation are not given, we will use the sample standard deviations to estimate them. The point estimate for the difference between population means is the difference between sample means, 882.2-401.8=426.4. In Table D, we find the reliability factor to be 2.58. The estimated standard error is

$$s_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{274.9^2}{75} + \frac{226.4^2}{112}} = 38.279$$

Our 99% confidence interval for the difference between population means is

$$426.4 \pm 2.58(38.279)$$

$$327.6, 525.2$$

We are 99% confident that the mean number of CD4+ cells in HIV +ve males differs from the mean for HIV –ve males by some where between 327.6 and 525.2.

Confidence interval using 't' distribution:

- Say the variances of the two sampled population are not known.
- The two sampled population are normally distributed.
- The difference between the two population means can be estimated through confidence interval using 't' distribution as the source of reliability factor.

Situation-I-Equal population variances:

The 100(1-α)% confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(1-\alpha/2)} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

If the two sample sizes are unequal , the weighted average takes advantages of the additional information provided by the larger sample. The pooled estimate is given by

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

The standard error of the estimate , then, is given by

$$S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

The ‘t’ follows a Student’s t distribution with n_1+n_2-2 d.f.

Example:

The purpose of a study by Stone et al.(A-6) was to determine the effects of long-term exercise intervention on corporate executives enrolled in a supervised fitness program.

Data were collected on 13 subjects (the exercise group) who voluntarily entered a supervised exercise program and remained active for an average of 13 years and 17 subjects (the secondary group) who elected not to join the fitness program. Among the data collected on the subjects was maximum number of sit-ups completed in 30 seconds. The exercise group had a mean and standard deviation for this variable of 21.0 and 4.9 respectively. The mean and standard deviation for the secondary group were 12.1 and 5.6 respectively. We assume that the two population of overall muscle condition measures are approximately normally distributed and that the two population variances are equal. We wish to construct a 95% confidence interval for the difference between the means of the populations represented by these two samples.

Solution:

The 95% confident that the difference between population means is somewhere between 4.9, 12.9 .

Situation-II-Unequal population variances:

An approximate 100(1-α)% confidence interval **for** $\mu_1 - \mu_2$ is given by

$$\bar{x}_1 - \bar{x}_2 \pm t'_{(1-\alpha/2)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The solution proposed by Cochran consists of computing the reliability factor $t'_{(1-\alpha/2)} = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$

Where $w_1 = s_1^2/n_1$, $w_2 = s_2^2/n_2$, $t_1 = t_{1-\alpha/2}$ for n_1-1 degrees of freedom, and $t_2 = t_{1-\alpha/2}$ for n_2-1 degrees of freedom.

Example:

in the study by Stone et al.(A-6)described in previous example , the investigator also reported the following information on a measure of overall muscle condition scores made by the subjects:

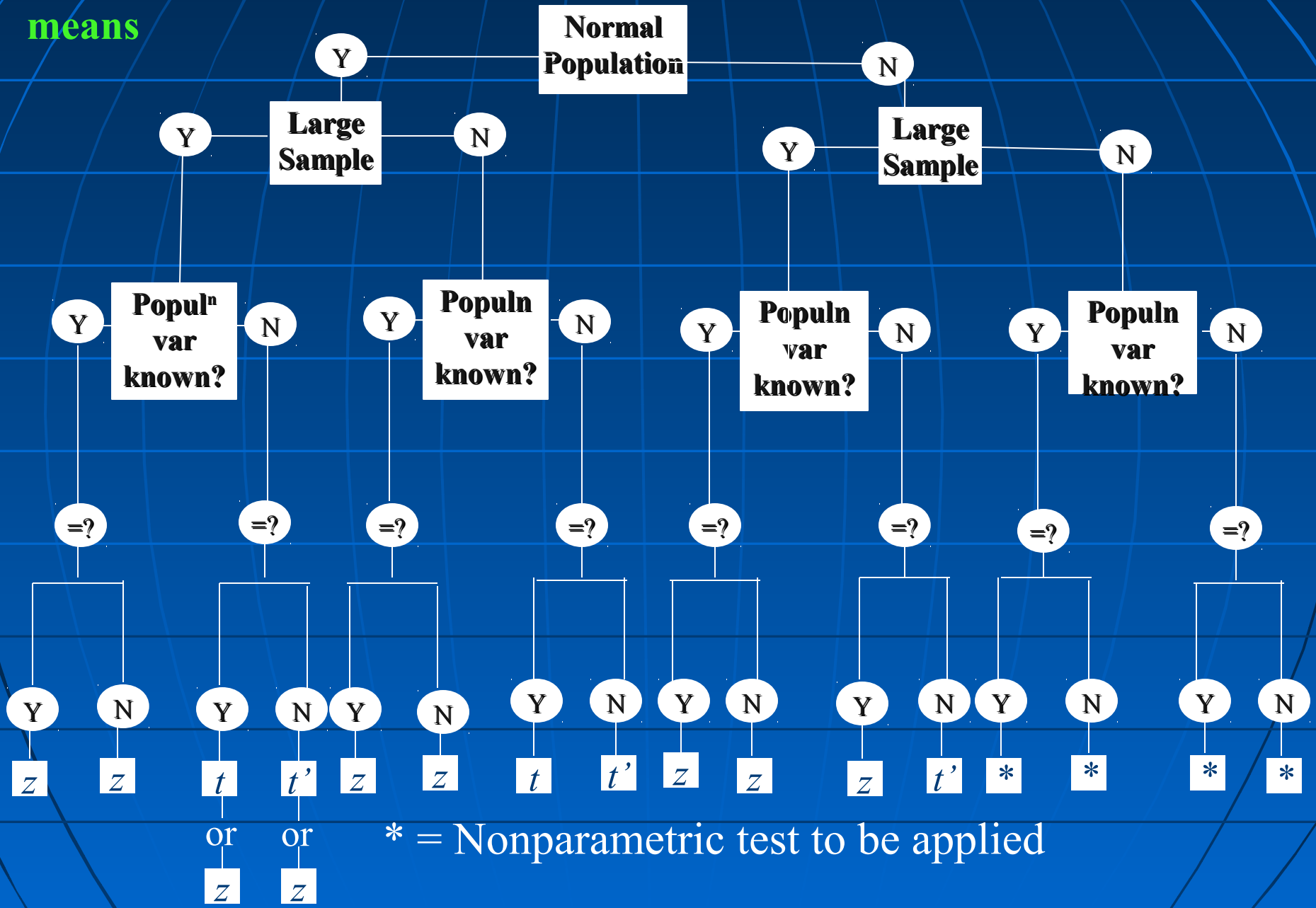
Sample	n	Mean	S.D
Exercise group	13	4.5	0.3
Sedentary group	17	3.7	1.0

We assume that the two populations of overall muscle condition scores are approximately normally distributed. We are unwilling to assume, however, that the two population variances are equal. We wish to construct a 95% confidence interval for the difference between the mean overall muscle condition scores of the two populations represented by the samples.

Solution:

The 95% confidence interval for the difference between the two population means when they are not equal is 0.25, 1.34

Flowchart in deciding whether the reliability factor should be z, t, or t' when making inferences about the difference between two population means



* = Nonparametric test to be applied

Confidence interval for a population proportion

- Population proportion in many situations is a matter of great interest to the health professional.
- What proportion of people suffers from different diseases, disability etc.? what proportion of patience response to a particular treatment? These are certain important questions for the health purpose.
- We estimate the population proportion as in the case of population mean.
- A random sample is drawn from the population of interest and the sample proportion $\hat{p}$ provides an unbiased estimate of the population proportion P.
- A confidence interval is obtained by the general formula estimator ± (reliability coefficient) × (standard error)
- when both np and n(1-p) are greater than 5, the sampling distribution of $\hat{p}$ is quite close to the normal distribution.

- When above condition is mate100(1-α) % confidence interval for p is given by

$$\hat{p} \pm z_{(1-\alpha/2)} \sqrt{\hat{p}(1-\hat{p})/n}$$

Example:

Mathers et al.(A-12) found that in a sample of 591 patients admitted to a psychiatric hospital, 204 admitted to using cannabis at least once in their lifetime. We wish to construct a 95% confidence interval for the proportion of lifetime cannabis users in the sampled population of psychiatric hospital admission.

Solution:

The 95% confident that the population proportion p is between 0.3069, 0.3835

Confidence interval for the difference between two population proportion

- An unbiased point estimator of the difference between two population proportions (p_1-p_2) is provided by the difference between sample proportion $\hat{p}_1 - \hat{p}_2$.
- When n1 and n2(sample size) are large and the population proportions are not too close to 0 and 1, normal distribution theory may be applied to obtain confidence interval.
- A 100(1-α)% confidence interval for p_1-p_2 is given by

$$\hat{p}_2 - \hat{p}_2 \pm z_{(1-\alpha/2)} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Example:

Borst et al. (A-16) investigated the relation of ego development, age, gender and diagnosis to suicidality among adolescent psychiatric inpatients. There sample consisted of 96 boys and 123 girls between the age of 12 and 16 selected from admissions to a child and adolescent unit of a private psychiatric hospital. Suicide attempts were reported by 18 of the boys and 60 of the girls. Let us assume the girls and the boys behave like a simple random sample from a population of similar girls and that the boys likewise may be considered a SRS from a population of similar boys. For these two population, we wish to construct a 99% confidence interval for the difference between between the proportions of suicide attempters.

Solution:

The 99% confident that for the sampled populations , the proportion of suicide attempts among girls exceeds the proportion of suicide attempts among boys by somewhere between 0.1450, 0.4556

Example:

Goldberg et al. (A-24) conducted a study to determine if an acute dose of dextroamphetamine might have positive effects on affect and cognition in schizophrenic patients maintained on a regimen of haloperidol. Among the variables measured was the change in patients' tension-anxiety states. For $n_2=4$ patients who responded to amphetamine, the standard deviation was 5.8. Let us assume that these patients constitute independent SRS from populations of similar patients. Let us also assume that change scores in tension-anxiety state is a normally distributed variable in both populations. We wish to construct a 95% confidence interval for the ratio of the variances of these two populations.

Solution:

$$0.2081 < \frac{\sigma_1^2}{\sigma_2^2} < 14.0554$$