Chapter 6

Estimation based on Normal Distribution

Overview

- Point Estimation
- Parameter and statistic
- Unbiased estimator
- Interval estimation
- Confidence Interval for the Mean
- Sample size
- Confidence Intervals for the Difference between Two Means
- Confidence Interval for Variances and Ratio of Variances

6.1 Point Estimation of Mean and Variance

6.1.1 Introduction

- Assume that **some characteristics** of the elements in a population can be represented by a **random variable** X whose p.d.f. (or p.f.) is $f_X(x;\theta)$,

- where the form of the probability density function (or probability function) is assumed known except that it contains an unknown parameter θ.

6.1.1 Introduction (Continued)

- Further assume that the values $x_1, x_2, \cdots, x_n$ of a random sample $X_1, X_2 \cdots, X_n$ from $f_X(x; \theta)$ can be observed.

- On the basis of the observed sample values $x_1, x_2, \cdots, x_n$, it is desired to estimate the value of the unknown parameter θ.

6.1.2 Estimation

The estimation can be made in two ways: **Point estimation** and **Interval estimation**

- **Point estimation** is to let the value of some statistic, say
$$\hat{\theta} = \hat{\Theta}(X_1, X_2, \cdots, X_n),$$
to estimate the unknown parameter θ;

  such a statistic
$$\hat{\Theta}(X_1, X_2, \cdots, X_n),$$
is called **a point estimator**.

Statistic

- A statistic is a function of the random sample which does not depend on any unknown parameters.
- For example

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

or

$$X_{(n)} = \max(X_1, X_2, \cdots, X_n)$$

are some examples of a statistic.

Statistic (Continued)

- Let

$$W = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$$

Then W is not a statistic if μ is not known.

- However W is a statistic if μ is known.

Point Estimate of Mean

- Suppose μ is the population mean.
- The statistic that one uses to obtain a point estimate is called **an estimator**,

  For example, $\bar{X}$ is an estimator of μ.

  The value of $\bar{X}$, denoted by $\bar{x}$, is an estimate of μ.

Point Estimate of Mean (Continued)

Example

- If the sample mean of a random sample taken from a population with mean μ is 5,
- Then a point estimate for the population mean μ is 5.

Note: Different random samples give different point estimates of μ.

Interval Estimation

- Interval estimation is to define two statistics, say,

$$\hat{\Theta}_L \quad \text{and} \quad \hat{\Theta}_U, \qquad\qquad \text{where } \hat{\Theta}_L < \hat{\Theta}_U$$

so that $(\hat{\Theta}_L,\ \hat{\Theta}_U)$ constitutes an interval for which the probability of containing the unknown parameter θ can be determined.

Interval Estimation (Continued)

For example

- Suppose σ^2 is known. Let

$$\widehat{\Theta}_L = \bar{X} - 2\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \widehat{\Theta}_U = \bar{X} + 2\frac{\sigma}{\sqrt{n}}.$$

- Then

$$\left(\bar{X} - 2\frac{\sigma}{\sqrt{n}}, \bar{X} + 2\frac{\sigma}{\sqrt{n}}\right)$$

is an interval estimator for μ.

6.1.3 Unbiased Estimator

Definition 6.1 (Unbiased estimator)

- A statistic $\widehat{\Theta}$ is said to be an **unbiased estimator** of the parameter θ if

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

Unbiased Estimator (Continued)

Example 1

$\bar{X}$ is an unbiased estimator of μ. That is, $E(\bar{X}) = \mu$.

Example 2

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is an **unbiased** estimator of σ^2.

That is,

$$E(S^2) = \sigma^2$$

Unbiased Estimator (Continued)

Example 3

- $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is a **biased** estimator of σ^2.

- It can be shown that,

$$E(T) = \frac{n-1}{n}\sigma^2 \neq \sigma^2.$$

6.2 Interval Estimation

- An interval estimate of a population parameter θ is an interval of the form

$$\hat{\theta}_L < \theta < \hat{\theta}_U,$$

where $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on

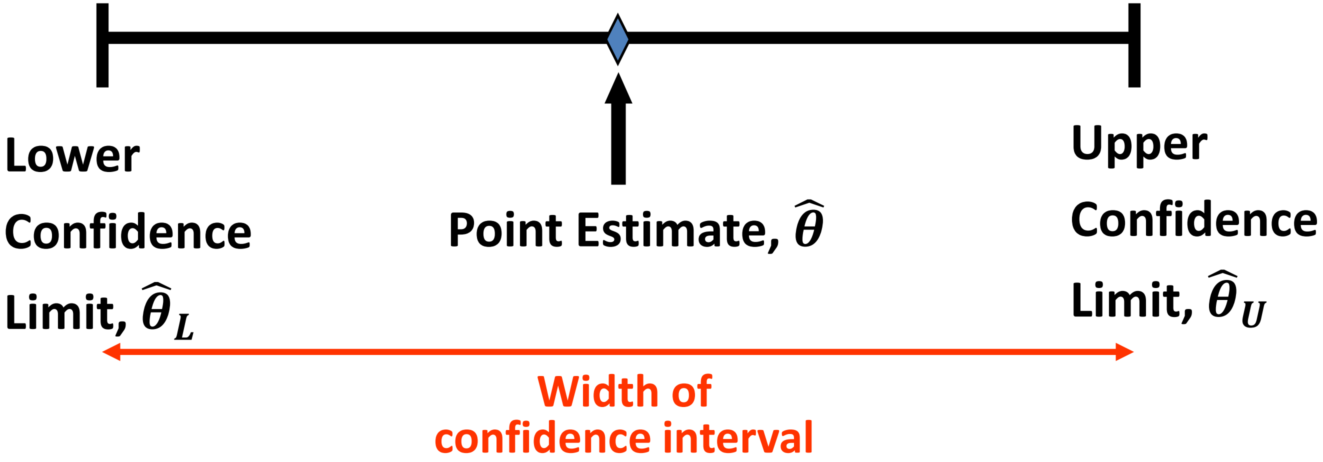
(1) the value of the statistic $\hat{\Theta}$ for a particular sample and

(2) the sampling distribution of $\hat{\Theta}$.

Interval Estimation (Continued)

- The interval estimate of a population parameter θ is an interval of the form

$$\hat{\theta}_L < \theta < \hat{\theta}_U,$$



Interval Estimation (Continued)

- Since different samples will generally yield different values of $\widehat{\Theta}$,
- therefore, different values of $\hat{\theta}_L$ and $\hat{\theta}_U$, these end points of the interval are values of corresponding random variables $\widehat{\Theta}_L$ and $\widehat{\Theta}_U$.
- These intervals may not contain the parameter θ as $\hat{\theta}_L$ and $\hat{\theta}_U$ vary.

Interval Estimation (Continued)

- We shall seek a random interval

$$(\hat{\Theta}_L, \hat{\Theta}_U)$$

containing θ with a given probability 1 – α.

- That is

$$\Pr(\hat{\Theta}_L < \theta < \hat{\Theta}_U) = 1 - \alpha.$$

Interval Estimation (Continued)

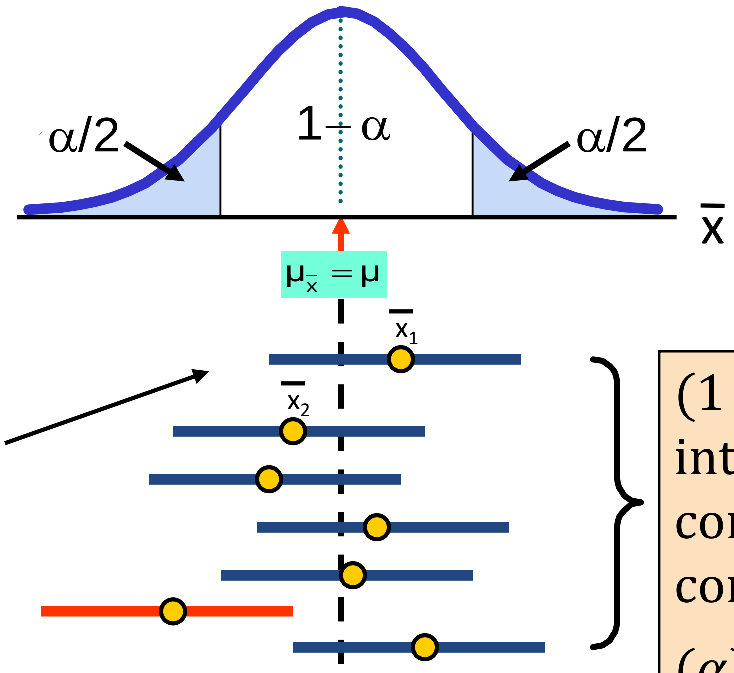
- Then the interval $\hat{\theta}_L < \theta < \hat{\theta}_U$, computed from the selected sample is called a $(1-\alpha)100\%$ confidence interval for θ,
- and the fraction $(1-\alpha)$ is called **confidence coefficient** or **degree of confidence**,
- and the end points $\hat{\theta}_L$ and $\hat{\theta}_U$ are called lower and upper confidence limits respectively.

Interval Estimation (Continued)

- This means that if samples of the same size n are taken,
- then in the long run, $(1-\alpha)100\%$ of the intervals will contain the unknown parameter θ,
- and hence with a confidence of $(1-\alpha)100\%$, we can say that the interval covers θ.

Intervals and Level of Confidence

Sampling Distribution of the Mean



Intervals extend from

$\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}$

to

$\bar{X} + z_{\alpha/2}\sigma/\sqrt{n}$

Confidence Intervals

$(1-\alpha)100\%$ of intervals constructed contain μ;

$(\alpha)100\%$ do not.

6.3 Confidence Intervals for the Mean

6.3.1 Known Variance Case

- Confidence interval for mean with
  (i) known variance and
  (ii) the population is normal
  or n is sufficiently large (say $n \geq 30$)

Confidence Intervals for the Mean (Continued)

Known Variance Case (Continued)

- When the population is normal or by the Central Limit Theorem, we can expect that

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Therefore

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

C.I. for Mean with Known Variance

- Hence

$$\Pr\left(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

or

$$\Pr\left(\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right) = 1-\alpha$$

C.I. for Mean with Known Variance (Continued)

- If $\bar{X}$ is the mean of a random sample of size n from a population with known variance σ^2,
- a $(1-\alpha)100\%$ confidence interval for μ is given by

$$\left(\bar{X} - z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)\right)$$

Sample Size for Estimating μ

- Most of the time, $\bar{X}$ will not be exactly equal to μ and the point estimate is in error.
- The size of this error will be $|\bar{X} - \mu|$.
- We know that

$$\Pr\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

or

$$\Pr\left(|\bar{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Sample Size for Estimating μ (Continued)

- Let e denote the margin of error.
- We want the error $|\bar{X} - \mu|$ does not exceed the margin of error, e, with a probability larger than $1 - \alpha$.
- That is,

$$\Pr(|\bar{X} - \mu| \leq e) \geq 1 - \alpha$$

Sample Size for Estimating μ (Continued)

- Since $\Pr\left(|\bar{X} - \mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$, therefore

$$e \geq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Hence for a given margin of error e, the sample size is given by

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

Example 1

The mean for the CAP of a random sample of 36 college seniors is calculated to be 3.5. Assuming that it is known from previous studies that $\sigma = 0.3$,

(i) Find a 95% confidence interval for the mean of the entire senior class;

(ii) How large a sample is required if we want to be 95% confidence that our estimate of μ is off by less than 0.05?

Solution to Example 1

(i) A 95% confidence interval for μ is

$$\bar{X} - z_{0.025}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{0.025}\frac{\sigma}{\sqrt{n}}$$

$$\Leftrightarrow \quad 3.5 - 1.96\frac{0.3}{6} < \mu < 3.5 + 1.96\frac{0.3}{6}$$

$$\Leftrightarrow \quad 3.402 < \mu < 3.598$$

Solution to Example 1 (Continued)

(ii) $e = 0.05, \sigma = 0.3$ and $\alpha = 0.05$ implies $z_{\alpha/2} = 1.96.$

$$n \geq \left(z_{\alpha/2}\frac{\sigma}{e}\right)^2 = \left(\frac{1.96(0.3)}{0.05}\right)^2 = 138.3.$$

Hence $n \geq 139$.

6.3.2 Unknown Variance Case

Confidence interval for mean with

(i) unknown population variance and

(ii) the population is normal or very closed to a normal distribution

(iii) the sample size is small

Unknown Variance Case (Continued)

- Let
$$T = \frac{(\bar{X} - \mu)}{S/\sqrt{n}},$$
where S^2 is the sample variance.
- We know that $T \sim t_{n-1}$.

C.I. for Mean with Unknown Variance

- Hence

$$\Pr\left(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha/2}\right) = 1-\alpha$$

or $$\Pr\left(-t_{n-1;\alpha/2} < \frac{(\bar{X}-\mu)}{S/\sqrt{n}} < t_{n-1;\alpha/2}\right) = 1-\alpha$$

or $$\Pr\left(-t_{n-1;\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \bar{X}-\mu < t_{n-1;\alpha/2}\left(\frac{S}{n}\right)\right) = 1-\alpha$$

or $$\Pr\left(\bar{X}-t_{n-1;\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X}+t_{n-1;\alpha/2}\left(\frac{S}{\sqrt{n}}\right)\right) = 1-\alpha$$

C.I. for Mean with Unknown Variance (Continued)

- If $\bar{X}$ and S are the sample mean and standard deviation of a random sample of size $n < 30$ from an approximate normal population with unknown variance σ^2,
- a $(1-\alpha)100\%$ confidence interval for μ is given by

$$\bar{X} - t_{n-1;\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + t_{n-1;\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

C.I. for Mean with Unknown Variance (Continued)

- For large n (say $n > 30$),
- the t-distribution is approximately the same as the $N(0, 1)$ distribution.
- Hence, when σ^2 is unknown,
- population is normal and
- $n > 30$,
- a $(1 - \alpha)100\%$ confidence interval for μ is given by

$$\bar{X} - z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + z_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right)$$

Example 1

- The contents of 7 similar containers of sulphuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 and 9.6 litres.

- Find a 95% confidence interval for the mean content of all such containers,

- assuming an approximate normal distribution for container contents.

Solution to Example 1

- From the data, we have $n = 7, \bar{x} = 10$ and $s^2 = 0.08$.
- Also from the statistical table, we have $t_{6,0.025} = 2.447$.
- Therefore a 95% confidence interval for the mean is given by

$$\bar{X} - t_{6;0.025}\left(\frac{S}{\sqrt{n}}\right) < \mu < \bar{X} + t_{6;0.025}\left(\frac{S}{\sqrt{n}}\right)$$

$$10 - 2.447\frac{0.2828}{\sqrt{7}} < \mu < 10 + 2.447\frac{0.2828}{\sqrt{7}}$$

$$10 - 0.262 < \mu < 10 + 0.2626$$

$$9.738 < \mu < 10.262.$$

Example 2

- A major department store chain is interested in estimating the average amount its credit card customers spent on their first visit to the chain's new store in the mall.
- Fifty credit card accounts were randomly sampled and analyzed with the following results:

$$\bar{x} = \$62.56 \text{ and } s^2 = 400.$$

Example 2 (Continued)

(a) Identify the population the department store chain is interested in learning about.

(b) Which population parameter does the chain wish to estimate?

(c) Construct a 90% confidence interval for the parameter identified in part (b).

Solution to Example 2

(a) The population is all its credit card customers.

(b) μ, the average amount spent on their first visit to the chain's new store in the mall.

Solution to Example 2 (Continued)

(c) Since n is large, we use z-value instead of t-value

From the statistical table, we have $z_{0.05} = 1.645$.

Therefore a 90% confidence interval for the mean is given by

$$\bar{X} - z_{0.05}(S/\sqrt{n}) < \mu < \bar{X} + z_{0.05}(S/\sqrt{n})$$

From the data, we have $\bar{x} = 62.56$ and $s = 400$. Hence, the 90% CI for μ is given by

$$62.56 - 1.645\sqrt{400/50} < \mu < 62.56 + 1.645\sqrt{400/50}$$

$$57.907 < \mu < 67.213.$$

6.4 Confidence Intervals for the Difference between Two Means

- If we have two populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 respectively,

- Then

$$\bar{X}_1 - \bar{X}_2$$

is the point estimator of $\mu_1 - \mu_2$.

6.4.1 Known Variances

- $\boldsymbol{\sigma_1^2}$ **and** $\boldsymbol{\sigma_2^2}$ **are known and not equal, and the two populations are normal,**

- **or when** $\boldsymbol{\sigma_1^2}$ **and** $\boldsymbol{\sigma_2^2}$ **are known and not equal, but** $\boldsymbol{n_1, n_2}$ **are sufficiently large (**$\boldsymbol{n_1 \geq 30, n_2 \geq 30}$**)**

- According to Section 5.5, we have

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right).$$

6.4.1 Known Variances

- We can assert that

$$\Pr\left(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}\right) = 1 - \alpha$$

Known Variances (Continued)

which leads to the following $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2$$

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

Example 1

- A study was conducted in which two types of engines, A and B, were compared.
- Gas mileage, in miles per gallon, was measured.
- 50 experiments were conducted using engine A
- 75 experiments were done for engine type B.
- The gasoline used and other conditions were held constant.

Example 1 (Continued)

- The average gas mileage for 50 experiments using engine A was 36 miles per gallon and
- The average gas mileage for the 75 experiments using machine B was 42 miles per gallon.
- Find a 96% confidence interval on $\mu_B - \mu_A$, where μ_A and μ_B are population mean gas mileage for machine types A and B, respectively.
- Assume that the population standard deviations are 6 and 8 for machine types A and B, respectively.

Solution to Example 1

- From the given info, we have $\bar{x}_A = 36$ and $\bar{x}_B = 42$.
- Hence the point estimate for $\mu_B - \mu_A$ is

$$\bar{x}_B - \bar{x}_A = 42 - 36 = 6.$$

- We also know that $\sigma_1 = 6$ and $\sigma_2 = 8$.
- Use $\alpha = 0.04$, we find $z_{0.02} = 2.05$.

Solution to Example 1 (Continued)

- Since the sample sizes are large, therefore the 96% confidence interval for $\mu_B - \mu_A$ is

$$6 - 2.05\sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_B - \mu_A < 6 + 2.05\sqrt{\frac{64}{75} + \frac{36}{50}}$$

$$3.428 < \mu_B - \mu_A < 8.571.$$

6.4.2 Large Sample C.I. for Unknown Variances

- $\boldsymbol{\sigma_1^2}$ **and** $\boldsymbol{\sigma_2^2}$ **are unknown**
- $\boldsymbol{n_1, n_2}$ **are sufficiently large** $\boldsymbol{(n_1 \geq 30, n_2 \geq 30)}$
- we may replace by σ_1^2 and σ_2^2 by their estimates, S_1^2 and S_2^2,

Large Sample C.I. for Unknown Variances (Continued)

- A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by:

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2$$

$$< (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}.$$

Example 2

- Two kinds of thread are being compared for strength.
- 50 pieces of each type of thread are tested under similar conditions.
- Brand A had an average tensile strength of 78.3 kg with a s.d. of 5.6 kg.
- Brand B had an average tensile strength of 87.2 kg with a s.d. of 6.3 kg.
- Construct a 95% confidence interval for the difference of the population means.

Solution to Example 2

- From the given info, we have $\bar{x}_A = 78.3$ and $\bar{x}_B = 87.2$.
- Hence the point estimate for $\mu_A - \mu_B$ is

$$\bar{x}_A - \bar{x}_B = 78.3 - 87.2 = -8.9.$$

- We also know from the data that $s_1 = 5.6$ and $s_2 = 6.3$.
- $\alpha = 0.05$ implies $z_{0.025} = 1.96$.

Solution to Example 2 (Continued)

- Since the sample sizes are large, therefore an approximate 95% confidence interval for $\mu_A - \mu_B$ is

$$-8.9 - 1.96\sqrt{\frac{5.6^2}{50} + \frac{6.3^2}{50}} < \mu_A - \mu_B < -8.9 + 1.96\sqrt{\frac{5.6^2}{50} + \frac{6.3^2}{50}}$$

$$-11.236 < \mu_A - \mu_B < -6.564.$$

6.4.3 Unknown but Equal Variances

- $\boldsymbol{\sigma_1^2}$ **and** $\boldsymbol{\sigma_2^2}$ **are unknown but equal and**
- **the two populations are normal**
- **Small sample sizes (**$\boldsymbol{n_1 \leq 30}$ **and** $\boldsymbol{n_2 \leq 30}$**)**
- Let $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then

$$(\bar{X}_1 - \bar{X}_2) \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right).$$

Unknown but Equal Variances (Continued)

- Therefore we obtain a standard normal variable in the form

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Unknown but Equal Variances (Continued)

- σ^2 can be estimated by the pooled sample variance

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

with S_1^2 and S_2^2 being the sample variances of the first and second samples respectively.

Unknown but Equal Variances (Continued)

- Note that if the two populations are normal with the same variance σ^2, then

$$\frac{(n_1 - 1)S_1^2}{\sigma^2} \sim \chi^2_{n_1 - 1} \quad \text{and} \quad \frac{(n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_2 - 1},$$

Hence

$$\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2} \sim \chi^2_{n_1 + n_2 - 2}.$$

Unknown but Equal Variances (Continued)

- Substituting S_p^2 for σ^2, we obtain the statistic

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}.$$

Unknown but Equal Variances (Continued)

- We can assert that

$$\Pr\left(-t_{n_1+n_2-2;\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} < t_{n_1+n_2-2;\alpha/2}\right)$$
$$= 1 - \alpha.$$

Unknown but Equal Variances (Continued)

- Therefore a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) - t_{n_1+n_2-2;\alpha/2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} < \mu_1 - \mu_2$$

$$< (\bar{X}_1 - \bar{X}_2) + t_{n_1+n_2-2;\alpha/2}\, S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Unknown but Equal Variances (Continued)

where S_p is the pooled estimate of the population standard deviation and $t_{n_1+n_2-2;\alpha/2}$ is the value from the t-distribution with the degrees of freedom $n_1 + n_2 - 2$, leaving an area of $\alpha/2$ to the right.

[i.e. $\Pr\left(T_{n_1+n_2-2} > t_{n_1+n_2-2\,;\alpha/2}\right) = \alpha/2$.]

Unknown but Equal Variances for Large Samples

- Note that for large samples such that $n_1 \geq 30$ and $n_2 \geq 30$, we can replace $t_{n_1+n_2-2;\alpha/2}$ by $z_{\alpha/2}$ in the above formula.
- Therefore a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2$$

$$< (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Example 3

- A course in mathematics is taught to 12 students by the conventional classroom procedure.
- A second group of 10 students was given the same course by means of programmed materials.
- At the end of the semester the same examination was given to each group.
- The 12 students meeting in the classroom made an average grade of 85 with standard deviation of 4.

Example 3 (Continued)

- The 10 students using programmed materials made an average of 81 with a standard deviation of 5.
- Find a 90% confidence interval for the difference between the population means,
- assuming the populations are approximately normally distributed with equal variances.

Solution to Example 3

- Let μ_1 and μ_2 represent the average grades of all students who might take this course by the classroom and programmed presentations respectively.
- So $\bar{x}_1 - \bar{x}_2 = 85 - 81 = 4$ is the point estimate for $\mu_1 - \mu_2$.
- Since $\sigma_1^2 = \sigma_2^2$, we estimate the population variance by the pooled variance

$$s_p^2 = \frac{11(16) + 9(25)}{12 + 10 - 2} = 20.05.$$

Solution to Example 3 (Continued)

- A 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) - t_{20;\,0.05}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2$$

$$< (\bar{X}_1 - \bar{X}_2) + t_{20;\,0.05}\, S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Solution to Example 3 (Continued)

- From the given data, we have a 90% confidence interval for $\mu_1 - \mu_2$ is given by

$$4 - 1.7247\sqrt{20.05}\sqrt{\frac{1}{12} + \frac{1}{10}} < \mu_1 - \mu_2$$

$$< 4 + 1.7247\sqrt{20.05}\sqrt{\frac{1}{12} + \frac{1}{10}}$$

$$0.693 < \mu_1 - \mu_2 < 7.307.$$

6.4.4 C.I. for the difference between two means for paired data (dependent data)

- If we run a test on a new diet using 15 individuals, the weights before (x_i) and after (y_i) completion of the test form our two samples.
- Observations in the two samples made on the same individual are related and hence form a pair.
- To determine if the diet is effective, we must consider the differences d_i $(= x_i - y_i)$ of paired observations.

C.I. for the difference between two means for paired data (Continued)

- These differences are the values of a random sample $d_1, d_2, \cdots, d_n$ from a population that we shall assume to be normal with mean μ_D and unknown variance σ_D^2.
- In fact $\mu_D = \mu_1 - \mu_2$ and the point estimate of μ_D is given by

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)$$

- The point estimate of σ_D^2 is given by

$$s_D^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2.$$

6.4.4.1 Small Sample and Approximate Normal Population

- A $(1-\alpha)100\%$ confidence interval for μ_D can be established by writing

$$\Pr\left(-t_{n-1;\alpha/2} < T < t_{n-1;\alpha/2}\right) = 1-\alpha,$$

where $T = \frac{\bar{d}-\mu_D}{s_d/\sqrt{n}} \sim t_{n-1}$ distribution.

- Therefore a $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is given by

$$\bar{d} - t_{n-1;\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right) < \mu_D < \bar{d} + t_{n-1;\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right).$$

For large sample (n > 30)

- For **sufficiently large** sample, we may replace $t_{n-1;\alpha/2}$ by $z_{\alpha/2}$ and

- a $(1 - \alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is given by

$$\bar{d} - z_{\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right) < \mu_D < \bar{d} + z_{\alpha/2}\left(\frac{s_D}{\sqrt{n}}\right).$$

Example 4

- Twenty students were divided into 10 pairs,
- each member of the pair having approximately the same IQ.
- One of each pair was selected at random and
- assigned to a mathematics section using programmed materials only.
- The other member of each pair was assigned to a section in which the professor lectured.
- At the end of the semester each group was given the same examination and the following results were recorded.

Example 4 (Continued)

Pair	1	2	3	4	5	6	7	8	9	10
P.M.	76	60	85	58	91	75	82	64	79	88
Lecture	81	52	87	70	86	77	90	63	85	83
d	–5	8	–2	–12	5	–2	–8	1	–6	5

- Find a 98% confidence interval for the true difference in the two learning procedures.

Solution to Example 4

- From the data, we have

$$\bar{d} = \frac{1}{10}\sum_{i=1}^{10} d_i = -1.6 \quad \text{and}$$

$$s_D^2 = \frac{1}{9}\left(\sum_{i=1}^{10} d_i^2 - 10\bar{d}^2\right) = 40.71.$$

$\alpha = 0.02$ implies that $t_{9;0.01} = 2.281.$

Solution to Example 4 (Continued)

- Therefore, a 98% confidence interval for μ_D is

$$\bar{d} - t_{9;0.01}\left(\frac{s_D}{\sqrt{10}}\right) < \mu_D < \bar{d} + t_{9;0.01}\left(\frac{s_D}{\sqrt{10}}\right)$$

$$-1.6 - 2.821\sqrt{\frac{40.71}{10}} < \mu_D < -1.6 + 2.821\sqrt{\frac{40.71}{10}}$$

$$-7.292 < \mu_D < 4.092.$$

6.5 C.I. for Variances and Ratio of Variances

6.5.1 Confidence intervals for a variance (of a normal population)

- Let $X_1, X_2, \cdots, X_n$ be a random sample of size n from a (approximately) $N(\mu, \sigma^2)$ distribution.
- Then the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)$$

is a **point estimate** of σ^2.

Case 1 μ is known

- When μ is known, we have

$$\frac{X_i - \mu}{\sigma} \sim N(0, 1) \qquad \text{for all } i$$

or

$$\left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(1) \qquad \text{for all } i$$

and hence

$$\sum_{i=1}^{n} \frac{(X_i - \mu)^2}{\sigma^2} \sim \chi^2(n).$$

Case 1 μ is known (Continued)

- Therefore

$$\Pr\left(\chi^2_{n;1-\alpha/2} < \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} < \chi^2_{n;\,\alpha/2}\right) = 1-\alpha$$

and hence

$$\Pr\left(\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n;\,\alpha/2}} < \sigma^2 < \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n;1-\alpha/2}}\right) = 1-\alpha$$

Case 1 μ is known (Continued)

- Therefore, a $(1-\alpha)100\%$ confidence interval for σ^2 of $N(\mu, \sigma^2)$ population with μ known is

$$\frac{\sum_{I=1}^{n}(X_i-\mu)^2}{\chi^2_{n;\,\alpha/2}} < \sigma^2 < \frac{\sum_{I=1}^{n}(X_i-\mu)^2}{\chi^2_{n;1-\alpha/2}}$$

Case 2 μ is unknown

- When μ is unknown, we have

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

Case 2 μ is unknown (Continued)

- The above results are true for both small and large n. Therefore,

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

and hence

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

Case 2 μ is unknown (Continued)

- Therefore, a $(1-\alpha)100\%$ confidence interval for σ^2 of $N(\mu, \sigma^2)$ population with μ unknown is

$$\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}$$

where S^2 is the sample variance.

Remarks

- A $(1-\alpha)100\%$ confidence interval for σ is obtained by taking the square root of each end point of the interval for σ^2.
- When μ is known, a $100(1-\alpha)\%$ C.I. for σ is

$$\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n;\,\alpha/2}}} < \sigma < \sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\chi^2_{n;1-\alpha/2}}}.$$

Remarks (Continued)

- When μ is unknown, a $100(1-\alpha)\%$ C.I. for σ is

$$\sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}}} < \sigma < \sqrt{\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}}.$$

- Notice that the parameter (or the degrees of freedom) of the χ^2-distribution changes from n to $n-1$ when μ is unknown.

Example 1

- The following are the volume, in decilitres, of 10 cans of peaches distributed by a certain company:

  46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2 and 46.0.
- Find a 95% confidence interval for the variance of all such cans of peaches distributed by this company,
- assuming volume to be a normally distributed variable.

Solution to Example 1

- From the data, we have
$$s^2 = \frac{1}{9}\left(\sum_{i=1}^{10} x_i^2 - 10\bar{x}^2\right) = 0.286$$
- From the statistical table, we have
$$\chi^2_{9;0.025} = 19.023 \text{ and } \chi^2_{9;0.975} = 2.7.$$

Solution to Example 1 (Continued)

- Substituting these values in the following formula, we have

$$\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}$$

- We obtain a 95% confidence interval for σ^2

$$\frac{9(0.2862)}{19.023} < \sigma^2 < \frac{9(0.2862)}{2.7} \text{ or } 0.135 < \sigma^2 < 0.954.$$

6.5.2 C.I. for the ratio of two variances (of normal population) with unknown means

- Let $X_1, X_2, \cdots, X_{n_1}$ be a random sample of size n_1 from a (or approximately) $N(\mu_1, \sigma_1^2)$ population and

- $Y_1, Y_2, \cdots, Y_{n_2}$ be a random sample of size n_2 from a (or approximately) $N(\mu_2, \sigma_2^2)$ population,

- where μ_1 and μ_2 are unknown.

C.I. for the ratio of two variances (of normal population) with unknown means (Continued)

- Then

$$\frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1-1) \text{ and } \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2-1)$$

where $S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(X_i - \bar{X})^2$ and

$S_2^2 = \frac{1}{n_2-1}\sum_{j=1}^{n_2}\left(Y_j - \bar{Y}\right)^2$.

C.I. for the ratio of two variances (of normal population) with unknown means (Continued)

- Hence

$$F = \frac{\dfrac{(n_1-1)S_1^2}{\sigma_1^2}\Big/(n_1-1)}{\dfrac{(n_2-1)S_2^2}{\sigma_2^2}\Big/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1).$$

- We then can assert that

$$\Pr\left(F_{n_1-1,n_2-1;1-\alpha/2} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{n_1-1,n_2-1;\,\alpha/2}\right) = 1-\alpha.$$

C.I. for the ratio of two variances (of normal population) with unknown means (Continued)

- Therefore

$$\Pr\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;\,\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\right)$$
$$= 1-\alpha.$$

where $\Pr\left(\boldsymbol{F_{n_1-1,n_2-1}} \geq F_{n_1-1,n_2-1;\,\alpha/2}\right) = \alpha/2$ with $\boldsymbol{F_{n_1-1,n_2-1}}$ denote a random variable following an F-distribution with parameters (n_1-1) and (n_2-1).

C.I. for the ratio of two variances (of normal population) with unknown means (Continued)

- Hence, a $100(1-\alpha)\%$ confidence interval for the ratio σ_1^2/σ_2^2 when μ_1 and μ_2 are unknown

$$\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;\,\alpha/2}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1;\,\alpha/2}$$

since $F_{n_1-1,n_2-1;1-\alpha/2} = \dfrac{1}{F_{n_2-1,n_1-1;\,\alpha/2}}$. (See p6.95)

C.I. for the ratio of two variances (of normal population) with unknown means (Continued)

$$\Pr\left(\boldsymbol{F_{n_1-1,n_2-1}} > F_{n_1-1,n_2-1;1-\alpha/2}\right) = 1-\alpha/2$$

$$\Longrightarrow \quad \Pr\left(\frac{1}{\boldsymbol{F_{n_1-1,n_2-1}}} < \frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\right) = 1-\alpha/2$$

$$\Longrightarrow \quad \Pr\left(\boldsymbol{F_{n_2-1,n_1-1}} < \frac{1}{F_{n_1-1,n_2-1;1-\alpha/2}}\right) = 1-\alpha/2$$

C.I. for the ratio of two variances (of normal population) with unknown means (Continued)

But

$$\Pr\left(\boldsymbol{F_{n_2-1,n_1-1}} < F_{n_2-1,n_1-1;\alpha/2}\right) = 1 - \alpha/2$$

Hence

$$F_{n_1-1,n_2-1;1-\alpha/2} = \frac{1}{F_{n_2-1,n_1-1;\,\alpha/2}}.$$

Remark

- A $(1-\alpha)100\%$ confidence interval for σ_1/σ_2 is obtained by taking the square root of each end point of the interval for σ_1^2/σ_2^2.
- When μ_1 and μ_2 are unknown, a $100(1-\alpha)\%$ C.I. for σ_1^2/σ_2^2 is

$$\sqrt{\frac{S_1^2}{S_2^2}\frac{1}{F_{n_1-1,n_2-1;\,\alpha/2}}} < \frac{\sigma_1}{\sigma_2} < \sqrt{\frac{S_1^2}{S_2^2}F_{n_2-1,n_1-1;\,\alpha/2}}.$$

Example 2

- A standardized placement test in mathematics was given to 25 boys and 16 girls.
- The boys made an average grade of 82 with a standard deviation of 8, while the girls made an average grade of 78 with a standard deviation of 7.
- Find a 98% confidence interval for σ_1^2/σ_2^2 and σ_1/σ_2,
- where σ_1^2 and σ_2^2 are the variances of the populations of grades for all boys and girls, respectively, who at some time have taken or will take this test.
- Assume the populations to be normally distributed.

Solution to Example 2

- From the data, we have $n_1 = 25, s_1 = 8, n_2 = 16$ and $s_2 = 7$.
- Take $\alpha = 0.02$.
- $F_{24,15;\,0.01} = (3.45 + 3.18)/2 = 3.305$ and
- $F_{15,24;0.01} = 3.03 - 3(3.03 - .66)/(24 - 12) = 2.94$.
- By the formula, we obtain the 98% confidence interval for σ_1^2/σ_2^2 is

$$\frac{64}{49}\frac{1}{3.305} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{64}{49}\,2.94 \quad \text{or} \quad 0.395 < \frac{\sigma_1^2}{\sigma_2^2} < 3.84.$$

Solution to Example 2 (Continued)

- the 98% confidence interval for σ_1/σ_2 is given by

$$\sqrt{0.395} < \sqrt{\frac{\sigma_1^2}{\sigma_2^2}} < \sqrt{3.84}$$

or

$$0.628 < \frac{\sigma_1}{\sigma_2} < 1.960.$$