1 One and two sample estimation problems

The distributions associated with populations are often known except for one or more parameters. Estimation problems deal with how best to estimate the value of these parameters and equally important, how to provide a measure of the confidence in that estimate. In other words, what can one say about the range of certainty in that estimate. We shall be concerned with the estimation of the following parameters: the mean, the variance and the proportion.

Notation1 We shall denote parameters by Greek letters such as θ. The statistic which serves to estimate θ is indicated by the capital $\hat{\Theta}$ and it is a random variable. A specific value of that random variable is indicated by $\hat{\theta}$. An estimator is a rule which produces a point estimate based on the observed data sample $X_1, ..., X_n$.

Notation2 The population mean will be denoted by μ, the population variance by σ^2, and the population proportion by p.

Notation3 There is some variation to the above notation. The sample mean based on a sample of size n is denoted by $\bar{X}_n$ and a specific observed value is denoted by $\bar{x}_n$. The sample variance based on a sample of size n is denoted by S_n^2 or simply S^2 if the meaning is clear. a specific observed value is denoted by s^2.

Notation4 The population proportion is denoted by p whereas the point estimator is denoted by $\hat{P}$. A specific observed value is denoted by $\hat{p}$.

Definition A statistic $\hat{\Theta}$ is said to an unbiased estimator of a parameter θ if

$$E\left(\hat{\Theta}\right) = \theta$$

The estimator with the smallest variance among all unbiased estimators is called the most efficient estimator of θ.

Theorem The sample mean $\bar{X}_n$ is an unbiased estimator for the population mean μ.

Theorem The sample variance S^2 is an unbiased estimator for the population variance σ^2.

1.1 Variance of a point estimator

Sometimes we have more than point estimator for the same parameter. In order to make a choice we impose certain restrictions. One is that the estimator should be unbiased. The other is the estimator should have a smaller variance. Among all unbiased estimators, the one with the smallest variance is called the most efficient estimator. As an example, for a sample of size 3, both $\bar{X}_3$ and $\hat{\Theta} = \frac{X_1+2X_2+3X_3}{6}$ are unbiased. But $\bar{X}_3$ is more efficient since

$$Var\left(\bar{X}_3\right) = \frac{\sigma^2}{3}, Var\left(\hat{\Theta}\right) = \frac{14}{36}\sigma^2$$

1.2 Interval estimates

In addition to point estimators for parameters, we are also interested in interval estimates which provide some measure of uncertainty. In order to construct interval estimates, we need to find two random variables, $\hat{\Theta}_L, \hat{\Theta}_U$ such that

$$P\left(\hat{\Theta}_L < \theta < \hat{\Theta}_U\right) = 1 - \alpha, 0 < \alpha < 1.$$

In other words, we would like the parameter to be contained in an interval. The interval computed from the sample, $\hat{\theta}_L < \theta < \hat{\theta}_U$ is called a $100\,(1-\alpha)\,\%$ confidence interval. The endpoints of the interval are called the confidence limits. In general, we would like a short confidence interval with a high degree of confidence.

1.3 Estimating the mean of a single population when the variance is known

Assume that either the sample comes from a normal distribution or the sample size n is very large. In either case, if $\bar{X}_n$ is used as an estimator of μ, then with probability $(1-\alpha)$, the mean μ will be included in the interval $\bar{X}_n \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$. Hence,

$$\hat{\Theta}_L = \bar{X}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\Theta}_U = \bar{X}_n + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The width of the confidence interval is

$$\hat{\Theta}_U - \hat{\Theta}_L = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

The error defined as the half width, will not exceed e when the sample size is

$$n = \left(\frac{z_{\alpha/2}\sigma}{e}\right)^2$$

We note that the width of the confidence interval decreases as the sample size increases. It increases as the standard deviation increases and as the the confidence level increases.

Defintion The standard error of an estimate is the standard deviation of the estimate.

Hence, the standard error of $\bar{X}$ deboted s.e.$\left(\bar{X}\right)=\frac{\sigma}{\sqrt{n}}$

Example The amount of butterfat in pounds produced by a cow during a 305 day milk production period is normally distributed with mean μ and standard deviation 89.75. Construct a 95% confidence for the mean μ on the basis of a random szmple of size 20 when the sample mean is 507.5.

Here, $\bar{x}_{20} = 507.5, \sigma = 89.75, \alpha = 0.05$. Hence $z_{0.025} = 1.96$. The interval becomes

$$507.5 \pm 1.96\frac{89.75}{\sqrt{20}} \quad = \quad 507.5 \pm 39.34$$

It is felt that the sample size is too large. If the level of confidence is decreased to 90%, and the margin of error increased to $e = 45$, what would be the sample size necessary?

$$n = \left(\frac{1.645\,(89.75)}{45} \right)^2 = 10.76 \cong 11$$

1.4 Estimating the mean of a single population when the variance is unknown

When the variance of a normal population is unknown, we may construct a confidence interval based on the distribution of the random variable

$$T = \frac{\sqrt{n}\,(\bar{X}_n - \mu)}{S}$$

which is Student t with $n-1$ degrees of freedom. In that case, we have that with probability $(1-\alpha)$, the mean μ will be included in the interval $\bar{X}_n \pm t_{\alpha/2}\frac{S}{\sqrt{n}}$. Hence,

$$\hat{\Theta}_L = \bar{X}_n - t_{\alpha/2}\frac{S}{\sqrt{n}}, \hat{\Theta}_U = \bar{X}_n + t_{\alpha/2}\frac{S}{\sqrt{n}}$$

Example The amount of butterfat in pounds produced by a cow during a 305 day milk production period is normally distributed with mean μ. On the basis of a random szmple of size 20, the sample mean is 507.5. and the sample standard deviation is 89.75. Construct a 95% confidence for the mean μ.

Here, $\bar{x}_{20} = 507.5, s = 89.75, \alpha = 0.05$. Hence $t_{0.025} = 2.093$ when the number of degrees of freedom is $n - 1 = 19$. The interval becomes

$$507.5 \pm 2.093\frac{89.75}{\sqrt{20}} \quad = \quad 507.5 \pm 42.00$$

We see that the interval is wider when the variance is unknown. We refer to $\frac{\sigma}{\sqrt{n}}$ as the standard error and to $\frac{s}{\sqrt{n}}$ as the estimate of the standard error.

1.5 Estimating the difference between two means

Example Suppose that we have two different feeding programs for fattening beef cattle. We are interested in seeing if there is a significant difference between them.

Suppose that we have two normal populations with means μ_1, μ_2 and variances σ_1^2,σ_2^2 respectively. We would like a confidence interval for the difference $\mu_1 - \mu_2$. Several cases present themselves which we can tabulate as follows. Let $\bar{X}_1, \bar{X}_2$ represent the sample means form the two populations based on samples $n_1 n_2$respectively.

Variances	Interval	Definitions	d.f.
σ_1^2, σ_2^2 known	$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$		
$\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown	$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2} S_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$S_P^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$	$n_1 + n_2 - 2$
σ_1^2, σ_2^2 unknown	$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$		*

$$* = \frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\left[\left(\frac{S_1^2}{n_1}\right)^2/(n_1-1)+\left(\frac{S_2^2}{n_2}\right)^2/(n_2-1)\right]}$$

1.6 Paired observations

Example An experiment is conducted to compare people's reaction time to a red light vs a green light in seconds.

	Red (x)	Green (y)	$d = x - y$
1	0.30	0.43	-0.13
2	0.23	0.32	-0.09
3	0.41	0.58	-0.17
4	0.53	0.46	0.07
5	0.24	0.27	-0.03
6	0.36	0.41	-0.05
7	0.38	0.38	0
8	0.51	0.61	-0.10

In the case of paired samples, we can no longer assume that the $X's$ and $Y's$ are independent. Instead we may assume that the differences $D_1, ..., D_n$ constitute a random sample from say a normal distribution with mean μ_D and variance σ_D^2. In that case we are back to the one sample problem. In our example, a confidence interval for the mean is given by

$$\bar{D} \pm t_{\alpha/2}\frac{S_d}{\sqrt{n}}$$

Here, we have

$$-0.0625 \pm 2.365 \frac{0.0765}{\sqrt{8}}$$

or

$$(-0.1265, 0.0015)$$

We see that 0 is in the interval so that we conclude there is no difference between the two reaction times. However, the upper limit is very small. Had it been negative, we would have concluded that people react faster to a red light.

Exercise Show that if this example were treated as a two sample problem, the variance would increase, the d.f. would increase but the t-value would decrease.

1.7 Single sample: estimating a proportion

Suppose that X is a random variable having a binomial distribution with parameters n, p. We are interested in estimating p. Here X represents the number of successes in mn repetitions of a Bernoulli experiment. As examples, we may be interested in estimating the proportion of voters in favour of a certain candidate or perhaps the success rate of a certain drug in curing the common cold.The construction of a confidence interval is based on the Central Limit approximation.

If $\hat{P}$ is used as an estimator of p, then a $100\,(1-\alpha)\,\%$ confidence interval for p is given by $\hat{P} \pm z_{\alpha/2}\sqrt{\hat{P}\hat{Q}/n}$.

The error will not exceed e when the sample size is

$$\begin{aligned} n &= \left(\frac{z_{\alpha/2}}{e}\right)^2 \hat{p}\hat{q} \\ Max\ n &= \frac{1}{4}\left(\frac{z_{\alpha/2}}{e}\right)^2 \end{aligned}$$

Example From a random sample of 400 citizens in Ottawa, there were 136 who indicated that the city's transpotation system is adequate. Construct a 99% confidence interval for the population proportion who feel the transportation system is adequate.

Here, $\hat{p} = \frac{136}{400} = 0.34$,$\sqrt{\hat{p}\hat{q}}$=0.2244,$z_{\alpha/2} = 2.575$. The confidence interval is then

$$0.34 \pm 2.575\sqrt{\frac{0.34\,(0.66)}{400}}$$

$$0.34 \pm 0.061$$

Example What is the sample size necessary to estimate the proportion of voters in favor of a candidate if we allow a maximum error of 3% and would like a confidence of 95%?

$$\begin{aligned} n &= max_p p\,(1-p)\left(\frac{z_{\alpha/2}}{e}\right)^2 \\ &= \frac{1}{4}\left(\frac{1.96}{0.03}\right)^2 \cong 1068 \end{aligned}$$

1.8 Estimating the difference between two proportions

For two proportions P_1, P_2 a point estimate of the difference $P_1 - P_2$ is given by

$$\hat{P}_1 - \hat{P}_2$$

A $100\,(1-\alpha)\,\%$ confidence interval for $p_1 - p_2$ is given by

$$\hat{P}_1 - \hat{P}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_1\hat{Q}_1}{n_1} + \frac{\hat{P}_2\hat{Q}_2}{n_2}}$$

Example Two detergents were tested for their ability to remove grease stains. Type A was successful on 63 stains out of 91 whereas type B was successful on 42 stains out of 79. We construct a 90% confidence interval for the difference Type A-Type B and obtain

$$(0.038, 0.282)$$

Since the interval does not include 0, it seems that

$$P_A - P_B > 0$$

and hence Type A is better than Type B.

1.9 Estimating the variance

Suppose that we are given a random sample of size n from a normal population with variance σ^2. Estimation of the variance is based on the statistic

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

We obtain from the chi square distribution two values $\chi^2_{n-1}(1-\alpha/2), \chi^2_{n-1}((\alpha/2))$ such that

$$P\left(\chi^2_{n-1}(1-\alpha/2) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{n-1}((\alpha/2))\right) = 1-\alpha$$

Reworking the inner expression we see that a $100(1-\alpha)\%$confidence interval for σ^2 is given by

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1}((\alpha/2))}, \frac{(n-1)S^2}{\chi^2_{n-1}(1-\alpha/2)}\right)$$

Example The time required in days based on a sample of size 13 for a maturation of seeds for a certain type of plant is 18.97 days with a sample variance of 10.7.

A 90% confidence interval for the population variance is

$$\left(\frac{128.41}{21.03}, \frac{128.41}{5.226}\right) = (6.11, 24.57)$$

2 One and two-sample tests of hypotheses

A statistical hypothesis is a conjecture concerning one or more populations. For example, we may believe that a coin is balanced or that the average height of soldiers in the Canadian army is 5'9". Statistical methods can be used to verify such claims.

There are two possible actions that one may take with respect to a claim or hypothesis:

i) rejection of the null hypothesis when it is true is a Type I error

ii) non rejection of the null hypothesis when it is false is a Type II error

	H_0 is true	H_0 is false
Do not reject H_0	Correct decision	Type II error
Reject H_0	Type I error	Correct decision

We may illustrate the situation using the normal distribution for the sample mean. Suppose that under the null hypothesis

$$\bar{X}_n \sim N\left(25, 4^2\right)$$

whereas under the alternative

$$\bar{X}_n \sim N\left(50, 5^2\right)$$

We illustrate graphically the various types of error as well as the notion of p-values and power.

Note that Table 6.3 on page 264 of the text summarizes all the tests con-

cerning means from normal distributions, i.e.

1. single mean, variance known
2. single mean,variance unknown
3. difference of two means, variances known
4. difference of two means, varainces equal to a common but unknown mean
5. difference of two means, variances unknown
6. paired observations