Estimation Theory

Statistical Inference – methods by which generalizations are made about a population

Two Major Areas of Statistical Inference

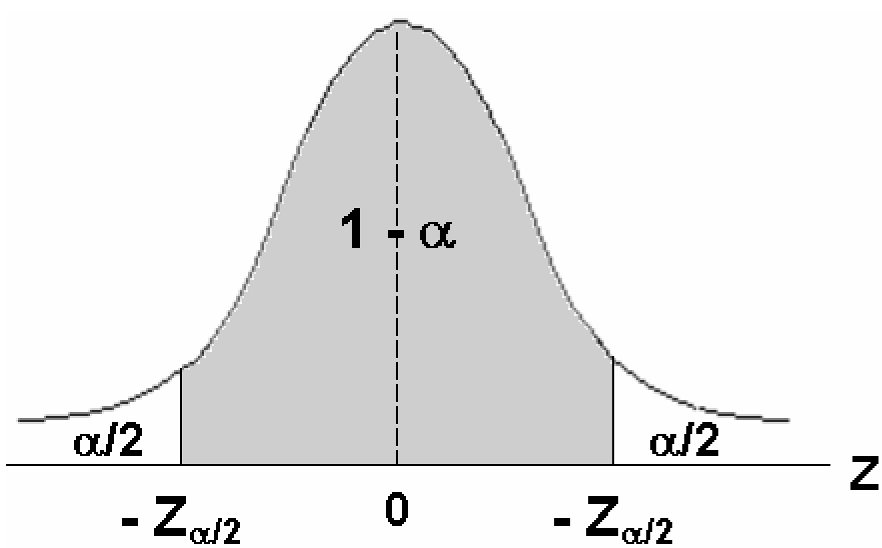
1. **Estimation** – a parameter is established based on the sampling distribution of a proportion, establishing a certain degree of accuracy from the estimate
2. **Test of Hypothesis** – a decision is arrived at about a prestated hypothesis, thereby accepting or rejecting the hypothesis.

I. Estimation of Means: Single Sample

Point estimate – a statistic taken from a sample and is used to estimate a population parameter. However, a point estimate is only as good as the representativeness of its sample. If other samples are taken from the population, the point estimates derived are likely to vary.

Interval estimate – a range of values within which the analyst can declare with some confidence the population parameter lies; also called confidence interval.

If our sample is selected from a normal population, or if the sample size n is sufficiently large, we can establish a confidence interval for μ by considering the sampling distribution of $\overline{x}$.



From the central limit theorem, we can expect the sampling distribution of $\overline{x}$ to be approximately normal with $\mu_{\overline{x}} = \mu$ and $\delta_{\overline{x}} = \dfrac{\delta}{\sqrt{n}}$.

Writing $z_{\alpha/2}$ for the z-value above which we find an area of $\alpha/2$, we see that:

$$P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$$

where: $z = \dfrac{\overline{x} - \mu}{\delta/\sqrt{n}}$, hence: $P(-z_{\alpha/2} < \dfrac{\overline{x} - \mu}{\delta/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$

Multiply each term of the inequality $\delta/\sqrt{n}$ and subtracting $\overline{x}$ from each term and multiplying by -1, we formally state the results as:

If $\overline{x}$ is the mean of a random sample of size *n* from a population with known variance δ^2, a (1-α)100% confidence interval for μ is given by

$$\boxed{\overline{x} - z_{\alpha/2}\frac{\delta}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\frac{\delta}{\sqrt{n}}}$$

confidence interval of μ;

δ known; large samples

Problems:

1. An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed with a standard deviation of 40 hours. If a sample of 30 bulbs has an average life of 780 hours, find a 96% confidence interval for the population mean of all bulbs produced by this firm. Ans: 765 < μ < 795
2. A random sample of 100 automobile owners shows that, in the state of Virginia, an automobile is driven on the average 23500 kilometers per year with a standard deviation of 3900 kilometers. Construct a 99% confidence interval for the average number of kilometers an automobile is driven annually in Virginia. Ans: 22496 < μ < 24504
3. The heights of a random sample of 50 college students showed a mean height of 174.5 cm and a standard deviation of 6.9 cm. Construct a 98% confidence interval for the mean height of all college students. Ans: 172.23 < μ < 176.77

Making use of the fact that for large samples from infinite populations, the sampling distribution is approximately a normal distribution with $\mu_{\overline{X}} = \mu$ and $\delta_{\overline{X}} = \frac{\delta}{\sqrt{n}}$, we find that the probability is 1 - α that the mean of a large sample from an infinite population will differ from the population mean by at most $z_{\alpha/2} \frac{\delta}{\sqrt{n}}$. In other words,

If $\overline{x}$ is used as an estimate of μ, we can be (1-α)100% confident that the error will not exceed $\boxed{z_{\alpha/2} \frac{\delta}{\sqrt{n}}}$ **Maximum error of estimate**

Problems:

4. With reference to problem 3, what can we assert with 98% confidence about the possible size of our error if we estimate the mean height of all college students to be 174.5 cm? Ans: error $\leq$ 2.27 cm
5. A team of efficiency experts intends to use the mean of a random sample of size 150 to estimate the average mechanical aptitude of assembly-line workers in a large industry. If, based on experience, the efficiency experts can assume that δ= 6.2 for such data, what can they assert with probability 0.99 about the maximum error of their estimate? Ans: max. error = 1.30

From, the preceding formula, we say:

If $\overline{x}$ is used as an estimate of μ, we can be (1-α)100% confident that the error will not exceed a specified amount ***e*** when the sample size is

$$\boxed{n = (\frac{z_{\alpha/2}\delta}{e})^2} \quad \textbf{Sample size for estimating } \mu$$

Problems:

6. With reference to problem 1, how large a sample is needed if we wish to be 96% confident that our sample mean will be within 10 hours of the true mean? Ans: 68
7. The dean of a certain college wants to use the mean of a random sample to estimate the average amount of time students take to get from one class to the next, and she wants to be able to assert with probability 0.95 that her error will be at most 0.25 minute. If she knows from studies of a similar kind that it is reasonable to let δ= 1.50 minutes, how large a sample will she need? Ans: 139

II. Estimation of Means (Small Samples)

To develop corresponding methods that also apply to small samples, it is necessary to assume that the populations we are sampling have roughly the shape of normal distributions. We then base our methods on the statistic

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

which is a value of a random variable having the t-distribution. This distribution is sometimes called the student t-distribution or student's t distribution; developed by the statistician W.S. Gosset, under the pen name "Student."

The exact shape of the t-distribution depends on a parameter called the number of *degrees of freedom*, equals $n - 1$, sample size less one.

Small-sample confidence interval for μ

$$\boxed{\bar{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}}$$

Problems:

8. A machine is producing metal pieces that are cylindrical in shape. A sample of pieces is taken and the diameters are 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01 and 1.03 centimeters. Find a 99% confidence interval for the mean diameter of pieces from this machine, assuming an approximate normal distribution. Ans: $0.978 < \mu < 1.033$
9. A random sample of 12 graduates of a certain secretarial school typed an average of 79.3 wpm with a standard deviation of 7.8 wpm. Assuming normal distribution for the number of words typed per minute, find a 95% confidence interval for the average number of words typed by all graduates of this school. Ans: $74.34 < \mu < 84.26$

III. **Estimating the Difference Between Two Means: Two Samples**

A. Confidence interval for $\mu_1 - \mu_2$; $\delta_1^{\,2}$ and $\delta_2^{\,2}$ known

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\delta_1^{\,2}}{n_1} + \frac{\delta_2^{\,2}}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\delta_1^{\,2}}{n_1} + \frac{\delta_2^{\,2}}{n_2}}$$

B. Confidence interval for $\mu_1 - \mu_2$; $\delta_1^{\,2} = \delta_2^{\,2}$ but unknown (also for equal samples)

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} * s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where: s_p = pooled estimate of the population standard deviation

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

$$\upsilon = n_1 + n_2 - 2 \quad \text{degrees of freedom}$$

C. Confidence interval for $\mu_1 - \mu_2$; $\delta_1^{\,2} \neq \delta_2^{\,2}$ and unknown

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2}\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2}\sqrt{\frac{s_1^{\,2}}{n_1} + \frac{s_2^{\,2}}{n_2}}$$

with $$\upsilon = \frac{\left(\dfrac{s_1^{\,2}}{n_1} + \dfrac{s_2^{\,2}}{n_2}\right)^2}{\left[\dfrac{\left(s_1^{\,2}/n_1\right)^2}{n_1 - 1} + \dfrac{\left(s_2^{\,2}/n_2\right)^2}{n_2 - 1}\right]}$$ degrees of freedom

D. Confidence interval for $\mu_D = \mu_1 - \mu_2$ (paired observations)

$$\bar{d} - t_{\alpha/2} * \frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2} * \frac{s_d}{\sqrt{n}}$$

where: $\bar{d}$ = mean of normally distributed differences of *n* random pairs of measurement

s_d = standard deviation of the normally distributed differences on *n* random pairs of measurement

υ = n – 1 degrees of freedom

Problems:

10. Two kinds of thread are being compared for strength. 50 pieces of each type of thread are tested under similar conditions. Brand A has an average tensile strength of 78.3 kg with a standard deviation of 5.6 kg., while Brand B has an average tensile strength of 87.2 kg with a standard deviation of 6.3 kg. Construct a 95% confidence interval for the difference of the population means. Ans: $6.56 < \mu_B - \mu_A < 11.24$
11. In a batch chemical process, two catalysts are being compared for their effect on the output of the process reaction. A sample of 12 batches was prepared using catalyst 1 and a sample of 10 batches was obtained using catalyst 2. The 12 batches for which catalyst 1 was used gave an average yield of 85 with a sample standard deviation of 4, and the average for the second sample gave 81 and a sample standard deviation of 5. Find a 90% confidence interval for the difference between the population means, assuming that the populations are approximately normally distributed with equal variances. Ans: $0.69 < \mu_1 - \mu_2 < 7.31$
12. A taxi company is trying to decide whether to purchase brand A or brand B tires for its fleet of taxis. To estimate the difference in the two brands, an experiment is conducted using 12 of each brand. The tires are run until they wear out. The results are:

Brand A: $\bar{x}_1 = 36{,}300$ km Brand B: $\bar{x}_2 = 38{,}100$ km

$S_1 = 5000$ km $S_2 = 6100$ km

Compute a 95% confidence interval for $\mu_1 - \mu_2$, assuming the populations to be approximately normally. You may not assume that the variances are equal. Ans: $-65.22 < \mu_1 - \mu_2 < 2922$

13. The following data represent the running times of films produced by two motion-picture companies:

Company	Time (minutes)						
I	103	94	110	87	98		
II	97	82	123	92	175	88	118

Compute a 90% confidence interval for the difference between the average running times of films produced by the two companies. Assume that the running time differences are approximately normally distributed with unequal variances. Ans: $-11.9 < \mu_{II} - \mu_I < 36.5$

14. Referring to problem 12, find a 99% confidence interval for $\mu_1 - \mu_2$, if a tire of each company is assigned at random to the rear wheels of 8 taxis and the following distance, in kilometers, recorded:

Taxi	Brand A	Brand B
1	34,300	36,700
2	45,500	46,800
3	36,700	37,700
4	32,000	31,100
5	48,400	47,800
6	32,800	36,400
7	38,100	38,900
8	30,100	31,500

Assume that the differences of the distances are approximately normally distributed. Ans: $-2939 < \mu_D < 689$

IV. **Estimation of Proportions**

A. Single sample

Large sample confidence interval for p

If $\hat{p}$ is the proportion of successes in a random sample of size *n*, and $\hat{q} = 1 - \hat{p}$, an approximate (1-α)100% confidence interval for the binomial parameter p is given by

$$\boxed{\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}}$$

where $z_{\alpha/2}$ is the value leaving an area of $\alpha/2$ to the right.

Theorems:

1. If $\hat{p}$ is used as an estimate of p, we can be (1-α)100% confident that the error will be less than a specified amount <u>*e*</u> when the sample size is approximately
$$n = \frac{{z_{\alpha/2}}^2 \hat{p}\hat{q}}{e^2}$$
2. If $\hat{p}$ is used as an estimate of p, we can at least be (1-α)100% confident that the error will be less than a specified amount <u>*e*</u> when the sample size is approximately
$$n = \frac{{z_{\alpha/2}}^2}{4e^2}$$

Problems:

15. Compute a 68% confidence interval for the proportion of defective items in a process when it is found that a sample size of 100 yields 8 defectives. How large a sample is needed if we wish to be 98% confident that our sample proportion will be within 0.05 of the true population defective? Ans: 0.017 < p < 0.143; 160

16. A geneticist is interested in the proportion of African males that have a certain minor blood disorder. In a random sample of 100 African males, 24 are found to be afflicted. (a) Compute a 99% confidence interval for the proportion of African males that have this blood disorder. (b) What can we assert with 99% confidence about the possible size of our error if we estimate the proportion of African males with this blood disorder to be 0.24? Ans: 0.130 < p < 0.35; error ≤ 0.11

17. A study is to be made to estimate the proportion of residents in a certain city and its suburbs who favor the construction of a nuclear power plant. How large a sample is needed if one wishes to be at least 95% confident that the estimate is within 0.04 of the true proportion of residents in this city and its suburb that favor the construction of the nuclear power plant? Ans: 601

B. Two samples: Estimating the difference between two proportions

Large-sample confidence interval for $p_1 - p_2$

If $\hat{p}_1$ and $\hat{p}_2$ are the proportion of successes in random samples of size n_1 and n_2, respectively, $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$, an approximate $(1-\alpha)$100% confidence interval for the difference of two binomial parameters $p_1 - p_2$ is given by

$$\boxed{(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}}$$

where $z_{\alpha/2}$ is the z-value leaving an area of $\alpha/2$ to the right.

Problems:

18. A certain geneticist is interested in the proportion of males and females in the population that have a certain minor blood disorder. In a random sample of 1000 males, 250 are found to be afflicted, whereas 275 of the 1000 females appear to have the disorder. Compute a 95% confidence interval for the difference between the proportion of males and females that have the blood disorder. Ans: -0.0136 < $p_F - p_M$ < 0.0636

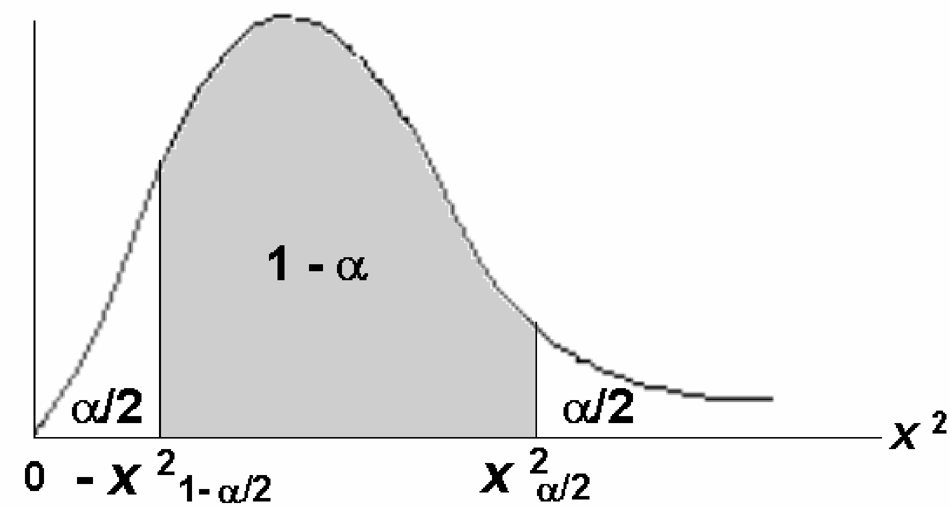
19. In a study, Germination and Emergence of broccoli, conducted by the Department of Horticulture Polytechnic Institute and State University, a researcher found that at 5°C, 10 seeds out of 20 germinated, while at 15°C, 15 out of 20 seeds germinated. Compute a 95% confidence interval for the difference between the proportion of germination at the two different temperatures and decide if there is a significant difference.

V. **Estimation of Variance**

To estimate δ based on s, we require that the population we are sampling has roughly the shape of a normal distribution. In that case, the statistic

$$\chi^2 = \frac{(n-1)s^2}{\delta^2}$$

called the chi-square statistic, is a value of a random variable having approximately the chi-square distribution. The parameter of this continuous distribution is called the number of degrees of freedom, just like the parameter of the t-distribution. Unlike the normal and the t-distributions, its domain consists only of non-negative real numbers.



χ^2 is the value such that the area under the curve to its right is $\alpha/2$, while $\chi_{1-\alpha/2}{}^2$ is such that the area under the curve to its left is $\alpha/2$.

$$P\,(\chi_{1-\alpha/2}{}^2 < \chi^2 < \chi_{\alpha/2}{}^2) = 1-\alpha$$

$$P(\chi_{1-\alpha/2}{}^2 < \frac{(n-1)s^2}{\delta^2} < \chi_{\alpha/2}{}^2) = 1-\alpha$$

Applying algebra, we get:

Small-sample confidence interval for δ^2

If s^2 is the variance of a random sample of size *n,* from a normal population, a $(1-\alpha)$ 100% confidence interval for δ^2 is

$$\boxed{\frac{(n-1)s^2}{\chi_{\alpha/2}^{\;2}} < \delta^2 < \frac{(n-1)s^2}{\chi_{1-\alpha/2}^{\;2}}}$$

where $\chi_{\alpha/2}^{\;2}$ and $\chi_{1-\alpha/2}^{\;2}$ are the χ^2-values with υ= n-1 degrees of freedom, leaving areas of $\alpha/2$ and 1-$\alpha/2$, respectively, to the right.

A $(1-\alpha)$ 100% confidence interval for δ is obtained by taking the square root of each endpoint of the interval for δ^2.

Problems:

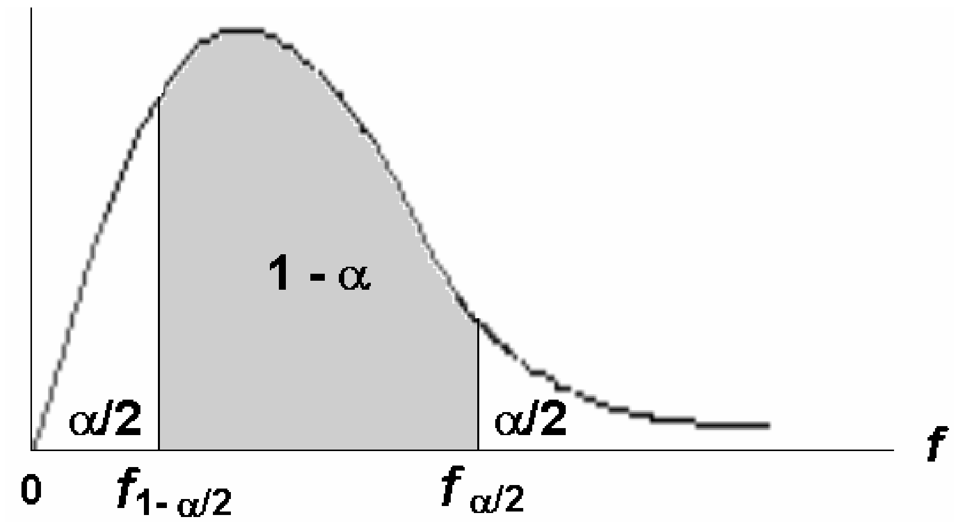
20. A random sample of size 20 obtained a mean of 72 and a variance 16 on a college placement test in mathematics. Assuming the scores to be normally distributed, construct a 98% confidence interval for δ^2. Ans: 8.4 < δ^2 < 39.827

21. A manufacturer of car batteries claims that his batteries will last, on average, 3 years with a variance of 1 year. If 5 of these batteries have lifetimes of 1.9, 2.4, 3.0, 3.5 and 4.2, construct a 95% confidence interval for δ^2 and decide the manufacturer's claim that δ^2 = 1 is valid. Assume the population of battery lives to be approximately distributed. Ans: 0.293 < δ^2 < 6.736

22. A random sample of 8 cigarettes of a certain brand has an average nicotine content of 2.6 mg and a standard deviation of 0.9 mg. Assuming the distribution of nicotine contents to be approximately normal, construct a 99% confidence interval for δ.

Estimating the ratio of two variances

Confidence interval for δ_1^2 / δ_2^2



If s_1^2 and s_2^2 are the variances of independent samples of size n_1 and n_2 respectively, from normal populations, then a $(1-\alpha)100\%$ confidence interval for δ_1^2 / δ_2^2 is

$$\frac{s_1^2}{s_2^2} * \frac{1}{f_{\alpha/2}(\upsilon_1, \upsilon_2)} < \frac{\delta_1^2}{\delta_2^2} < \frac{s_1^2}{s_2^2} * f_{\alpha/2}(\upsilon_2, \upsilon_1)$$

where $f_{\alpha/2}(\upsilon_1, \upsilon_2)$ is an f-value with $\upsilon_1 = n_1 - 1$ and $\upsilon_2 = n_2 - 1$ degrees of freedom leaving an area of $\alpha/2$ to the right and $f_{\alpha/2}(\upsilon_2, \upsilon_1)$ is a similar f-value with $\upsilon_2 = n_2 - 1$ and $\upsilon_1 = n_1 - 1$ degrees of freedom.

A $(1-\alpha)100\%$ confidence interval for δ_1 / δ_2 is obtained by taking the square root of each endpoint of the interval for δ_1^2 / δ_2^2

Problems:

23. Let us suppose that 12 Volkswagen trucks average 16 km per liter with a standard deviation of 1 km per liter and the 10 Toyota trucks average 11 km per liter with a standard deviation of 0.8 km per liter. Construct a 98% confidence interval for δ_1 / δ_2 where δ_1 and δ_2 respectively, are the standard deviations for the distances obtained per liter of fuel by the Volkswagen and Toyota mini-trucks.

24. Construct a 90% confidence interval for δ_1^2 / δ_2^2 in problem 12. Were we justified in assuming that $\delta_1 = \delta_2$ when we constructed our confidence interval for $\mu_1 - \mu_2$.