3. Estimation and Testing

3.1 Estimation

Estimation is the process of finding approximate values of the population parameters. This consists of two methods, point estimation and interval estimation.

3.1.1 Point Estimation

A point estimate of a parameter θ is a single number that can be regarded as a sensible value for θ. A point estimate is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the point estimator of θ.

Example

Suppose, for example, that the parameter of interest is μ, the true average lifetime of batteries of a certain type. A random sample of n = 3 batteries might yield observed lifetimes (hours)

$x_1 = 5:0$; $x_2 = 6:4$; $x_3 = 5:9$.

The computed value of the sample mean lifetime is $\bar{x} = 5:77$.

It is reasonable to regard 5.77 as a very plausible value of μ our 'best guess' for the value of μ based on the available sample information.

3.1.2. Interval Estimation

A point estimator cannot be expected to provide the exact value of the population parameter. An interval estimate can be computed by adding and subtracting a margin of error to the point estimate.

Point Estimate +/- Margin of Error

The purpose of an interval estimate is to provide information about how close the point estimate is to the value of the parameter.

In the process of calculating the confidence interval for means, the information available on population variance should also be considered. The margin of error must be computed using either:

- the population standard deviation σ, or
- the sample standard deviation *s.*

Interval Estimate of Population Mean with Known Variance σ

Let us denote the 100(1 −α⁄2) percentile of the standard normal distribution as $z_{\alpha/2}$. For a random sample of sufficiently large size, the end points of the interval estimate at (1 − α) confidence level is given as follows:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Example :

Assume the population standard deviation σ of the student height in survey is 9.48. Find the margin of error and interval estimate at 95% confidence level. (Consider the data omitting the missing values)

```
> library(MASS)
> attach(survey)
> height<- na.omit(Height)
> MOE<- qnorm(.975)*9.48/sqrt(length(height))
> MOE
[1] 1.285237
> CI<-mean(height)+c(-MOE,MOE)
> CI
[1] 171.0956 173.6661
```

Assuming the population standard deviation σ being 9.48, the margin of error for the student height survey at 95% confidence level is 1.2852 centimeters. The confidence interval is between 171.10 and 173.67 centimeters.

Interval Estimate of Population Mean with Unknown Variance

Let us denote the 100(1 −α⁄2) percentile of the Student t distribution with n− 1 degrees of freedom as $t_{\alpha/2}$. For random samples of sufficiently large size, and with standard deviation s, the end points of the interval estimate at (1 −α) confidence level is given as follows:

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}}$$

Example:

Without assuming the population standard deviation of the student height in survey, find the margin of error and interval estimate at 95% confidence level.

```
> MOE<- qt(.975,length(height)-1)*sd(height)/sqrt(length(height))
> MOE
[1] 1.342878
> CI<-mean(height)+c(-MOE,MOE)
> CI
[1] 171.0380 173.7237
```

Alternative Way:

```
> t.test(height)

        One Sample t-test

data:  height
t = 253.0667, df = 208, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 171.0380 173.7237
sample estimates:
mean of x
 172.3809
```

Interval Estimate of Population Proportion

Let us denote the 100(1 −α⁄2) percentile of the standard normal distribution as $z_{\alpha/2}$. If the samples size n and population proportion p satisfy the condition that np ≥ 5 and n(1 − p) ≥ 5, then the end points of the interval estimate at (1 − α) confidence level is defined in terms of the sample proportion as follows.

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

Example:

Compute the margin of error and estimate interval for the female students' proportion in survey at 95% confidence level.

```
> gender<- na.omit(survey$Sex)
> k <- sum(gender == "Female")
> pbar <- k/length(gender)
```

```
> pbar
[1] 0.5
> MOE <- qnorm(.975)*sqrt(pbar*(1-pbar)/length(gender))
> MOE
[1] 0.06379139
> CI<-pbar+c(-MOE,MOE)
> CI
[1] 0.4362086 0.5637914
```

At 95% confidence level, between 43.6% and 56.3% of the university students are female, and the margin of error is 6.4%.

Alternative Way:

```
> prop.test(k, length(gender))

        1-sample proportions test without continuity correction

data:  k out of length(gender), null probability 0.5
X-squared = 0, df = 1, p-value = 1
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.4367215 0.5632785
sample estimates:
  p
0.5
```

Estimating the Difference between Two Population Means

Let μ_1 equal the mean of population 1 and μ_2 equal the mean of population 2. The difference between the two population means is μ_1 - μ_2. To estimate μ_1 - μ_2, we will select a simple random sample of size n_1 from population 1 and a simple random sample of size n_2 from population 2. Let $\bar{x}_1$ equal the mean of sample 1 and $\bar{x}_2$ equal the mean of sample 2. Then the point estimator of the difference between the means of the populations 1 and 2 is, $\bar{x}_1 - \bar{x}_2$.

Let σ_1 and σ_2 equal the standard deviation of of population 1 and population 2 respectively.

Interval Estimate of the Difference between Two Independent Population Means : σ_1 and σ_2 Known

Let us denote the 100(1 −α⁄2) percentile of the standard normal distribution as $z_{\alpha/2}$. For a random sample of sufficiently large size, the end points of the interval estimate at (1 − α) confidence level is given as follows:

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

Interval Estimate of the Difference between Two Independent Population Means : σ_1 and σ_2 unknown

When σ_1 and σ_2 are unknown, we will use the sample standard deviations s_1 and s_2 as estimates of σ_1 and σ_2 , and replace $z_{\alpha/2}$ with $t_{\alpha/2}$. Then the end points of the interval estimate at (1 −α) confidence level is given as follows:

$$\bar{x}_1 - \bar{x}_2 \pm t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Here the degree of freedom to be used for t value is,

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

Interval Estimate of the Difference between Two Independent Population Means : σ_1 and σ_2 Unknown but Equal

The end points of the interval estimate at (1 −α) confidence level in this situation is given as follows:

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

Where s_p is called the pooled sample variance and can be calculated using the formula :

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

Here the degree of freedom for the t value equals, $n_1 - n_2 - 2$.

Estimating the Difference between Two Population Means: Matched Samples

With a matched-sample design each sampled item provides a pair of data values. This design often leads to a smaller sampling error than the independent-sample design because variation between sampled items is eliminated as a source of sampling error.

Calculation of the confidence interval is similar to the single population σ unknown case. Only difference is, the sample mean and the sample standard deviation are calculated using a new sample constructed by taking the differences of each pair of data values.

Interval Estimate of the Difference between Two Population Proportions

If the sample sizes are large, the sampling distribution of $\bar{p}_1 - \bar{p}_2$ can be approximated by a normal probability distribution. The sample sizes are sufficiently large if <u>all</u> of these conditions are met: $n_1p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2p_2 \geq 5$, $n_2(1 - p_2) \geq 5$. Then the end points of the interval estimate at $(1 - \alpha)$ confidence level is defined in terms of the sample proportion as follows.

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